

# Revisiting MLLM Based Image Quality Assessment: Errors and Remedy

Zhenchen Tang<sup>1,2</sup>, Songlin Yang<sup>3</sup>, Bo Peng<sup>1\*</sup>, Zichuan Wang<sup>1,2</sup>, Jing Dong<sup>1\*</sup>

<sup>1</sup>New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

<sup>2</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences

<sup>3</sup>MMLab@HKUST, The Hong Kong University of Science and Technology

{tangzhenchen2024, wangzichuan2024}@ia.ac.cn, syangds@connect.ust.hk, {bo.peng, jdong}@nlpr.ia.ac.cn

## Abstract

The rapid progress of multi-modal large language models (MLLMs) has boosted the task of image quality assessment (IQA). However, a key challenge arises from the inherent mismatch between the discrete token outputs of MLLMs and the continuous nature of quality scores required by IQA tasks. This discrepancy significantly hinders the performance of MLLM-based IQA methods. Previous approaches that convert discrete token predictions into continuous scores often suffer from conversion errors. Moreover, the semantic confusion introduced by level tokens (e.g., “good”) further constrains the performance of MLLMs on IQA tasks and degrades their original capabilities to related tasks. To tackle these problems, we provide a theoretical analysis of the errors inherent in previous approaches and, motivated by this analysis, propose a simple yet effective framework, Q-Scorer. This framework incorporates a lightweight regression module and IQA-specific score tokens into the MLLM pipeline. Extensive experiments demonstrate that Q-Scorer achieves state-of-the-art performance across multiple IQA benchmarks, generalizes well to mixed datasets, and further improves combined with other methods.

**Code** — <https://github.com/2kxx/Q-Scorer>

## 1 Introduction

Image Quality Assessment (IQA) is a fundamental task in computer vision, aiming to quantify the perceptual quality of images in a way that aligns closely with human visual perception. With the recent advances in reinforcement learning-based post-training paradigms, IQA has become increasingly important as a reward signal for downstream tasks, such as image generation (Liang et al. 2024; Yu et al. 2024; Li et al. 2025c,b; Han et al. 2025) and enhancement (Zheng et al. 2021; Zhou and Wang 2022; Wang et al. 2025).

Recently, leveraging MLLMs (Liu et al. 2023; Ye et al. 2024; Bai et al. 2025; Chen et al. 2024c) to improve IQA performance has shown promise, as MLLMs can understand visual images using linguistic intelligence, which greatly facilitates aligning model predictions with human evaluations across diverse image types. However, existing MLLM-based

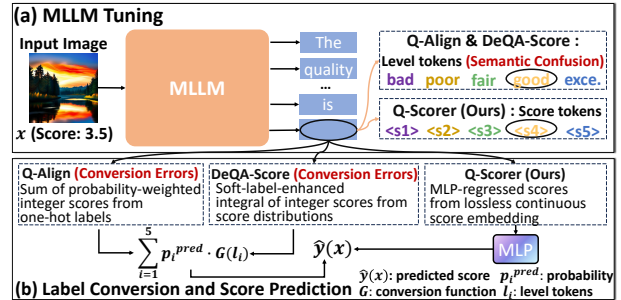


Figure 1: Overview of MLLM-based IQA methods and error analysis. The figure shows how MLLMs are adapted for label conversion and score prediction, and highlights steps causing conversion errors and semantic confusion.

IQA methods (Wu et al. 2023a,b, 2024) typically formulate the task as learning the correlation between images and discrete textual tokens (e.g., “poor,” “good”), which fits the token output nature of MLLMs but conflicts with the intrinsic nature of IQA — predicting continuous quality scores. As shown in Fig.1, although several studies (Wu et al. 2023b; You et al. 2025; Li et al. 2025a) has explored techniques to bridge this mismatch, two key problems remain:

(1) Conversion Errors: Existing token-based approaches suffer from inevitable theoretical errors during the label conversion process. Since real-valued Mean Opinion Scores (MOSs) are quantized into discrete tokens, even perfect classification cannot exactly recover the ground-truth MOS. We theoretically analyze this issue and prove that such errors are intrinsic to token-based IQA formulations in Sec. 3.1.

(2) Semantic Confusion: Most quality-related textual descriptions (i.e., level tokens) used in current methods are drawn from pre-trained LLM vocabularies, whose semantics are not optimized for IQA. This introduces two problems: pre-tuning semantic confusion, where these tokens carry prior meanings unrelated to visual quality; post-tuning semantic confusion, where training on IQA-specific tasks may overwrite the semantic integrity of these tokens.

To address these problems, we propose **Q-Scorer**, a simple yet effective framework that incorporates a lightweight regression module and IQA-specific score tokens into the MLLM pipeline. For conversion errors, we adopt an MLP-based module to regress the continuous IQA scores. Al-

\*Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

though prior work used a linear layer on the final token (He et al. 2024), its entanglement with IQA-irrelevant information limits performance compared to token-based methods (even the baseline Q-Align), which causes insufficient embedding expressiveness. To address this issue and mitigate semantic confusion, we introduce a set of IQA-specific score tokens. These tokens guide the MLP module to improve IQA score prediction without compromising the original capabilities of MLLMs.

Equipped with these designs, our method achieves state-of-the-art performance using only LoRA-based fine-tuning on limited number of model parameters. Q-Scorer consistently outperforms baseline approaches in score regression, and despite being trained solely on the KonIQ dataset, it also exhibits strong generalization to other datasets. Furthermore, it generalizes well to mixed-dataset settings and can be seamlessly integrated with other methods to achieve even better performance.

Our contributions can be summarized as follows:

- We conduct error analysis and find that existing MLLM-based IQA methods suffer from conversion errors and semantic confusion when using discrete level tokens.
- To address these challenges, we propose Q-Scorer, a simple yet effective framework that leverages IQA-specific score tokens and an MLP regressor.
- Extensive experiments demonstrate that our framework achieves state-of-the-art performance across multiple benchmarks, generalizes well to mixed datasets, and further improves combined with other methods.

