VisualQuality-R1: Reasoning-Induced Image Quality Assessment via Reinforcement Learning to Rank

Tianhe Wu[1,2], **Jian Zou**[1], **Jie Liang**[2], **Lei Zhang**[2,3],*and **Kede Ma**[1]*

[1]City University of Hong Kong
[2]OPPO Research Institute
[3]The Hong Kong Polytechnic University
{tianhewu, jianzou5, kede.ma}@cityu.edu.hk, liang27jie@gmail.com,
cslzhang@comp.polyu.edu.hk
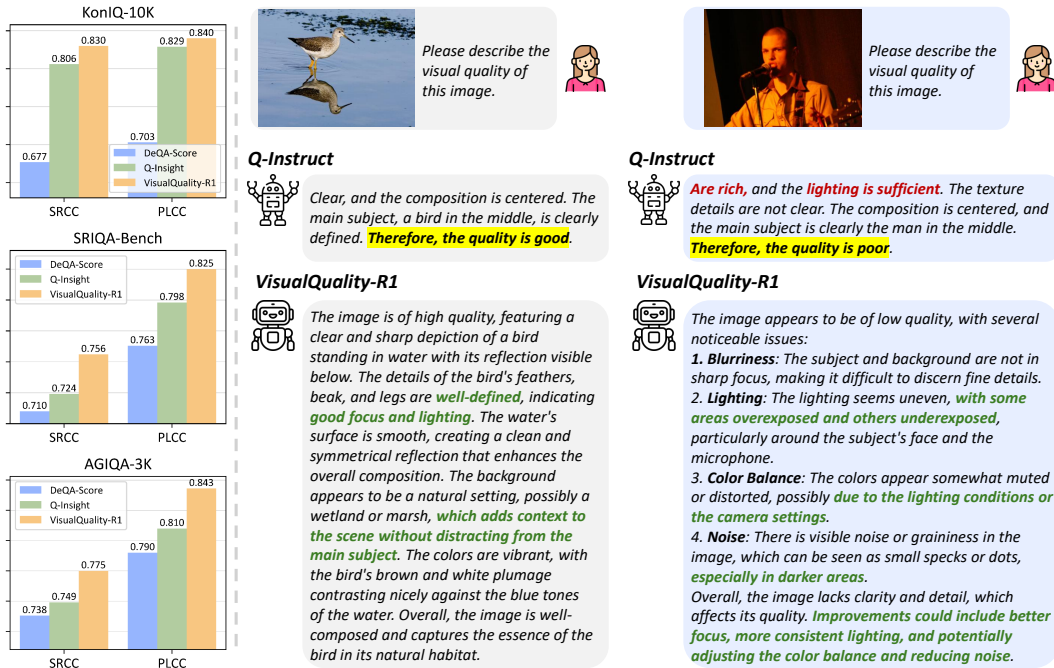https://github.com/TianheWu/VisualQuality-R1

Figure 1: VisualQuality-R1 excels at image quality scoring, while generating contextually rich, human-aligned quality descriptions.

Abstract

DeepSeek-R1 has demonstrated remarkable effectiveness in incentivizing reasoning and generalization capabilities of large language models (LLMs) through reinforcement learning. Nevertheless, the potential of reasoning-induced computational modeling has not been thoroughly explored in the context of image quality assessment (IQA), a task critically dependent on visual reasoning. In this paper, we introduce VisualQuality-R1, a reasoning-induced no-reference IQA (NR-IQA) model, and we train it with reinforcement learning to rank, a learning algorithm tailored to the intrinsically relative nature of visual quality. Specifically, for a pair of images, we employ group relative policy optimization to generate multiple quality scores for each image. These estimates are then used to compute comparative

---

*Corresponding authors.

probabilities of one image having higher quality than the other under the Thurstone model. Rewards for each quality estimate are defined using continuous fidelity measures rather than discretized binary labels. Extensive experiments show that the proposed VisualQuality-R1 consistently outperforms discriminative deep learning-based NR-IQA models as well as a recent reasoning-induced quality regression method. Moreover, VisualQuality-R1 is capable of generating contextually rich, human-aligned quality descriptions, and supports multi-dataset training without requiring perceptual scale realignment. These features make VisualQuality-R1 especially well-suited for reliably measuring progress in a wide range of image processing tasks like super-resolution and image generation.

1 Introduction

Image quality assessment (IQA) aims to quantify the visual quality of digital images consistent with human perceptual judgments. Commonly, IQA models are classified into full-reference (FR) and no-reference (NR) approaches [46], depending on the availability of pristine-quality reference images. In this paper, we focus on NR-IQA due to its practical relevance in real-world scenarios where reference images are unavailable. Over the decades, NR-IQA has evolved from knowledge-driven [32, 12] to data-driven approaches [29, 19, 52], and shifted from regression-based to ranking-based [56, 57] techniques. Nevertheless, achieving strong model generalization (*e.g.*, generalization to unseen image distortions) remains a significant, unresolved challenge, driving recent research toward multi-dataset training [6], active fine-tuning [43], and continual model adaptation [55].

The rapid advancement of vision-language models (VLMs) offers promising avenues for enhancing NR-IQA generalization by contextualizing it into broader vision tasks [50]. VLMs can effectively integrate multi-modal information, enabling understanding of both low-level image distortions, such as noise and blur, and high-level perceptual attributes, such as aesthetics and content semantics. This multi-modal semantic contextualization enables VLMs to articulate nuanced quality descriptions with stronger generalization.

