

Reinforcing Vision-Language Models for Image Quality Assessment with Grounding Process Rewards

Jingwei Liu^{1,2,3}, Bo Liu^{1,2}, Hongyan Li^{1,2}, Tianlin Zhang³, Yifei Qian³,
Congyang Zhao³, Junhong Liu, Yang Cai³, and Ling Yang⁴ *

¹ State Key Laboratory of General Artificial Intelligence, Peking University

² School of Intelligence Science and Technology, Peking University

³ Amap, Alibaba Group

⁴ Department of Electrical and Computer Engineering, Princeton University

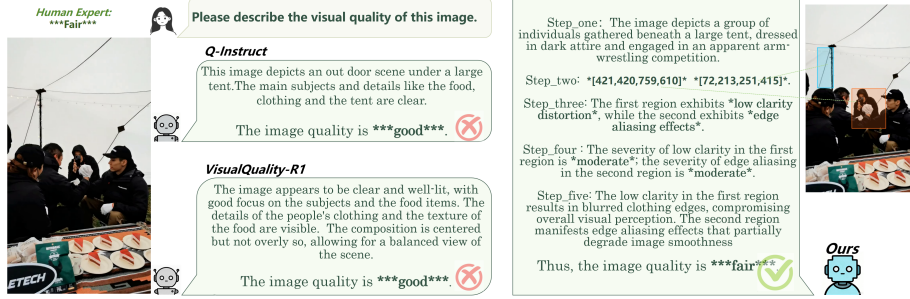


Fig. 1: RGPR utilizes RL to achieve fine-grained reasoning alignment in IQA, improving accuracy while generating high-fidelity explanations. Unlike existing models that yield inflated scores on real-world images by ignoring distortions, RGPR ensures a more stable and human-aligned assessment.

Abstract. Image Quality Assessment (IQA) requires aligning model outputs with human perception. While Vision-Language Models (VLMs) show promise, current methods struggle with generalization due to a lack of fine-grained reasoning alignment. Existing RL-based approaches primarily rely on outcome rewards; however, their inherent sparsity often results in accurate predictions based on erroneous logical attributions. This phenomenon creates a critical bottleneck that stifles the development of robust reasoning capabilities and limits performance gains in out-of-distribution scenarios. To address this problem, we propose RGPR, which utilizes *grounding process rewards* to orchestrate VLM trajectories along a human-emulated reasoning chain. To maintain optimization stability, we introduce *step-wise reward normalization* to preserve training signal precision. RGPR achieves state-of-the-art results across three tasks, demonstrating superior reasoning fidelity and robust generalization.

* Corresponding authors: Yang Cai, Ling Yang.

1 Introduction

The proliferation of User-Generated Content (UGC) and the rapid advancement of Artificial Intelligence Generated Content (AIGC) have positioned No-Reference Image Quality Assessment (IQA) as an essential task within image processing systems and global communication networks [12, 16, 25, 45]. Traditional NR-IQA methods historically relied on Natural Scene Statistics (NSS) and hand-crafted features to predict quality scores [26–29, 52]. The subsequent deep learning based approaches have significantly improved performance, particularly in handling complex visual distortions [10, 18, 24, 32, 34, 45, 55].

More recently, Vision-Language Models (VLMs) have introduced novel capabilities for IQA by demonstrating potential in fusing high-level semantic understanding with the perception of low-level visual distortions [2, 15, 30, 31, 51]. While initial efforts leveraged Supervised Fine-Tuning (SFT) to specialize VLMs for IQA [40, 46, 48, 58, 62], these methods remain constrained by inherent challenges, such as catastrophic forgetting and limited generalization.

Reinforcement Learning (RL) has emerged as a more powerful paradigm for fine-tuning VLMs by aligning model outputs with human annotations via reward-driven optimization. For instance, Q-Insight [19] defines rewards based on evaluative consistency, while VisualQuality-R1 [41] utilizes rewards derived from ranking-based preferences. However, these outcome-based reward designs suffer from inherent sparsity, providing insufficient guidance for the intermediate reasoning steps. The sparsity leaves the model’s internal logic largely underspecified, frequently allowing it to garner high rewards for “correct answers derived from flawed internal logic” [36, 57]. Consequently, these logical flaws hinder robust reasoning, leading to poor out-of-distribution generalization (Fig. 1). Moreover, while naive multi-dimensional rewards may mitigate sparsity, they often trigger training instability and degrade signal discriminability [17].