## 2 Related Work

### 2.1 Traditional IQA Methods

Prior to the emergence of MLLMs, traditional IQA methods were generally categorized into full-reference (FR) and no-reference (NR) approaches. Classical FR methods rely on handcrafted similarity metrics that compare a distorted image with its high-quality reference, such as PSNR (Hore and Ziou 2010) and SSIM (Wang et al. 2004). In contrast, NR methods estimate perceptual quality directly from a single image using natural scene statistics, as exemplified by BRISQUE (Mittal, Moorthy, and Bovik 2012) and NIQE (Mittal, Soundararajan, and Bovik 2012). With the advancement of deep learning, data-driven IQA approaches have achieved significant improvements by directly regressing human perceptual scores from raw image inputs (Bosse et al. 2017; Talebi and Milanfar 2018; Su et al. 2020). More recent studies further boost performance and generalization by incorporating multi-scale features (Ke et al. 2021), co-training on multiple datasets (Zhang et al. 2021), multitask learning (Zhang et al. 2023), or leveraging pretrained vision-language models such as CLIP (Wang, Chan, and Loy 2023; Tang et al. 2024).

### 2.2 MLLM-Based IQA Methods

MLLM-based IQA methods leverage the foundational knowledge embedded in MLLMs to improve IQA performance and enhance generalization. Q-Bench (Wu et al.

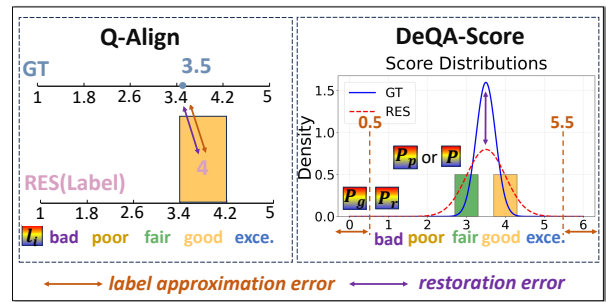


Figure 2: Visual illustration and detailed explanation of conversion errors. The figure provides examples detailing two main sources of conversion errors: label approximation (from discretizing MOS) and restoration error (from imperfect score restoration). GT means assumed ground-truth Gaussian distribution. RES means distribution restored from soft labels (DeQA). See Sec. 3.2 and Fig.3 for details.

2023a) demonstrates that MLLMs possess low-level visual perception and understanding, enabling them to predict quantifiable quality scores. Q-Instruct (Wu et al. 2024) improves question-answering accuracy including IQA tasks, by introducing instruction-response style description dataset. DepictQA (You et al. 2024b) and its extension (You et al. 2024a) generates detailed quality assessments by training on large-scale explanation dataset. Moreover, Q-Ground (Chen et al. 2024a) and SEAGULL (Chen et al. 2024b) extend IQA to more fine-grained tasks, including visual quality grounding and region-aware scoring.

### 2.3 IQA Scorers

Although numerous MLLM-based IQA methods have emerged recently, their scoring strategies are limited and can only be categorized into **token-based** and **regression-based approaches**, depending on whether they utilize an MLP (Li et al. 2025a). We refer to the methods proposing these scoring strategies as IQA Scorers. First are token-based methods. Initially, Q-Bench (Wu et al. 2023a) introduces a binary softmax strategy to predict quality scores by two discrete levels. Compare2Score (Zhu et al. 2024) adopts a pairwise comparison strategy to infer image quality scores. Building on Q-Bench, the widely adopted method Q-Align (Wu et al. 2023b) improves scoring process by discretizing quality into five levels and training MLLMs using one-hot labels. DeQA-Score (You et al. 2025) further enhances Q-Align by introducing soft labels based on score distributions. RealQA (Li et al. 2025a) takes a different approach by directly predicting two extra significant digits of the score.

In contrast, regression-based methods are rarely explored in MLLM-based IQA. While MLPs are commonly used in traditional IQA models, their application within MLLMs remains limited due to the dominance of the next-token prediction paradigm, which makes it challenging to integrate continuous regression objectives effectively. Among the few existing attempts, VideoScore (He et al. 2024) performs direct numerical regression via a linear output layer. However, its insufficient embedding expressiveness limits its performance compared to token-based methods.

---

**Algorithm 1: Q-Align**

---

```

1: Input: Image  $x$ 
2: Target: Normalized ground-truth MOS  $S \in [1, 5]$ 
3: Output: Predicted quality score  $\hat{y}(x)$ 
4: Training Phase:
5: Divide the range  $[1, 5]$  into 5 intervals  $I_1, I_2, \dots, I_5$ 
6: for  $i = 1$  to 5 do
7:   if  $1 + \frac{4}{5}(i - 1) < S \leq 1 + \frac{4}{5}i$  then
8:     Assign label  $L(S) \leftarrow l_i$ 
9:   end if
10: end for
11: Level tokens:  $\{l_i\} = \{bad, poor, fair, good, excellent\}$ 
12: Inference Phase:
13: Predict probability  $p_i^{pred}$  for each level token  $l_i$ 
14:  $G(l_i)$  convert  $l_i$  to discrete score  $i$ 
15: Predict final score:  $\hat{y}(x) = \sum_{i=1}^5 p_i^{pred} \cdot G(l_i)$ 
16: return  $\hat{y}(x)$ 

```

---

### 3 Error Analysis: Conversion Errors and Semantic Confusion

Previous methods for formulating MLLM-based IQA typically adopt discrete level tokens (i.e., “bad”, “poor”, “fair”, “good”, “excellent”) to describe image quality. However, this level-token-based formulation introduces two potential sources of errors: conversion errors (Sec. 3.1) and semantic confusion (Sec. 3.2). For conversion errors, we analyze the limitations of discrete token-based score fitting and explain why modeling IQA as a regression task can mitigate this issue. For semantic confusion, we present insights from both pre-tuning and post-tuning perspectives, highlighting the problems arising from directly using level tokens.