However, current NR-IQA methods mainly leverage VLMs through supervised fine-tuning (SFT), which face several critical limitations [48, 54]. First, constructing informative quality descriptions demands extensive human effort, rendering the annotation process labor-intensive and expensive[2]. Second, models trained via SFT often overfit to the biases and idiosyncrasies present in training data, and may unintentionally encounter catastrophic forgetting of acquired knowledge during pre-training. Third, SFT typically yields overly rigid and templated outputs (see Fig. 1) that may be less useful.

Reinforcement learning (RL) has recently emerged as a powerful alternative, enhancing the reasoning capabilities of LLMs, while aligning their responses with human preferences [34, 13]. In particular, DeepSeek-R1 [13] demonstrates the effectiveness of RL in promoting generalization by encouraging automated exploration of plausible reasoning paths and employing rule-based rewards to prevent reward hacking [36]. However, a direct adaptation of RL techniques to NR-IQA, as exemplified by the recent Q-Insight model [21], has been limited by its reliance on dataset-specific reward design and additional distortion-type classification. These constraints stem from its treatment of visual quality as an *absolute* perceptual quantity, thereby framing NR-IQA simplistically as a regression task.

In this paper, we introduce **VisualQuality-R1**, a reasoning-induced NR-IQA model, and we train it via reinforcement learning to rank (RL2R), a learning algorithm explicitly designed to capture the inherently *relative* nature of visual quality. Specifically, we employ group relative policy optimization (GRPO) [35] to derive multiple quality scores for each image in a pair. We then compute comparative probabilities between images using the Thurstone model [40] by assessing the difference between the mean quality score of one image and individual quality scores of another, normalized by their added sample variance. Unlike previous methods, we define the reward function using the continuous fidelity measure [41], which provides precise guidance to facilitate quality ranking. Extensive experiments confirm that VisualQuality-R1 effectively assesses visual quality across a diverse range of distortion scenarios, outperforming discriminative deep learning-based NR-IQA models as well as a recent reasoning-induced quality regression method. Moreover, VisualQuality-R1 generates contextually

---

[2]Utilizing state-of-the-art proprietary VLMs such as GPT-4o [16] for automated annotation suffers from similar scalability challenges due to high computational costs and financial burdens.

rich, human-aligned quality descriptions (see Fig. 1), which can be leveraged to provide targeted feedback for downstream image processing algorithms, and support fine-grained quality control in digital photography pipelines. Additionally, we demonstrate that VisualQuality-R1 remains effective across multi-dataset training scenarios without requiring perceptual scale realignment. As a result, VisualQuality-R1 represents a substantial advancement in NR-IQA, and establishes new benchmarks for reasoning and generalization performance.

2 Related Work

This section provides a structured review of related NR-IQA models, emphasizing recent advancements, particularly those leveraging VLMs.

Regression-based Models NR-IQA models primarily employed regression-based approaches, wherein image quality was treated as an absolute perceptual quantity directly estimated from extracted "quality-aware" features. Initially, features were handcrafted based on natural scene statistics [32, 33], degradation-specific characteristics [45, 47, 24], and perceptual models inspired by the human visual system [44]. However, these NR-IQA methods were limited by the representational capacity of handcrafted features. Later, deep learning-based regression models emerged as the dominant paradigm, using end-to-end trainable neural networks to directly predict quality scores (or, in some cases, quality distributions) [18, 29, 3, 38, 19, 52]. These models typically utilize standard regression losses such as the mean squared error and mean absolute error, or statistical distances such as the earth mover's distance [38] and Kullback–Leibler (KL) divergence [53]. Regression-based models often struggle with generalization issues, and require labor-intensive perceptual scale realignment [30] when training on multiple IQA datasets.

Ranking-based Models To address these shortcomings, ranking-based NR-IQA models were introduced, modeling visual quality as an intrinsically relative perceptual quantity. Gao *et al.* [10] pioneered the concept of quality ranking in NR-IQA, although their initial implementation relied on predefined anchor images and was not end-to-end optimized. Ma *et al.* [28] adapted RankNet [4] to NR-IQA by training (though not fully end-to-end) on quality-discriminable image pairs. Their subsequent work established the first end-to-end ranking-based NR-IQA method grounded in the Thurstone model [40]. Nonetheless, their approach suffers from scaling ambiguity during variance estimation. Zhang *et al.* [56] incorporated a hinge loss to regularize variance estimation, yet the scaling ambiguity persisted. Their study demonstrated the superiority of the fidelity loss [41] over the conventional cross-entropy loss in ranking-based NR-IQA. Subsequent research has adopted a simpler approach by fixing the variance parameter to one (corresponding to the Thurstone Case V model), facilitating active fine-tuning of NR-IQA models on challenging examples [43, 42] and allowing continual adaptation to novel distortion scenarios [55]. Other losses that enable quality ranking include the margin ranking loss [25], differentiable approximations of Spearman's rank correlation coefficient (SRCC) [2], Pearson linear correlation coefficient (PLCC) [51], and statistical distances between permutation probabilities [5, 37, 17].

VLM-based Models The integration of VLMs into NR-IQA has recently gained traction, particularly due to their proficiency in capturing semantic context through multi-modal representation learning. Early attempts include multitask adaptation of CLIP [57], as well as SFT-based methods like Q-Align [49], Compare2Score [58], DepictQA [54], and DeQA-Score [53], which trained VLMs to generate either quality scores, distributions, or descriptions. Closest to ours, Q-Insight [21] explored reasoning-induced quality regression through RL. However, Q-Insight struggles with dataset-specific reward calibration, generalization to novel distortion scenarios, and the added complexity of auxiliary distortion-type classification. In contrast, our VisualQuality-R1 redefines the use of VLMs in NR-IQA by shifting from absolute regression to relative ranking, leading to enhanced generalization across distortion scenarios with better quality justifications.