To address these issues, we propose Reinforcement Learning with Grounding Process Reward (RGPR), a novel RL-based framework that resolves the limitations of outcome-based rewards by aligning VLM trajectories with the explicit grounding of visual distortions. Specifically, we introduce *Grounding Process Rewards* (GPR) to enforce a “ground-before-assess” principle that provides the critical intermediate guidance necessary to mitigate reward sparsity and rectify erroneous logical attributions. This mechanism ensures that quality assessments are strictly predicated on grounded visual evidence, thereby fostering the robust reasoning capabilities required for superior out-of-distribution generalization. To stabilize the inherent multi-dimensional optimization, we further introduce *step-wise reward normalization*, which independently calibrates rewards at each reasoning stage to prevent reward aliasing and preserve the discriminative precision of training signals. Comprehensive evaluations across three tasks demonstrate that RGPR achieves state-of-the-art (SOTA) results, substantiating its superior reasoning fidelity and robust generalization.

Our contributions are summarized as follows:

- We propose RGPR, the first framework to achieve fine-grained alignment between VLM reasoning trajectories and human preferences in NR-IQA by shifting from outcome-based to process-based reinforcement.
- We introduce the Grounding Process Rewards based on a human-emulated reasoning chain. This mechanism effectively mitigates reasoning hallucinations and enhances the fidelity of explanatory assessments.
- To address the optimization imbalances inherent in multi-step supervision, we introduce the Step-wise Reward Normalization mechanism that prevents reward aliasing and preserves the discriminative precision of training signals, ensuring stable and effective policy updates.
- Extensive evaluations demonstrate that RGPR establishes new SOTA benchmarks across three tasks, substantiating its robust generalization and superior ability to generate interpretable IQA outcomes.

2 Related Work

No-Reference Image Quality Assessment No-Reference Image Quality Assessment (NR-IQA) aims to quantify perceptual degradation in alignment with human consensus. Traditional methods primarily relied on Natural Scene Statistics (NSS) and hand-crafted features [27, 28, 52], yet struggled with the diversity of real-world distortions. The advent of deep learning transitioned the field toward regression-based [34, 45] and ranking-based architectures [55, 56]. While these models achieve high correlation with human ratings, their black-box nature often leads to poor generalization across unseen domains.

Recently, Vision-Language Models (VLMs) have revolutionized IQA by enabling descriptive assessments via Supervised Fine-Tuning (SFT) [40, 47, 58]. Significant efforts have been dedicated to data curation, such as Q-Instruct [39] for low-level visual feedback and Vida-UGC [21] for detailed reasoning chains. However, SFT-based models are prone to shortcut learning, where they mimic the linguistic style of human annotations without truly mastering the underlying visual logic, leading to marginal gains in out-of-distribution (OOD) generalization. Although recent RL-based methods [19, 41] have improved scoring stability, their outcome-only reward signals fail to supervise the intermediate reasoning process. Our method fills this gap by introducing GPR, ensuring that the model’s internal reasoning trajectories are explicitly aligned with human-like distortion discovery.

VLM as Visual Content Judge VLMs are increasingly utilized as "judges" to evaluate visual content. One branch of research integrates external heads (e.g., Bradley-Terry layers) for preference regression [23, 42]. However, such methods bypass the VLM’s inherent generative potential, sacrificing interpretability. Conversely, generative judges like WorldPM [33] and UnifiedReward [37] leverage long-chain reasoning to provide evaluative feedback.

Despite their success in high-level semantic evaluation, these models struggle to transfer to the NR-IQA domain. IQA necessitates an acute sensitivity to fine-grained perceptual artifacts (e.g., aliasing, compression noise), which fundamentally differs from the high-level semantic understanding (e.g., object relations, scene logic) emphasized in general VLM pre-training. Unlike universal judges that provide a holistic preference, RGPR establishes a fine-grained alignment

between the VLM and human reasoning paths specifically for IQA, yielding both accurate quality metrics and verifiable, grounding-based explanations.