#### 3.1 Conversion Errors

To align with the discrete token outputs of MLLMs, ground-truth IQA scores are typically quantized into class-level labels. However, as illustrated in Fig.2, this design introduces two types of conversion errors: **label approximation error** and **restoration error**. The label approximation error, which arises from converting a continuous ground-truth MOS  $S$  into a discrete label that MLLMs can process. The restoration error, referring to the discrepancy between the predicted score  $\hat{y}(x)$  (where  $x$  is an input image) and the ground-truth MOS  $S$  due to imperfect restoration from the learned label representation. In this section, we first formulate the errors of representative methods and then analyze why regression-based modeling can effectively mitigate conversion errors. The proof details are provided in Appendix.

##### (a) Error Formulation: Q-Align and DeQA-Score

**Q-Align** serves as the foundational paradigm for most current MLLM-based IQA methods. As shown in Algorithm 1, it learns from one-hot score labels and predicts the final score using a weighted sum of integer scores from level token probabilities. However, this process involves only discrete textual supervision, leading to the loss of continuous score information during training.

Assuming that the model makes a perfectly accurate pre-

---

**Algorithm 2: DeQA-Score**

---

```

1: Input: Image  $x$ 
2: Target: Assume  $S$  follows Gaussian distribution (GT):  $s \sim \mathcal{N}(\mu, \sigma^2)$ . MOS as  $\mu$  and annotated variance as  $\sigma^2$ 
3: Output: Predicted quality score  $\hat{y}(x)$ 
4: Training Phase:
5: Divide the range  $[0.5, 5.5]$  into 5 intervals  $I'_1, I'_2, \dots, I'_5$ 
6: Define soft label  $P_r = \{p_1^{raw}, p_2^{raw}, p_3^{raw}, p_4^{raw}, p_5^{raw}\}$ 
7: Define interval midpoint  $c_i \in \{1, 2, 3, 4, 5\}$ 
8: for  $i = 1$  to 5 do
9:    $p_i^{raw} = \int_{c_i - \frac{1}{2}}^{c_i + \frac{1}{2}} f(s) ds$ 
10: end for
11: Enhancing soft label accuracy:
12: Linearly transform  $P_r$  to enhanced soft label  $P$ :
13:  $p_i = \alpha \cdot p_i^{raw} + \beta$ , with  $\sum_i p_i = 1$ ,  $\sum_i p_i c_i = \mu$ 
14: Use model to predict level token distribution  $P_p$ 
15:  $P_p = \{p_1^{pred}, p_2^{pred}, p_3^{pred}, p_4^{pred}, p_5^{pred}\}$ 
16: Minimize KL divergence:  $\mathcal{L}_{KL}(P \| P_p)$ 
17: Inference Phase:
18: Predict final score:  $\hat{y}(x) = \sum_{i=1}^5 p_i^{pred}(x) \cdot c_i$ 
19: return  $\hat{y}(x)$ 

```

---

diction (i.e., assigning probability 1 to the correct discrete label), the expected predicted score  $\hat{y}(x)$  will be  $G(l_j)$ , where  $j$  denotes the index of the predicted level token. In Fig.2, the model predicts level token  $l_4$ , and the resulting score is  $\hat{y}(x) = G(l_4) = 4$ . In this case, the label approximation error can be regarded as equivalent to the restoration error. The one-sample error  $\epsilon(x)$  can thus be calculated as:

$$\epsilon(x) = |G(l_j) - S| \geq 0. \quad (1)$$

Under this assumption, we treat the error  $\epsilon(x)$  as uniformly distributed within the interval. Therefore, the expected theoretical error  $E[\epsilon(x)^2] = \frac{18}{125} > 0$ .

**DeQA-Score**, as shown in Algorithm 2, learns from soft-label-enhanced integral of integer scores from the score distribution and predicts the final score via the same probability-weighted summation as Q-Align, which is proposed to mitigate the label approximation error in Q-Align.

However, this formulation introduces certain theoretical errors. Assuming the model makes a perfectly accurate prediction, two sources of error remain: (1) The Gaussian prior itself is an approximation of the true, human rating distribution. This modeling discrepancy leads to an inherent systematic error  $\epsilon_1(x)$ . (2) The interval-based truncation leads to  $\sum_i p_i^{raw} < 1$ , which means that raw soft label  $P_r$  does not form a strictly valid discrete probability distribution.

As a result, the label and the predicted score  $\hat{y}(x)$  inevitably deviate from the ground-truth MOS  $S$ , leading to a label approximation error  $\epsilon_2(x)$ :

$$\epsilon_2(x) = |\hat{y}(x) - S| = \left| \sum_{i=1}^5 p_i^{raw} c_i - \int_{-\infty}^{\infty} s f(s) ds \right| > 0, \quad (2)$$

Although DeQA-Score applies a linear transformation to refine  $P_r$  into enhanced soft label  $P$ , theoretically reducing

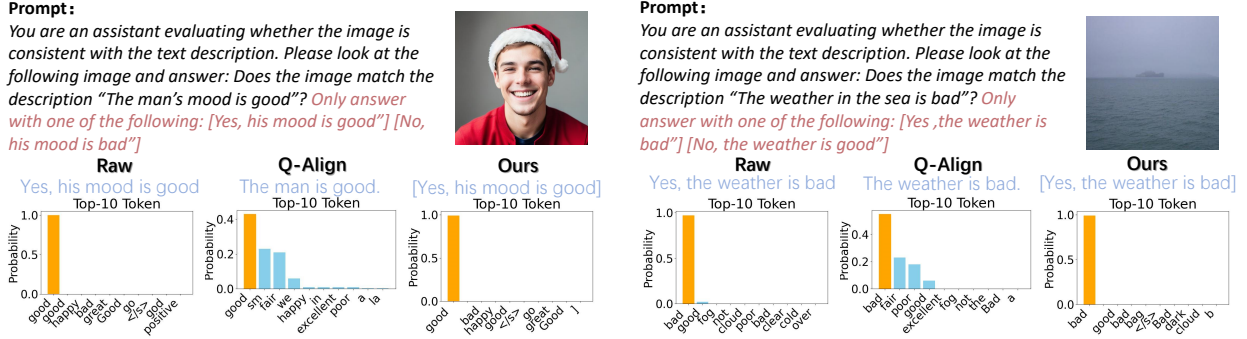


Figure 3: Examples of post-tuning semantic confusion, showing how different token strategies affect T2I alignment assessment.