3 Reasoning-Induced NR-IQA

To harness both the powerful reasoning-inducing capabilities of RL and the intrinsically relative nature of visual quality, we propose an NR-IQA model—VisualQuality-R1—and an RL2R method
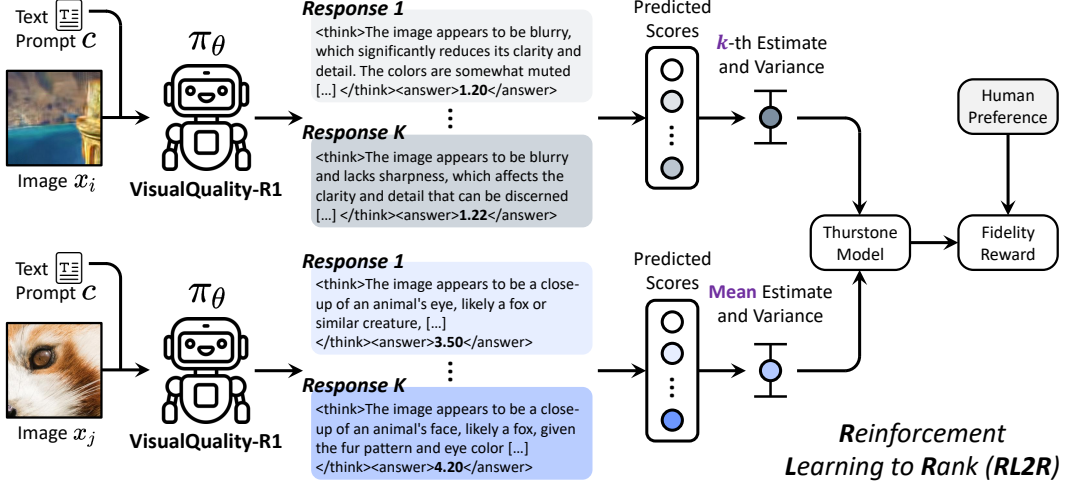
Figure 2: System diagram of the proposed VisualQuality-R1 trained via RL2R. Given an image pair (x_i, x_j) with a shared text prompt c, VisualQuality-R1 generates K responses. Following GRPO [35], each response includes a detailed reasoning process and a predicted quality score. To assess relative visual quality, we calculate the asymmetric comparative probability that image x_i is perceived better than x_j under the Thurstone model [40]. This involves subtracting the mean predicted score of x_j from the k-th score of x_i, standardized by their added sample variance. A fidelity reward is derived from human preference, providing continuous supervisory signals for policy optimization.

of training it that seamlessly integrates the Thurstone model within GRPO. Fig. 2 shows the system diagram of VisualQuality-R1 trained via RL2R.

3.1 VisualQuality-R1 via RL2R

Given a text prompt c and an image x, our goal is to fine-tune a pre-trained VLM, with policy $\pi_\theta(\cdot|c, x)$, to produce a scalar quality score in the range of $[1, 5]$, following a step-by-step reasoning process, encapsulated within specially designated tags for explicit instruction and enhanced interpretability. The complete structured text prompt is provided in Table 1.

More specifically, for a training batch of images $\{x_1, x_2, \ldots, x_B\}$, where B is the minibatch size, we apply GRPO to generate K quality predictions for x_i, $q(x_i) = [q_1(x_i), q_2(x_i), \ldots, q_K(x_i)]^\mathsf{T}$. This output naturally encodes predictive uncertainty, which is crucial for making reliable relative quality ranking. Under the Thurstone model [40], the visual quality of an image is assumed to follow a Gaussian distribution. Thus, we are able to compute the asymmetric comparative probability for each of the $B \times (B-1)$ ordered image pairs by subtracting the mean quality score of x_j from the k-th quality score of x_i, standardized by their added sample variance:

$$p_k(x_i, x_j) = \Phi\left(\frac{q_k(x_i) - \mu(q(x_j))}{\sqrt{\sigma^2(q(x_i)) + \sigma^2(q(x_j)) + \gamma}}\right), \quad \text{for } i \neq j, \tag{1}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. $\mu(q(x_j))$ and $\sigma^2(q(x_j))$ represent the mean and variance of the quality predictions for x_j, respectively. γ is a small positive constant to avoid any potential division by zero. Compared to previous ranking-based NR-IQA models that fix the variance parameter in Eq. (1) to one, we explicitly leverage sample variances derived from GRPO. This gives us an opportunity to dynamically accommodate predictive uncertainty for different images. Meanwhile, using the sample mean for quality comparison stabilizes the asymmetric probability estimate and the subsequent reward calculation by appropriately penalizing outlier predictions.

The true preference $p(x_i, x_j)$ is derived from human mean opinion scores (MOSs):

$$p(x, y) = \begin{cases} 1 & \text{if } \mathrm{MOS}(x) > \mathrm{MOS}(y) \\ 0.5 & \text{if } \mathrm{MOS}(x) = \mathrm{MOS}(y) \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

4

Table 1: Structured text prompt used in VisualQuality-R1.