3 Preliminary

3.1 Image Quality Assessment with Explanations

VLMs typically generate text responses $\mathcal{C}$ autoregressively based on input images $\mathcal{I}$ and prompts $\mathcal{P}$:

$$\mathcal{C} \sim \pi_\theta(\cdot \mid \mathcal{X}) \quad (1)$$

where $\mathcal{V}_{\pi_\theta}$ is the VLM governed by policy π_θ . While IQA tasks require VLMs to output evaluation results $\mathcal{R}$ that align with human references, IQA with explanations requires that $\mathcal{V}_\theta$ simultaneously output the explanation behind their assessments. In this paper, we define this task as:

$$\mathcal{C} \supseteq \{g_i, \tau_i, \delta_i\}_{i=1}^k \cup \{\mathcal{R}\} \quad (2)$$

where k represents the number of identified distortions, $g_i \in \mathbb{R}^4$ denotes the bounding box coordinates of the i -th distortion, τ_i signifies the distortion type, and δ_i refers to the distortion severity. Distinct from prior approaches [19] that constrain the model to a single distortion ($k = 1$), our formulation allows $\mathcal{C}$ to contain multiple distortion tuples, thereby reflecting the compositional nature of real-world image degradations.

3.2 Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) is a value-free policy optimization method that aligns model outputs with rewards by estimating advantages *relatively* within a group of rollouts sampled for the same input. Concretely, for each training sample $\mathcal{X}_i$, we draw a group of G trajectories $\{\mathcal{C}^{(i,j)}\}_{j=1}^G$ from the current (behavior) policy $\pi_{\theta_{\text{old}}}$:

$$\mathcal{C}^{(i,j)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \mathcal{I}, \mathcal{P}), \quad j = 1, \dots, G. \quad (3)$$

For each trajectory, a total reward is computed based on I individual reward objectives:

$$r^{(i,j)} = \sum_{b=1}^B r_b(\mathcal{X}_i, \mathcal{C}^{(i,j)}) \quad (4)$$

GRPO constructs a group-relative advantage by normalizing rewards within the same group:

$$\hat{A}_{i,j} = \frac{r^{(i,j)} - \text{mean}(\{r^{((i,1))}, \dots, r^{((i,G))}\})}{\text{std}(\{r^{((i,1))}, \dots, r^{((i,G))}\}) + \epsilon}. \quad (5)$$

Let $s_{i,j,t}(\theta)$ denote the token-level importance ratio for the t -th token of trajectory $\mathcal{C}^{(i,j)}$:

$$s_{i,j,t}(\theta) = \frac{\pi_\theta(c_{i,j,t} \mid \mathcal{X}_i, c_{i,j,<t})}{\pi_{\theta_{\text{old}}}(c_{i,j,t} \mid \mathcal{X}_i, c_{i,j,<t})}. \quad (6)$$

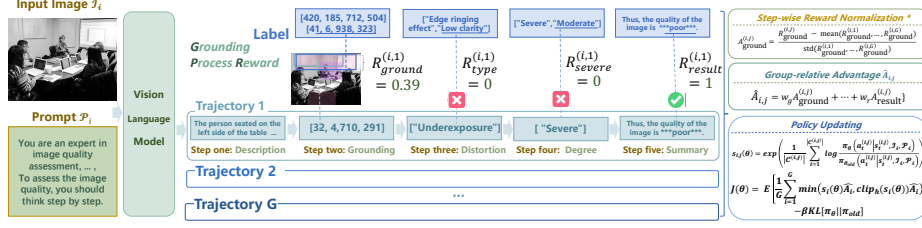


Fig. 2: Overview of the RGPR. The * in the upper-right corner indicates that the same operation is also applied to the rewards of the other steps. Models can often produce label-consistent assessment results despite incorrect reasoning, making outcome-only training limited. We mitigate this issue with process rewards that enforce fine-grained alignment along the reasoning trajectory.