the label approximation error to zero. As shown in Fig.2, the process of using the discrete distribution  $P$  to recover the original continuous Gaussian distribution  $P_g$  (GT) inevitably causes a restoration error:

$$\mu^{res} = \sum p_i^{pred} c_i, \quad (\sigma^{res})^2 = \sum p_i^{pred} (c_i - \mu^{res})^2 \quad (4)$$

This error is severe when the original variance is small, resulting in a significant increase in error  $\epsilon_1$  (You et al. 2025).

Combining these, the error is always greater than zero:

$$\epsilon(x) = \epsilon_1(x) + \epsilon_2(x) > 0, \quad (4)$$

Assuming  $\epsilon_1$  and  $\epsilon_2$  are independent, the total theoretical error is always greater than zero:

$$E[\epsilon(x)^2] = E[\epsilon_1^2] + E[\epsilon_2^2] > 0. \quad (5)$$

### (b) Error Mitigation: Regression-Based Modeling

Given the errors discussed above, we explore regression-based modeling as a mitigation. In regression-based models, the predicted score  $\hat{y}(x)$  is obtained via a regression function  $\mathcal{F}(x; \theta)$  over the input feature  $x$ :

$$\hat{y}(x) = \mathcal{F}(x; \theta). \quad (6)$$

Since regression-based models use the ground-truth MOS  $S$  directly as supervision without discretization, their theoretical label approximation error is inherently zero. Therefore, only the restoration error needs to be considered. Let  $g(x)$  denote the function mapping input  $x$  to the corresponding MOS  $S$ . Then, for any  $\epsilon > 0$ , there exists a feedforward neural network with a single hidden layer of the form:

$$\mathcal{F}(x) = \sum_{i=1}^n \alpha_i \cdot \sigma(w_i^\top x + b_i), \quad (7)$$

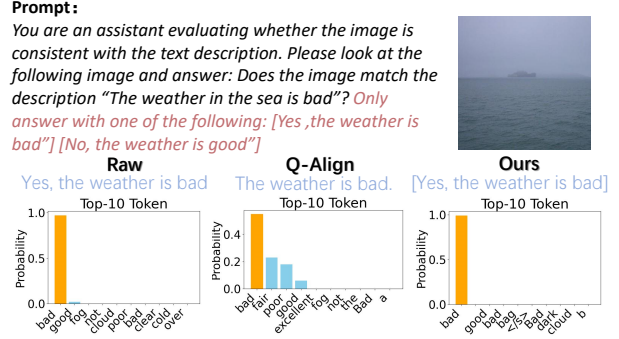
where  $\sigma(\cdot)$  is an activation function (e.g., Sigmoid or ReLU), and  $\{\alpha_i, w_i, b_i\}_{i=1}^n$  are the learnable parameters.

By Universal Approximation Theorem (UAT) (Hornik, Stinchcombe, and White 1989), such a network satisfies:

$$\epsilon(x) = \sup_x |\mathcal{F}(x) - g(x)| = \sup_x |\hat{y}(x) - S| < \epsilon, \quad (8)$$

The uniform approximation also implies the expected error:

$$E[\epsilon(x)^2] = \inf_{\theta} E_x \left[ (\hat{y}(x) - S)^2 \right] < \epsilon^2, \quad (9)$$



Category	Level Texts	KonIQ
Common	excellent / good / fair / poor / bad	0.932 / 0.938
Reverse	bad / poor / fair / good / excellent	0.922 / 0.931
Number	five / four / three / two / one	0.924 / 0.934
Random	apple / fog / tree / bag / car	0.925 / 0.936
<b>Ours</b>	<b>&lt;s5&gt;/ &lt;s4&gt;/ &lt;s3&gt;/ &lt;s2&gt;/ &lt;s1&gt;</b>	<b>0.944 / 0.942</b>

Table 1: Comparison of different token settings on KonIQ. ‘‘Ours’’ replaces level tokens with score tokens, while other settings remain consistent with Q-Align. (PLCC / SRCC)

As  $\epsilon \rightarrow 0$ , the bound vanishes. We obtain  $E[\epsilon(x)^2] = 0$ , indicating that with sufficiently many model parameters, the theoretical error of regression-based methods can be reduced to zero, whereas the error of previous token-based methods remains strictly positive. As for RealQA (Li et al. 2025a), it directly outputs scores with a maximum valid digit of two, which inherently introduces restoration error.

This result guarantees that, with sufficiently rich embeddings  $x$  and enough hidden units  $\theta$ , an MLP can theoretically approximate the ground-truth MOS function  $g(x)$  arbitrarily well. This continuous mapping naturally offers finer granularity and better capacity to represent real-valued scores.

### 3.2 Semantic Confusion

Token-based IQA methods often directly use existing vocabulary tokens to represent quality levels. However, this introduces a phenomenon we refer to as semantic confusion, which occurs **Pre-tuning** and **Post-tuning**. It fundamentally limits the effectiveness and transferability of such methods.

**Pre-Tuning Confusion.** This arises because MLLMs already associate certain tokens with inherent semantic meanings. For example, words such as ‘‘good’’ or ‘‘bad’’, beyond representing quality, are strongly associated with general sentiment and weather conditions. When these tokens are only reused to represent discrete quality levels in IQA tasks, the pre-trained semantics can interfere with accurate quality judgment. As shown in Tab.1, using level tokens from Q-Align (Common) underperforms our method due to prior meanings unrelated to visual quality. Moreover, reversing semantic alignment (e.g., mapping ‘‘good’’ to ‘‘poor’’) significantly degrades performance. Both highlight the impact of pre-trained semantics. Similarly, using numeric or random tokens can also introduce noise into the learning process.