You are doing the image quality assessment task. Here is the question:
Rate the overall image visual quality. The rating should be a float between 1 and 5, with 1 representing very poor quality and 5 representing excellent quality. First output the thinking process in `<think>` `</think>` tags and then output the final answer with only one score in `<answer>` `</answer>` tags.

An important aspect of our RL2R algorithm is that we define the reward function $r_k(x_i)$ for each quality estimate $q_k(x_i)$ as the fidelity measure [41]—a continuous analogue of the discretized binary reward [13, 21], averaged across all $B - 1$ image pairs:

$$r_k(x_i) = \frac{1}{B-1} \sum_{j \neq i} \left(\sqrt{p(x_i, x_j) p_k(x_i, x_j)} + \sqrt{(1 - p(x_i, x_j))(1 - p_k(x_i, x_j))} \right). \quad (3)$$

This continuous reward feedback provides precise guidance during RL2R by capturing subtle distinctions in quality ranking, thus boosting generalization across diverse distortion scenarios. We collect K fidelity rewards for x_i into the vector $r(x_i) = [r_1(x_i), r_2(x_i), \ldots, r_K(x_i)]^\mathsf{T}$, and compute the relative advantage $a_k(x_i)$ by standardizing rewards within group:

$$a_k(x_i) = \frac{r_k(x_i) - \mu(r(x_i))}{\sigma(r(x_i))}. \quad (4)$$

The final policy update of $\pi_\theta(\cdot|c, x_i)$ is guided by the regularized objective in GRPO:

$$\ell(\theta) = \frac{1}{BK} \sum_{i=1}^{B} \sum_{k=1}^{K} \left(\min \left(\frac{\pi_\theta(o_k|c, x_i)}{\pi_{\theta_{\text{old}}}(o_k|c, x_i)} a_k(x_i), \text{clip} \left(\frac{\pi_\theta(o_k|c, x_i)}{\pi_{\theta_{\text{old}}}(o_k|c, x_i)}, 1 - \epsilon, 1 + \epsilon \right) a_k(x_i) \right) \right.$$
$$\left. - \beta \, D_{\text{KL}} \left(\pi_\theta(o_k|c, x_i) \| \pi_{\text{ref}}(o_k|c, x_i) \right) \right).$$
$$(5)$$

Here, $\pi_{\theta_{\text{ref}}}(\cdot|c, x_i)$ denotes the stable reference policy obtained after VLM pre-training, and $\pi_{\theta_{\text{old}}}(\cdot|c, x_i)$ is the policy from the previous RL2R training epoch, from which we sample K reasoning trajectories $o = \{o_k\}_{k=1}^{K}$. The second KL divergence term is approximated by

$$D_{\text{KL}} \left(\pi_\theta(o_k|c, x_i) \| \pi_{\text{ref}}(o_k|c, x_i) \right) = \frac{\pi_{\text{ref}}(o_k|c, x_i)}{\pi_\theta(o_k|c, x_i)} - \log \frac{\pi_{\text{ref}}(o_k|c, x_i)}{\pi_\theta(o_k|c, x_i)} - 1, \quad (6)$$

incorporated to ensure that the updated policy $\pi_\theta(\cdot|c, x_i)$ does not deviate excessively from $\pi_{\text{ref}}(\cdot|c, x_i)$. ϵ is the clipping threshold to prevent large and potentially destabilizing updates to the policy. The coefficient β serves as a balancing factor between the reward-weighted likelihood term and the KL regularization term.

We conclude this section by highlighting the key strengths of the resulting VisualQuality-R1. First, VisualQuality-R1 inherits all the advantages of ranking-based NR-IQA models, enabling effective multi-dataset training, active fine-tuning, and continual model adaptation without requiring perceptual scale realignment [30], a feature notably absent in regression-based NR-IQA approaches. Second, trained via RL2R, VisualQuality-R1 mitigates the scalability and overfitting issues inherent in SFT-based models. Third, VisualQuality-R1 promises to both improve model generalizability and furnish contextually rich textual justifications alongside numerical quality scores, thereby boosting its practical relevance in real-world IQA applications.

4 Experiments

To validate VisualQuality-R1, we conduct comprehensive experiments across diverse distortion scenarios, ablation studies on key design components, and in-depth analysis of model behaviors.

4.1 Experimental Setups

Competing Models and Training Details Competing methods encompass three categories: 1) handcrafted models: NIQE[33] and BRISQUE [32]; 2) discriminative deep-learning-based models:

Table 2: SRCC and PLCC results of NR-IQA models trained on KADID-10K. Exceptions include Q-Insight[†] and VisualQuality-R1[†], which use a combined training set (KADID-10K and SPAQ). Top two results are highlighted in **bold** and <u>underline</u>, respectively.