GRPO then maximizes a clipped objective using the group-relative advantages:

$$\mathcal{L}_{\text{clip}}(\theta; \hat{A}) = \frac{1}{G} \sum_{j=1}^G \frac{1}{|\mathcal{C}^{(i,j)}|} \sum_{t=1}^{|\mathcal{C}^{(i,j)}|} \min(s_{i,j,t}(\theta) \hat{A}_{i,j}, \text{clip}(s_{i,j,t}(\theta), \epsilon_c) \hat{A}_{i,j}) - \beta \mathcal{L}_{\text{KL}}(\theta). \quad (7)$$

where ϵ_c is the clipping threshold, β controls the KL regularization strength, and ϵ in Eq. (5) is a small constant for numerical stability.

4 Method

This section delineates the proposed RGPR framework. We begin by employing SFT to cold-start the VLM and internalize a structured multi-step reasoning format (Sec. 4.1). Subsequently, we introduce the core of our approach: a RL framework utilizing fine-grained Grounding Process Rewards (Sec. 4.2) and a step-wise reward normalization mechanism (Sec. 4.3). Finally, we describe the trajectory-level policy optimization strategy designed for stable training (Sec. 4.4).

4.1 Structured Multi-Step Reasoning with SFT

Real-world images often exhibit complex, compositional distortions, making IQA with high-quality explanations a formidable challenge. To address this, VLMs require structured multi-step reasoning to systematically decompose the assessment process [20]. While recent VLMs incorporate inherent thinking modes [13], their performance often plateaus in tasks like NR-IQA.

Following established practices [4, 9], we employ SFT to cold-start the VLM and internalize a fixed multi-step reasoning paradigm. Unlike mathematical reasoning tasks [44, 53], the IQA task allows for a structured, fixed reasoning format [37]. Drawing upon annotation experience from professional NR-IQA experts, we devise the following five-step reasoning trajectory: description, distortion grounding, distortion type discrimination, distortion severe assessment, and summarization. This trajectory ensures that global quality ratings are strictly

predicated on verifiable local evidence. By enforcing a "ground before assess" principle, the framework effectively mitigates reasoning hallucinations and prevents the occurrence of correct scores derived from erroneous logical attributions.

We curate data from [21] and reorganize it into this multi-step chain. Specifically, the grounding step is annotated with M bounding boxes delineating degraded regions. The subsequent steps provide per-box labels for category and severity, followed by a final quality rating. Each step is encapsulated by structured delimiters. The SFT objective is defined as:

$$\mathcal{L}_{\text{SFT}}(\theta) = - \sum_{t=1}^T \log p(c_t \mid \mathcal{I}, \mathcal{P}, c_{<t}; \theta) \quad (8)$$

4.2 Design of Grounding Process Rewards

Motivation. Compared to sparse outcome-based rewards, process-based supervision provides dense and interpretable signals by verifying intermediate reasoning steps [22, 54]. This advantage is vital for NR-IQA to ensure a rigorous alignment between evaluative outcomes and underlying visual evidence. Motivated by this, we propose *Grounding Process Rewards* to enforce a "ground-before-assess" principle. GPR compels the model to master robust visual cues rather than exploiting spurious correlations, thereby enhancing both reasoning fidelity and generalization [1, 61].

GPR comprise the grounding reward R_{ground} , the type reward R_{type} , the severe reward R_{severe} , and the result reward R_{result} . To compute R_{ground} , we first match the predicted bounding boxes $\mathcal{G} = \{g_n\}_{n=1}^N$ with ground-truth labels $L = \{l_m\}_{m=1}^M$ using an Intersection over Union (IoU) threshold τ :

$$\mathcal{M}(g_n, L) = \mathbb{I} \left[\max_{m \in [M]} \text{IoU}(g_n, l_m) > \tau \right] \quad (9)$$

where $\mathbb{I}$ refers to the indicator function. The grounding reward comprises two components: precision-based IoU (S_{IoU}) and recall (S_{recall}). S_{IoU} measures the cumulative maximum IoU for all predicted boxes that satisfy the matching criterion:

$$S_{\text{IoU}} = \frac{1}{N} \sum_{n=1}^N \mathcal{M}(g_n, L) \cdot \text{IoU}(g_n, L) \quad (10)$$

, and S_{recall} quantifies the matching probability between model-predicted bounding boxes and ground-truth labels:

$$S_{\text{Recall}} = \frac{1}{M} \sum_{m \in [M]} \mathbb{I}[\exists n : \text{IoU}(g_n, l_m) > \tau] \quad (11)$$