**Post-Tuning Confusion.** This refers to the fact that fine-tuning quality-related tokens for IQA can overwrite their semantic integrity and eventually disrupt their use in other tasks. In Fig.3, we show that after Q-Align training, the model’s ability to assess even related tasks like text-to-image (T2I) alignment using these tokens is noticeably impaired. The model tends to overgenerate the IQA-specific sentence pattern (e.g., “The xxx is...”), and the output probability of all level tokens increases sharply, even when unrelated to the given prompt. To mitigate both forms of semantic confusion, we introduce IQA-specific tokens that do not carry pre-existing semantic interference. As evidenced in Tab.1 and Fig.3, our proposed tokens not only achieve better performance than level tokens in the IQA task, but also preserve the model’s capabilities in other tasks.

## 4 Error Remedy: Q-Scorer

To alleviate the errors discussed in Sec. 3, we propose our simple yet effective solution, Q-Scorer.

### 4.1 Model Architecture

**Visual Encoder and Abstractor.** As illustrated in Fig.4, we adopt mPLUG-Owl2 (Ye et al. 2024) as the base architecture to construct our model. We inherit both the visual encoder and the visual abstractor modules from mPLUG-Owl2 to process visual information. Specifically, the input image is first encoded into 1024 visual tokens via the visual encoder. Then, a visual abstractor is applied to compress these tokens into a more compact set of 64 tokens. Finally, reduced visual tokens are fused with prompt-generated text tokens and fed jointly into the LLM for response generation.

**IQA-Specific Score Token.** Although VideoScore (He et al. 2024) employs a regression module, experimental results show that due to the limited representation capacity of the embedding  $x$ , its performance is inferior even compared to token-based methods that suffer from errors. To address the aforementioned issues of semantic confusion and insufficient embedding expressiveness in regression-based IQA scoring, we introduce a set of IQA-specific score tokens directly into the language model’s vocabulary.

While adding a special token has been effective in many MLLM-based downstream tasks, using a single score token  $\{score\}$  often leads to highly similar embeddings across different inputs, because the model is trained to consistently produce the same token regardless of input variations. This makes it difficult for a shallow MLP to regress accurate scores from the similar embeddings. To mitigate this, we define a group of discrete tokens  $\{score1, score2, score3, score4, score5\}$ , each corresponding to a specific score interval, which mirrors the approach in Q-Align. During training, the model learns to select the appropriate score token  $\langle score_x \rangle$  based on the quality level, and a lightweight MLP regressor predicts a fine-grained score within that interval by this score token. This approach reduces semantic interference from existing vocabulary and enhances the model’s ability to distinguish between fine-level quality differences, enabling more accurate score prediction.

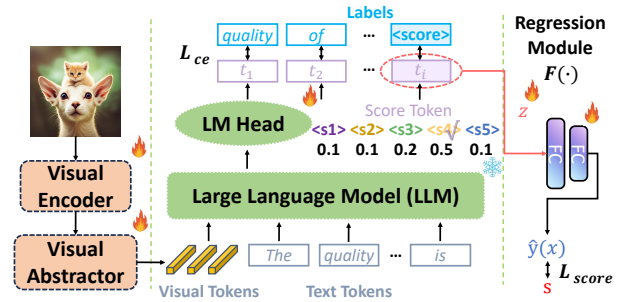


Figure 4: Overview of Q-Scorer. It uses  $\mathcal{L}_{ce}$  learn to output an interval-specific score token. The token’s embedding is then passed to an MLP to regress the continuous quality score, optimized with  $\mathcal{L}_{score}$  to preserve the lossless MOS.

**Regression Module.** Motivated by the theoretical conversion errors discussed earlier, we design a lightweight regression module that directly predicts image quality scores from score token embeddings. Specifically, we extract the embedding of the corresponding IQA-specific score token, which encodes both visual and quality-related semantics, and feed it into an MLP with a  $4096 \rightarrow 2048 \rightarrow 1024 \rightarrow 1$  architecture to produce a scalar quality prediction. This direct regression strategy eliminates the need for token-to-score conversions in token-based methods, which are often lossy and irreversible, effectively narrowing the gap between predicted scores and human perception.

### 4.2 Training Loss

We use the standard cross-entropy loss  $\mathcal{L}_{ce}$  to train the model to predict the correct score token. Following typical LLM training (Touvron et al. 2023), we compute the next-token prediction loss as:

$$\mathcal{L}_{ce} = - \sum_{i=1}^T \log P(t_i | t_1, t_2, \dots, t_{i-1}) \quad (10)$$

where  $T$  is the sequence length,  $t_i$  is the  $i$ -th token, and  $P(t_i | t_1, \dots, t_{i-1})$  denotes the probability of predicting token  $t_i$  given the preceding tokens.

In addition, we extract the embedding  $z$  of the predicted score token  $\langle score_x \rangle$  and use an MLP  $F(\cdot)$  to predict the final score. We apply a MSE loss between the predicted score and ground-truth MOS:

$$\mathcal{L}_{score} = (F(z) - S)^2 \quad (11)$$

This dual-loss design enables continuous MOS supervision to be seamlessly embedded into discrete score tokens, achieving lossless label within the MLLM framework.