Method	Imaging-Related Distortion				Processing-Related Distortion				Avg
	BID	CLIVE	KonIQ	SPAQ	De-blurring	Super-Res.	De-hazing	Image Gen.	
SRCC									
Handcrafted									
NIQE [33]	0.515	0.450	0.421	0.676	0.360	0.557	0.343	0.533	0.482
BRISQUE [32]	0.522	0.314	0.385	0.614	0.389	0.482	0.242	0.497	0.431
Discriminative Deep-Learning-based									
UNIQUE [56]	0.412	0.470	0.649	0.751	0.669	0.649	0.577	0.608	0.598
MUSIQ [19]	0.327	0.284	0.473	0.720	0.656	0.404	0.458	0.494	0.477
MANIQA [52]	0.420	0.487	0.213	0.745	0.726	0.263	0.608	0.422	0.486
VLM-based									
LIQE [57]	0.677	0.719	0.684	0.815	0.797	0.743	**0.646**	0.653	0.717
Q-Align [49]	0.576	0.554	0.573	0.767	0.761	0.684	0.455	0.682	0.632
DeQA-Score [53]	0.702	0.743	0.677	0.852	0.785	0.710	<u>0.643</u>	0.738	0.731
Qwen2.5-VL-7B [1]	0.711	0.733	0.754	0.848	0.820	0.603	0.458	0.735	0.708
Q-Insight [21]	0.784	0.761	0.806	0.872	0.831	0.724	0.601	0.749	0.766
Q-Insight[†]	<u>0.806</u>	<u>0.804</u>	0.812	<u>0.907</u>	**0.846**	0.700	0.539	0.657	0.759
VisualQuality-R1	0.790	0.750	<u>0.830</u>	0.875	0.838	**0.756**	0.598	**0.775**	<u>0.777</u>
VisualQuality-R1[†]	**0.811**	**0.811**	**0.855**	**0.913**	<u>0.845</u>	<u>0.752</u>	0.588	<u>0.754</u>	**0.791**
PLCC									
Handcrafted									
NIQE [33]	0.527	0.494	0.439	0.683	0.376	0.587	0.482	0.560	0.519
BRISQUE [32]	0.528	0.362	0.400	0.624	0.380	0.556	0.468	0.541	0.482
Discriminative Deep-Learning-based									
UNIQUE [56]	0.385	0.472	0.590	0.708	0.654	0.668	0.578	0.581	0.580
MUSIQ [19]	0.280	0.325	0.435	0.666	0.563	0.441	0.455	0.434	0.450
MANIQA [52]	0.512	0.571	0.257	0.753	0.728	0.243	0.663	0.448	0.522
VLM-based									
LIQE [57]	0.680	0.726	0.652	0.814	0.712	0.775	0.661	0.653	0.709
Q-Align [49]	0.651	0.643	0.612	0.779	0.802	0.713	0.525	0.705	0.679
DeQA-Score [53]	0.743	0.795	0.703	0.858	0.838	0.763	**0.688**	0.790	0.772
Qwen2.5-VL-7B [1]	0.725	0.760	0.810	0.854	0.852	0.653	0.553	0.810	0.752
Q-Insight [21]	0.796	0.795	0.829	0.872	0.857	0.798	0.669	0.810	0.803
Q-Insight[†]	<u>0.818</u>	<u>0.837</u>	0.809	<u>0.912</u>	0.861	0.779	0.626	0.705	0.793
VisualQuality-R1	0.806	0.794	<u>0.840</u>	0.878	<u>0.872</u>	**0.825**	0.651	**0.843**	<u>0.814</u>
VisualQuality-R1[†]	**0.820**	**0.844**	**0.870**	**0.917**	**0.879**	<u>0.824</u>	<u>0.674</u>	<u>0.820</u>	**0.831**

UNIQUE [56], MUSIQ [19], and MANIQA [52]; 3) VLM-based models: LIQE [57], Q-Align [49], DeQA-Score [53], Q-Insight [21], as well as the pre-trained Qwen2.5-VL-7B [1] baseline.

We fine-tune Qwen2.5-VL-7B [1] as the backbone for VisualQuality-R1 using GRPO [35], without freezing any layers. The AdamW optimizer [27] is employed with an initial learning rate of 1×10^{-6} and a linear decay schedule. For GRPO, we generate six candidate responses per prompt (*i.e.*, $K = 6$) and set the balance coefficient β to 0.04. Training runs on 16 NVIDIA A100 GPUs with a minibatch size of eight per GPU, taking approximately five hours for a total of 10 epochs.

4.2 Main Results

Single-Dataset Training We first train NR-IQA models exclusively on the synthetic KADID-10K [23] training set ($6 : 2 : 2$ split while ensuring content independence) and tested in a zero-shot setting across eight datasets with distortions arising from digital imaging and (post-)processing stages: BID [7], CLIVE [11], KonIQ-10k [15], SPAQ [8], Liu13 (deblurring) [26], SRIQA-Bench (super-resolution) [6], Min19 (dehazing) [31], and AGIQA-3K (image generation) [20].

The SRCC and PLCC results presented in Table 2 reveal several key observations. <u>First</u>, all VLM-based models outperform traditional and discriminative deep-learning-based ones, with the base Qwen2.5-VL-7B achieving an SRCC of 0.708 despite no IQA-specific training. This underscores

Table 3: PLCC results of VisualQuality-R1 with varying K in GRPO. The default setting is highlighted in **bold**.

# of Generated Responses	Imaging-Related Distortion				Processing-Related Distortion				Avg
	BID	CLIVE	KonIQ	SPAQ	De-blurring	Super-Res.	De-hazing	Image Gen.	
$K = 4$	0.805	0.795	0.839	0.875	0.875	0.815	0.643	**0.844**	0.811
$K = 5$	**0.806**	**0.804**	**0.840**	**0.879**	0.867	**0.826**	0.639	0.840	0.813
K = 6	**0.806**	0.794	**0.840**	0.878	**0.872**	0.825	**0.651**	0.843	**0.814**

Table 4: Comparison of different Thurstone model variants [40] in the GRPO of VisualQuality-R1.

Method	Imaging-Related Distortion				Processing-Related Distortion				Avg
	BID	CLIVE	KonIQ	SPAQ	De-blurring	Super-Res.	De-hazing	Image Gen.	
SRCC									
Q-Insight [21]	0.784	**0.761**	0.806	0.872	0.831	0.724	0.601	0.749	0.766
Binary Reward	0.780	0.756	0.821	**0.877**	0.834	0.748	0.587	0.771	0.772
Probability Average (Eq. (7))	0.785	**0.761**	0.836	0.875	0.835	0.747	0.574	**0.775**	0.774
Fixed Variance of One	0.778	0.750	0.818	0.871	0.830	0.744	**0.606**	0.760	0.770
VisualQuality-R1	**0.790**	0.750	0.830	0.875	**0.838**	**0.756**	0.598	**0.775**	**0.777**
PLCC									
Q-Insight [21]	0.796	0.795	0.829	0.872	0.857	0.798	**0.669**	0.810	0.803
Binary Reward	0.790	0.792	0.833	0.876	0.867	**0.825**	0.646	0.840	0.809
Probability Average (Eq. (7))	0.796	**0.797**	**0.844**	0.875	0.861	0.817	0.621	0.831	0.805
Fixed Variance of One	0.791	0.785	0.817	0.873	0.852	0.802	0.655	0.818	0.799
VisualQuality-R1	**0.806**	0.794	0.840	**0.878**	**0.872**	0.825	0.651	**0.843**	**0.814**

the power of current VLMs in capturing generalizable quality cues. Second, reasoning-induced models such as Q-Insight and VisualQuality-R1 surpass SFT-based counterparts like Q-Align and DeQA-Score. Third, the proposed VisualQuality-R1 achieves the best results on average, validating that RL2R aligns better with human perception of image quality than regression-based approaches.

Multi-Dataset Training   Our RL2R approach enables multi-dataset training without the need for perceptual scale realignment. To exploit this, we train VisualQuality-R1[†] on a combination of KADID-10K [23] and SPAQ [8] (again $6 : 2 : 2$ split while ensuring content independence). As shown in Table 2, VisualQuality-R1[†] yields consistent performance gains. Despite a minor dip in the image generation scenario, the average SRCC/PLCC rises from $0.777/0.814$ to $0.791/0.831$. In stark contrast, Q-Insight[†] [21] fails to benefit from multi-dataset training due to its inability to address perceptual scale variations[3]: KADID-10K uses ratings from 1 to 5, while SPAQ spans 0 to 100.

4.3 Ablation Studies

Effect of K in GRPO   We vary the number of generated responses, K, while keeping all other settings fixed during GRPO. Table 3 shows that reducing K from six (default) to four or five has only a marginal effect, offering a favorable trade-off between computational cost and accuracy.

Binary Reward vs. Continuous Fidelity Reward   Table 4 shows that, within the same RL2R framework, our continuous fidelity reward generalizes better than the binary reward adopted in GRPO [35]. Moreover, both reward variants consistently outperform the regression-based Q-Insight [21], underscoring the effectiveness of our RL2R optimization.

Thurstone Model Variants   To evaluate the effectiveness of mean quality computation in Eq. (1), we compare it with an alternative that averages probabilities across individual quality comparisons:

$$p_k(x_i, x_j) = \frac{1}{K} \sum_{k'=1}^{K} p_{k,k'}(x_i, x_j) = \frac{1}{K} \sum_{k'=1}^{K} \Phi \left(\frac{q_k(x_i) - q_{k'}(x_j)}{\sqrt{\sigma^2(q(x_i)) + \sigma^2(q(x_j)) + \gamma}} \right). \quad (7)$$

---

[3]To enable multi-dataset training of Q-Insight, we linearly rescale MOSs from different IQA datasets to $[1, 5]$, and apply a dataset-agnostic threshold to compute binary rewards.
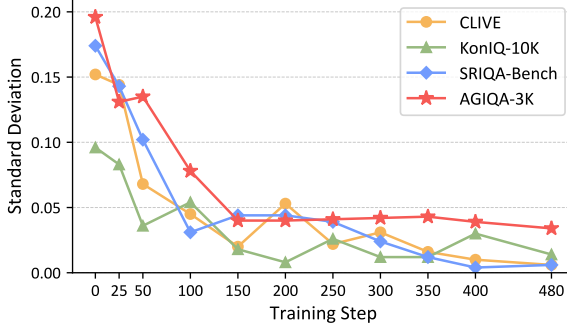
Figure 3: Prediction variability decreases during GRPO. We randomly select 20 images from each of CLIVE [11], KonIQ-10k [15], SRIQA-Bench [6], and AGIQA-3K [20]. At successive training steps, we generate multiple responses per image, compute the std of the predicted quality scores, and plot the average std across images. The uniformly downward trend confirms that VisualQuality-R1 becomes steadily more stable in assessing image quality as training progresses.

As reported in Table 4, averaging quality scores rather than probabilities yields higher performance across distortion scenarios, indicating more reliable comparative probability estimates and reward assignments. Taking a step further, we fix the variances in Eq. (7) to one—reducing the model to Thurstone Case V [40]. The constant-variance simplification degrades performance on nearly all datasets. This provides a strong indication that sample variances are capable of capturing the perceptual difficulty of image pairs, thus improving comparison reliability and stabilizing fidelity reward computation. Together, these findings verify that RL2R effectively embeds the Thurstone model within GRPO.