### 4.3 Conversation Formats

We define a concise conversation format tailored for the IQA task. Let the image token be  $\langle img \rangle$  and the predicted score token be  $\langle score_x \rangle$ . The dialogue format is as follows:

```
#User: <img> How would you rate the quality of this image?
#Assistant: The quality of this image is <score_x>.
```

Category	Methods	KonIQ	SPAQ <sup>cr</sup>	KADID <sup>cr</sup>	LIVE-Wild <sup>cr</sup>	AGIQA-3K <sup>cr</sup>	CSIQ <sup>cr</sup>
Traditional	NIQE	0.533 / 0.530	0.679 / 0.664	0.468 / 0.405	0.493 / 0.449	0.560 / 0.533	0.718 / 0.628
	BRISQUE	0.225 / 0.226	0.490 / 0.406	0.490 / 0.406	0.361 / 0.313	0.541 / 0.497	0.740 / 0.556
Learning-based	NIMA	0.896 / 0.859	0.838 / 0.856	0.532 / 0.535	0.814 / 0.771	0.715 / 0.654	0.695 / 0.649
	HyperIQA	0.917 / 0.906	0.791 / 0.788	0.506 / 0.468	0.772 / 0.749	0.702 / 0.640	0.752 / 0.717
	DBCNN	0.884 / 0.875	0.812 / 0.806	0.497 / 0.484	0.773 / 0.755	0.730 / 0.641	0.586 / 0.572
	MUSIQ	0.924 / 0.929	0.868 / 0.863	0.575 / 0.556	0.789 / 0.830	0.722 / 0.630	0.771 / 0.710
	CLIP-IQA+	0.909 / 0.895	0.866 / 0.864	0.653 / 0.654	0.832 / 0.805	0.736 / 0.685	0.772 / 0.719
	ManIQA	0.849 / 0.834	0.768 / 0.758	0.499 / 0.465	0.849 / 0.832	0.723 / 0.636	0.623 / 0.627
MLLM-based	Compare2Score	0.923 / 0.910	0.867 / 0.860	0.500 / 0.453	0.786 / 0.772	0.777 / 0.671	0.735 / 0.705
	Q-Align	0.941 / 0.940	0.886 / 0.887	0.674 / 0.684	0.853 / 0.860	0.772 / 0.735	0.785 / 0.737
	DeQA-Score	0.953 / 0.941	0.895 / 0.896	<b>0.694 / 0.687</b>	0.892 / 0.879	0.809 / 0.729	0.787 / 0.744
	Q-Align (LoRA)	0.932 / 0.938	0.874 / 0.886	0.624 / 0.632	0.858 / 0.859	0.806 / 0.735	0.772 / 0.730
	<b>Ours (5)</b>	0.959 / 0.948	0.898 / 0.898	0.676 / 0.671	0.889 / 0.870	<b>0.821 / 0.736</b>	<b>0.796 / 0.746</b>
	<b>Ours (1)</b>	<b>0.960 / 0.950</b>	<b>0.900 / 0.899</b>	0.660 / 0.645	<b>0.903 / 0.888</b>	0.811 / 0.722	0.795 / 0.733

Table 2: Performance comparison across multiple IQA datasets. “cr” denotes cross-dataset evaluation. “Ours(1)” refers to predicting the same score token regardless of input. “Ours(5)” denotes our base method. (PLCC / SRCC)

	Method	KonIQ	SPAQ	KADID
1	Q-Align	0.943 / 0.940	0.933 / 0.931	0.692 / 0.708
	DeQA-Score	0.953 / 0.943	0.936 / 0.933	<b>0.724 / 0.719</b>
	<b>Ours (5)</b>	0.954 / 0.941	0.936 / 0.932	0.672 / 0.661
	<b>Ours (1)</b>	0.950 / 0.936	0.934 / 0.931	0.646 / 0.628
	<b>Ours (F-loss)</b>	<b>0.961 / 0.950</b>	<b>0.937 / 0.934</b>	0.702 / 0.689
2	Q-Align	0.945 / 0.938	0.933 / 0.931	0.935 / 0.934
	DeQA-Score	0.957 / 0.944	<b>0.938 / 0.934</b>	0.955 / 0.953
	<b>Ours (5)</b>	0.953 / 0.939	0.936 / 0.932	0.961 / 0.958
	<b>Ours (1)</b>	0.952 / 0.938	0.934 / 0.930	0.959 / 0.956
	<b>Ours (F-loss)</b>	<b>0.961 / 0.951</b>	<b>0.938 / 0.935</b>	<b>0.963 / 0.961</b>

Table 3: Comparison of multi-dataset training results across different IQA datasets. “1” denotes training on KonIQ and SPAQ; “2” denotes training on KonIQ, SPAQ and KADID. “F-loss” denotes Fidelity loss. (PLCC / SRCC)

## 5 Experiments

In this section, to demonstrate the effectiveness and generalization ability of Q-Scorer, we conduct experiments on both single-dataset settings (Sec. 5.4) and multi-dataset settings (Sec. 5.5). In addition, we carry out ablation studies to verify the effectiveness of the two key components in Q-Scorer (Sec. 5.6). To further validate the efficiency of the score token embedding, we perform comprehensive comparisons with the VideoScore method from multiple perspectives (Sec. 5.7). Finally, we explore the integration of Q-Scorer with other methods and observe its compatibility and extensibility (Sec. 5.8).

### 5.1 Experimental Settings

We fine-tune our model using the LoRA strategy (Hu et al. 2022), starting from the pre-trained mPLUG-Owl2 (Ye et al. 2024) weights. The visual encoder is initialized with CLIP-pretrained ViT-L (Radford et al. 2021), and the language backbone is LLaMA-2-7B (Touvron et al. 2023). Training is performed for 3 epochs with a batch size of 16 using the AdamW optimizer (Loshchilov and Hutter 2017), an initial learning rate of  $2e-5$ , and cosine decay scheduling. All ex-

periments run on 2 NVIDIA RTX A100 GPUs, and training on KonIQ completes in about 30 minutes.

### 5.2 Datasets and Baselines

We train our model on three IQA datasets: KonIQ (Hosu et al. 2020), SPAQ (Fang et al. 2020), and KADID (Lin, Hosu, and Saupe 2019), following the setup in Q-Align. To assess generalization, we evaluate on four unseen datasets: LIVE-Wild (Ghadiyaram and Bovik 2015), AGIQA-3K (Li et al. 2023), and CSIQ (Larson and Chandler 2010). The MOSs of these datasets are normalized to  $[1, 5]$ .