4.4 Further Analysis

Predicted Score Variability over Iterations We randomly sample 20 images from each of CLIVE [11], KonIQ-10k [15], SRIQA-Bench (super-resolution) [6], and AGIQA-3K (image generation) [20], respectively. At successive training checkpoints, we generate multiple responses per image and compute the standard deviation (std) of the resulting K quality scores. As illustrated in Fig. 3, the std falls steadily across all datasets, indicating that predictions of VisualQuality-R1 become progressively more stable and confident. This confirms that RL2R effectively reduces predictive uncertainty, improving the sensitivity of VisualQuality-R1 to subtle quality differences.

Visual Reasoning Evolution over Iterations Fig. 4 tracks how the visual reasoning capabilities of VisualQuality-R1 mature over the course of training. The test image is super-resolved by SwinIR [22], which contains subtle, processing-related artifacts, making it an informative probe. Q-Insight notices that the image is "blurry" and "overexposed," but assigns an extremely low score (*i.e.*, 2.00), indicating limited sensitivity to super-resolution artifacts. The base model Qwen2.5-VL-7B [1] swings to the opposite extreme: it praises the "clear details" and "vibrant colors," declares the absence of blur or noise, and outputs an inflated score (*i.e.*, 4.80). The model clearly over-trusts superficial sharpness cues and misses hidden processing traces. In contrast, the proposed VisualQuality-R1 progressively refines its visual reasoning over iterations. At the 50-th step, it starts to suspect artificial stylization and questions the image's realism, yet it still values the apparent clarity. By the 200-th step, the description becomes more balanced. It acknowledges the level of detail and clarity, yielding a slightly higher but still cautious rating. At the last step, the explanation is now decidedly nuanced. VisualQuality-R1 attributes the remaining softness to possible filtering or to the object's inherent structure, labels the appearance "surreal," and reduces the score to 3.00, reflecting a judicious penalty for unnatural post-processing. In summary, RL2R guides VisualQuality-R1 from naïve, superficial remarks to sophisticated, human-aligned reasoning that correctly identifies subtle super-resolution artifacts and calibrates quality scores accordingly.

5 Conclusion and Discussion

We have introduced VisualQuality-R1, a reasoning-induced NR-IQA model optimized via RL2R. Our approach is grounded in the intrinsic relativity of visual quality, seamlessly integrating the Thurstone model within GRPO to capture predictive uncertainty. By introducing the continuous fidelity reward, VisualQuality-R1 delivers more precise policy-gradient signals.

Q-Insight

<think>**The image appears to be somewhat blurry**, which affects the clarity and detail that can be discerned. The colors are vibrant, but the lack of sharpness reduces the overall quality. There's also a slight **overexposure** in some areas, which might be due to lighting conditions or camera settings.</think>
<answer>2.00</answer>

Qwen2.5-VL-7B (**0**-th Step)

<think>The image appears to be of high quality with clear details and vibrant colors. The texture and patterns on the objects are well-defined, suggesting good resolution and lighting. **There are no visible distortions or blurriness** that would typically indicate lower quality. The contrast and saturation levels seem appropriate, enhancing the visual appeal without overwhelming the viewer.</think><answer>4.80</answer>

VisualQuality-R1 (**50**-th Step)

<think>The image appears to be somewhat **abstract or stylized**, which might not be ideal for a realistic quality assessment. However, it is clear and the **details are well-defined**, suggesting good resolution. The colors are vibrant and the contrast is high, which can be visually appealing but may not reflect naturalistic quality. The image does not contain any discernible text or identifiable objects that could provide further context.</think><answer>3.50</answer>

VisualQuality-R1 (**200**-th Step)

<think>The image appears to have a high level of detail and clarity, with distinct patterns and colors that are well-defined. However, there is a **slight lack of sharpness around the edges**, which might be due to the texture or the way the image was captured. The lighting seems even, and the colors are vibrant, contributing positively to the overall quality.</think><answer>3.80</answer>

VisualQuality-R1 (**Last** Step)

<think>The image appears to be a close-up of **some kind of fruit or vegetable with a patterned surface**. The colors are vibrant, but the image seems to have a filter applied that gives it a somewhat **surreal appearance**. **The details are not very sharp, which might be due to the nature of the subject or the processing of the image**. However, the saturation and contrast are high, making the image visually striking.</think><answer>3.00</answer>

Figure 4: Evolution of the reasoning capabilities of VisualQuality-R1 on an image super-resolved by SwinIR [22]. Initially, VisualQuality-R1 overlooks artifacts and overestimates quality; at later stages, it progressively detects stylization, blur, and filtering effects, yielding more accurate quality scores and human-aligned textual justifications. Zoom in for improved visibility.

Extensive experiments validate that VisualQuality-R1 consistently surpasses strong discriminative deep learning-based methods and a reasoning-induced baseline. Notably, it bridges the performance gap between synthetic and realistic distortions, demonstrating robustness to dataset inductive biases and noise. In addition to quantitative improvements, VisualQuality-R1 generates contextually rich, human-aligned quality descriptions, which not only enhance transparency and interpretability but also facilitate user trust and post-hoc diagnosis in downstream tasks, such as content filtering, local enhancement prioritization, and quality-aware image retrieval.