We primarily compare against Q-Align (Wu et al. 2023b), its LoRA-adapted variant, and its improved method DeQA-Score (You et al. 2025). Besides, we also report results from representative traditional and deep learning-based methods, including handcrafted metrics (NIQE (Mittal, Soundararajan, and Bovik 2012), BRISQUE (Mittal, Moorthy, and Bovik 2012)) and learning-based models (NIMA (Talebi and Milanfar 2018), HyperIQA (Su et al. 2020), DBCNN (Zhang et al. 2018), MUSIQ (Ke et al. 2021), CLIP-IQA+ (Wang, Chan, and Loy 2023), ManIQA (Yang et al. 2022), and Compare2Score (Zhu et al. 2024)).

### 5.3 Metrics

We use the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SRCC) to evaluate score regression performance. PLCC assesses the linear correlation between predicted scores and ground-truth MOSs, while SRCC evaluates the consistency of their ranking order.

### 5.4 Single-Dataset Training Results

We first train our model solely on the KonIQ dataset and evaluate its generalization on all other datasets, as summarized in Tab.2. To fairly assess LoRA-based fine-tuning, we include a LoRA-adapted Q-Align as a baseline. Our method consistently outperforms baselines in score regression across all datasets except KADID, demonstrating

Type	Ablation	KonIQ	SPAQ	KADID	LIVE-Wild	AGIQA-3K	CSIQ
Regression module	without	0.944 / 0.942	0.887 / 0.893	0.629 / 0.622	0.871 / 0.855	0.816 / 0.735	0.755 / 0.717
	[512, 256]	0.957 / 0.948	0.895 / 0.893	0.643 / 0.629	0.883 / 0.869	0.796 / 0.706	0.782 / 0.731
	[2048, 1024, 512, 256]	0.938 / 0.927	0.869 / 0.866	0.586 / 0.576	0.843 / 0.844	0.810 / 0.689	0.743 / 0.697
Score token	level token (mix)	0.948 / 0.940	0.896 / 0.896	0.653 / 0.635	0.873 / 0.853	0.772 / 0.685	0.759 / 0.684
	level token (bare)	0.946 / 0.935	0.887 / 0.883	0.607 / 0.592	0.862 / 0.854	0.817 / 0.707	0.774 / 0.727
	<b>Our (base)</b>	<b>0.959 / 0.948</b>	<b>0.898 / 0.898</b>	<b>0.676 / 0.671</b>	<b>0.889 / 0.870</b>	<b>0.821 / 0.734</b>	<b>0.796 / 0.732</b>

Table 4: Ablation study results. “without” denotes using weighted sum instead of an MLP. “[\*]” indicates different MLP settings. “mix” uses both level and score tokens; “bare” uses only level token embeddings for regression. (PLCC / SRCC)

strong generalization. On the KonIQ, it achieves state-of-the-art results, improving PLCC by 2% over fully fine-tuned Q-Align, and by 2.9% over its LoRA-adapted variant. The performance drop on KADID likely stems from LoRA’s side effects, as Q-Align with LoRA shows a similar decline. Interestingly, using a single score token outperforms the enhanced token set here, possibly due to the smaller data scale and simpler score distribution of a single dataset, where one token is sufficient for effective mapping.

### 5.5 Multi-Dataset Training Results

We present multi-dataset co-training results in Tab.3, where our method consistently outperforms Q-Align across various dataset combinations. However, when trained on all three datasets simultaneously, our model lags behind the DeQA-Score baseline, which leverages fidelity loss to improve cross-dataset generalization. This is expected, as different IQA datasets often exhibit domain shifts, where the same MOS may represent different perceptual qualities (Zhu et al. 2024). Using a unified set of score tokens and a single MLP head to represent all datasets may cause interference, especially reflected in the KonIQ performance drop. By simply adding fidelity loss, our method achieves state-of-the-art results (see the last row of Tab.3; details in the Appendix). Additionally, relying on a single score token significantly degrades performance in multi-dataset settings, underscoring the necessity of adopting a set of interval-specific score tokens to better model challenging datasets.

### 5.6 Ablation Studies

As shown in Tab.4, we conduct ablation studies to evaluate the impact of the regression module and score token design. Our proposed configuration consistently achieves the best performance. Notably, replacing the MLP with a probability-weighted sum over score tokens results in a significant performance drop, demonstrating that regression-based modeling is not only theoretically optimal but also empirically effective. Moreover, mixing level tokens with score tokens or directly using level token embeddings for regression degrades performance, further highlighting the semantic confusion introduced by level tokens.

### 5.7 Score Token Embedding Efficiency

RealQA (Li et al. 2025a) notes that while existing regression-based method (VideoScore) improves with more training, they still lag behind next-token prediction (NTP)

Method	Epochs	KonIQ
VideoScore	2	0.890 / 0.882
VideoScore	6	0.923 / 0.908
Ours	1	0.929 / 0.919
Ours	2	0.957 / 0.945
Ours (base)	3	<b>0.959 / 0.948</b>
Ours	4	<b>0.959 / 0.949</b>
Ours	5	0.957 / 0.947
Ours	6	0.957 / 0.948

Table 5: Performance comparison under different training epochs and methods. (PLCC / SRCC)

approaches. This is mainly due to the insufficient embedding expressiveness of the last token to model score distributions, leading to slow convergence and degraded performance. As shown in Tab.5, our model alleviates these issues through a dedicated score token design, achieving accurate score prediction within 1 epoch and full convergence in 3–4 epochs.

### 5.8 Combination with Other Methods

Our method can be further improved by integrating existing techniques. To mitigate potential overlap among score tokens, we incorporate the KL divergence loss from DeQA-Score. We also explore enhancements with external modules such as norm-in-norm loss (Peng et al. 2023), hyper network (Su et al. 2020) and ranking loss (Liu, Van De Weijer, and Bagdanov 2017). These combinations yield additional gains on several datasets, highlighting the compatibility and extensibility of our approach (details in the Appendix).