Limitations and Future Directions Despite the generalization capabilities demonstrated by VisualQuality-R1, several limitations and promising research directions merit further discussion. First, as a specific case of test-time scaling, VisualQuality-R1 is slow, expensive, and memory-hungry; it may also compound early errors into confidently wrong predictions. It is thus desirable to incorporate sample-adaptive reasoning, rationale compression or distillation, and self-consistency sampling to make VisualQuality-R1 faster, cheaper, and more robust. Second, VisualQuality-R1 relies on a single, fixed text prompt for all images, regardless of the underlying distortion scenario or application context. Incorporating application-aware prompt adaptation, for example via learned prompt-tuning or dynamic template selection, could tailor VisualQuality-R1's reasoning and scoring to specific image processing tasks, therefore improving its flexibility and accuracy. Third, VisualQuality-R1 is currently formulated as an NR-IQA model, focusing solely on distorted inputs without access to pristine counterparts. It is interesting to extend VisualQuality-R1 to a reference-based setting, which allows a (possibly corrupted) reference image [50]—potentially differing in resolution, color gamut, dynamic range, or bit depth—to serve as a flexible anchor for content fidelity. Last, we foresee adapting the proposed RL2R learning algorithm to other perceptual assessment tasks, including image aesthetics assessment [39], human age estimation [14], and perceptual similarity ranking [9]. Collectively, these promising directions aspire to foster more intelligent, transparent, and adaptable perceptual systems.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959, 2020.

[3] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2017.

[4] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *International Conference on Machine Learning*, pages 89–96, 2005.

[5] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *International Conference on Machine Learning*, pages 129–136, 2007.

[6] Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. *arXiv preprint arXiv:2503.11221*, 2025.

[7] Alexandre Ciancio, André Luiz N Targino Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, 2010.

[8] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020.

[9] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, pages 50742–50768, 2023.

[10] Fei Gao, Dacheng Tao, Xinbo Gao, and Xuelong Li. Learning to rank for blind image quality assessment. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2275–2290, 2015.

[11] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

[12] Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, 17(1):32–56, 2017.

[13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[14] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S Huang. Human age estimation using bio-inspired features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 112–119, 2009.

[15] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[17] Ekhine Irurozki, Borja Calvo, and Jose A. Lozano. Mallows and generalized mallows model for matchings. *Bernoulli*, 25(2):1160–1188, 2019.

[18] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[19] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality Transformer. In *IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021.

[20] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: An open database for AI-generated image quality assessment. *arXiv preprint arXiv:2306.04717*, 2023.

[21] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-Insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025.

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin Transformer. In *IEEE/CVF International Conference on Computer Vision Workshops*, pages 1833–1844, 2021.

[23] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10K: A large-scale artificially distorted IQA database. In *IEEE International Conference on Quality of Multimedia Experience*, pages 1–3, 2019.

[24] Hantao Liu, Nick Klomp, and Ingrid Heynderickx. A no-reference metric for perceived ringing artifacts in images. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(4):529–539, 2009.

[25] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *IEEE/CVF International Conference on Computer Vision*, pages 1040–1049, 2017.

[26] Yiming Liu, Jue Wang, Sunghyun Cho, Adam Finkelstein, and Szymon Rusinkiewicz. A no-reference metric for evaluating the quality of motion deblurring. *ACM Transactions on Graphics*, 32(6):1–12, 2013.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[28] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017.

[29] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

[30] Aliaksei Mikhailiuk, María Pérez-Ortiz, Dingcheng Yue, Wilson Suen, and Rafał K Mantiuk. Consolidated dataset and metrics for high-dynamic-range image quality. *IEEE Transactions on Multimedia*, 24:2125–2138, 2021.

[31] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yucheng Zhu, Jiantao Zhou, Guodong Guo, Xiaokang Yang, Xinping Guan, and Wenjun Zhang. Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Transactions on Multimedia*, 21(9):2319–2333, 2019.

[32] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[33] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[34] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, pages 53728–53741, 2023.

[35] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[36] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, pages 9460–9471, 2022.

[37] Hal Stern. Models for distributions on permutations. *Journal of the American Statistical Association*, 85(410):558–564, 1990.

[38] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[39] Xiaoou Tang, Wei Luo, and Xiaogang Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, 2013.

[40] Louis L Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1927.

[41] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 383–390, 2007.

[42] Zhihua Wang and Kede Ma. Active fine-tuning from gMAD examples improves blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4577–4590, 2021.

[43] Zhihua Wang, Haotao Wang, Tianlong Chen, Zhangyang Wang, and Kede Ma. Troubleshooting blind image quality models in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16256–16265, 2021.

[44] Zhou Wang and Alan C Bovik. Reduced-and no-reference image quality assessment. *IEEE Signal Processing Magazine*, 28(6):29–40, 2011.

[45] Zhou Wang, Alan C Bovik, and Brian L Evan. Blind measurement of blocking artifacts in images. In *IEEE International Conference on Image Processing*, pages 981–984, 2000.

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[47] Zhou Wang and Eero P Simoncelli. Local phase coherence and the perception of blur. In *Advances in Neural Information Processing Systems*, pages 1435–1442, 2003.

[48] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25490–25500, 2024.

[49] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-ALIGN: Teaching LMMs for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, pages 54015–54029, 2024.

[50] Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. A comprehensive study of multimodal large language models for image quality assessment. In *European Conference on Computer Vision*, pages 143–160, 2024.

[51] Kangmin Xu, Liang Liao, Jing Xiao, Chaofeng Chen, Haoning Wu, Qiong Yan, and Weisi Lin. Boosting image quality assessment through efficient Transformer adaptation with local feature enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2662–2672, 2024.

[52] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1191–1200, 2022.

[53] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. *arXiv preprint arXiv:2501.11561*, 2025.

[54] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, pages 259–276, 2024.

[55] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. Continual learning for blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2864–2878, 2022.

[56] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021.

[57] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023.

[58] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *arXiv preprint arXiv:2405.19298*, 2024.