## 6 Conclusions

Our work identifies conversion errors and semantic confusion in existing MLLM-based IQA methods, and introduce Q-Scorer, a simple yet effective framework that leverages IQA-specific score tokens and a lightweight regression module to predict continuous quality scores. Q-Scorer achieves state-of-the-art performance across multiple IQA benchmarks, offering a promising direction for enhancing the quality assessment capabilities of MLLMs.

**Limitations and Future Work.** Most IQA datasets primarily reflect annotators’ personal preferences, so models trained on their diverse distributions of their MOSs may fail to capture how people in general perceive image quality. In future work, we plan to incorporate self-learning mechanisms into MLLMs to improve their generalization beyond dataset-specific biases.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62272460, Beijing Natural Science Foundation under Grant No. 4232037

## References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Bosse, S.; Maniry, D.; Müller, K.-R.; Wiegand, T.; and Samek, W. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1): 206–219.
- Chen, C.; Yang, S.; Wu, H.; Liao, L.; Zhang, Z.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2024a. Q-ground: Image quality grounding with large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 486–495.
- Chen, Z.; Wang, J.; Wang, W.; Xu, S.; Xiong, H.; Zeng, Y.; Guo, J.; Wang, S.; Yuan, C.; Li, B.; et al. 2024b. SEAGULL: No-reference Image Quality Assessment for Regions of Interest via Vision-Language Instruction Tuning. *arXiv preprint arXiv:2411.10161*.
- Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3677–3686.
- Ghadiyaram, D.; and Bovik, A. C. 2015. Live in the wild image quality challenge database. *Online: <http://live.ece.utexas.edu/research/ChallengeDB/index.html> [Mar, 2017]*.
- Han, X.; Yang, S.; Wang, W.; Li, Y.; and Dong, J. 2025. Adaptive Median Smoothing: Adversarial Defense for Unlearned Text-to-Image Diffusion Models at Inference Time. In *Forty-second International Conference on Machine Learning*.
- He, X.; Jiang, D.; Zhang, G.; Ku, M.; Soni, A.; Siu, S.; Chen, H.; Chandra, A.; Jiang, Z.; Arulraj, A.; et al. 2024. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*.
- Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*, 2366–2369. IEEE.
- Hornik, K.; Stinchcombe, M.; and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5): 359–366.
- Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.
- Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1): 011006–011006.
- Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; and Lin, W. 2023. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 6833–6846.
- Li, M.; Wang, R.; Sun, L.; Bai, Y.; and Chu, X. 2025a. Next Token Is Enough: Realistic Image Quality and Aesthetic Scoring with Multimodal Large Language Model. *arXiv:2503.06141*.
- Li, Y.; Yang, S.; Han, X.; Wang, W.; Dong, J.; Lyu, Y.; and Xue, Z. 2025b. Instant Preference Alignment for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2508.17718*.
- Li, Y.; Yang, S.; Wang, W.; and Dong, J. 2025c. Beyond Inserting: Learning Subject Embedding for Semantic-Fidelity Personalized Diffusion Generation. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19401–19411.
- Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 1–3. IEEE.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36: 34892–34916.
- Liu, X.; Van De Weijer, J.; and Bagdanov, A. D. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, 1040–1049.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12): 4695–4708.
- Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.
- Peng, B.; Sun, X.; Wang, C.; Wang, W.; Dong, J.; Sun, Z.; Zhang, R.; Cong, H.; Fu, L.; Wang, H.; et al. 2023. DFGC-VRA: DeepFake Game Competition on Visual Realism Assessment. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 1–9. IEEE.

- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. 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*, 3667–3676.
- Talebi, H.; and Milanfar, P. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8): 3998–4011.
- Tang, Z.; Wang, Z.; Peng, B.; and Dong, J. 2024. CLIP-AGIQA: boosting the performance of ai-generated image quality assessment with clip. In *International Conference on Pattern Recognition*, 48–61. Springer.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2555–2563.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.
- Wang, Z.; Peng, B.; Yang, S.; Tang, Z.; and Dong, J. 2025. HandEval: Taking the First Step Towards Hand Quality Evaluation in Generated Images. *arXiv preprint arXiv:2510.08978*.
- Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; et al. 2023a. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*.
- Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. 2024. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25490–25500.
- Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023b. Q-align: Teaching llms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.
- Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1191–1200.
- Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13040–13051.
- You, Z.; Cai, X.; Gu, J.; Xue, T.; and Dong, C. 2025. Teaching Large Language Models to Regress Accurate Image Quality Scores using Score Distribution. *arXiv preprint arXiv:2501.11561*.
- You, Z.; Gu, J.; Li, Z.; Cai, X.; Zhu, K.; Dong, C.; and Xue, T. 2024a. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*.
- You, Z.; Li, Z.; Gu, J.; Yin, Z.; Xue, T.; and Dong, C. 2024b. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, 259–276. Springer.
- Yu, Z.; Guan, F.; Lu, Y.; Li, X.; and Chen, Z. 2024. Sfiqa: Quality and similarity integration for ai generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6692–6701.
- Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2018. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1): 36–47.
- Zhang, W.; Ma, K.; Zhai, G.; and Yang, X. 2021. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30: 3474–3486.
- Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14071–14081.
- Zheng, H.; Yang, H.; Fu, J.; Zha, Z.-J.; and Luo, J. 2021. Learning conditional knowledge distillation for degraded-reference image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10242–10251.
- Zhou, W.; and Wang, Z. 2022. Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity. In *Proceedings of the 30th ACM International Conference on Multimedia*, 934–942.
- Zhu, H.; Wu, H.; Li, Y.; Zhang, Z.; Chen, B.; Zhu, L.; Fang, Y.; Zhai, G.; Lin, W.; and Wang, S. 2024. Adaptive image quality assessment via teaching large multimodal model to compare. *arXiv preprint arXiv:2405.19298*.