To prevent the model from exploiting the reward via blind guessing (e.g., outputting excessive or over-sized boxes), we introduce an additional penalty term:

$$\begin{aligned}\mathcal{P} &= \lambda_n \left[\frac{N-M}{M} - \tau_n \right]_+ + \lambda_a \left[\frac{\bar{\mathcal{G}} - \bar{L}}{\bar{L}} - \tau_a \right]_+ \\ \bar{\mathcal{G}} &\triangleq \frac{1}{N} \sum_{n=1}^N \text{area}(g_n), \quad \bar{L} \triangleq \frac{1}{M} \sum_{m=1}^M \text{area}(\ell_m).\end{aligned}\tag{12}$$

where $[\cdot]_+$ denotes $\max(0, \cdot)$, τ_n, τ_a denotes the penalty threshold and $\text{area}(\cdot)$ denotes the area covered by the bounding box. The grounding reward is formally defined as follows:

$$R_{\text{ground}} = S_{\text{recall}} + S_{\text{IoU}} - \mathcal{P}\tag{13}$$

For the subsequent steps, R_{type} , R_{severe} , and R_{result} are binary rewards based on exact label matching. Crucially, we implement an *early-stopping principle*: if the model fails to correctly ground any distortions ($S_{\text{recall}} = 0$), all subsequent process rewards are set to zero. This ensures that the model is penalized for providing correct assessments derived from erroneous grounding.

During training, we iteratively compute the aforementioned rewards for response $\mathcal{C}^{(i,j)}$ to obtain $(R_{\text{ground}}^{(i,j)}, R_{\text{type}}^{(i,j)}, R_{\text{severe}}^{(i,j)}, R_{\text{result}}^{(i,j)})$, which is subsequently utilized for advantage estimation.

4.3 Step-wise Reward Normalization

Motivation. Despite providing dense guidance, multi-step process rewards often suffer from reward aliasing [5, 17, 50], collapsing of different rewards into indistinguishable cumulative values. This phenomenon erodes the discriminative precision of training signals, making it difficult to penalize nuanced errors like ‘‘correct grounding with faulty analysis.’’ To address this, we propose *step-wise reward normalization* to independently calibrate each reasoning stage. By normalizing rewards prior to aggregation, this mechanism ensures balanced policy updates and prevents any single dimension from dominating, thereby maintaining the structural integrity of the reasoning chain.

The step-wise reward normalization independently calibrate the advantage for each reasoning step within the group:

$$A_b^{(i,j)} = \frac{R_b^{(i,j)} - \mu(\mathcal{R}_b^{(i)})}{\sigma(\mathcal{R}_b^{(i)}) + \epsilon}, \quad b \in \{\text{ground, type, } \dots\}\tag{14}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation of the reward set $\mathcal{R}_b^{(i)} = \{R_b^{(i,1)}, \dots, R_b^{(i,G)}\}$, respectively.

The final group-relative advantage is then aggregated via a weighted summation:

$$\hat{A}_{i,j} = w_g A_{\text{ground}}^{(i,j)} + \dots + w_r A_{\text{result}}^{(i,j)}\tag{15}$$

Specifically, we assign a higher weight to $A_{\text{ground}}^{(i,j)}$ to enhance the model’s grounding capabilities. Alternatively, we investigated a dual-model relay mechanism as a strategy for refining optimization precision. More analyses of this approach are provided in Appendix.

4.4 Trajectory-Level Policy Optimization

Standard GRPO employs token-level importance sampling. However, we observed that this induces significant instability in IQA tasks. Low-frequency tokens, such as numerical coordinates in bounding boxes, often exhibit disproportionately large likelihood ratios compared to common text, leading to extreme gradient magnitudes and training collapse. To mitigate this, we adopt trajectory-level policy optimization [60], averaging the importance weights across the entire response:

$$s_{i,j}(\theta) = \exp \left(\frac{1}{|\mathcal{C}^{(i,j)}|} \sum_{t=1}^{|\mathcal{C}^{(i,j)}|} \log \frac{\pi_{\theta}(a_t^{(i,j)} | s_t^{(i,j)}, \mathcal{I}_i, \mathcal{P}_i)}{\pi_{\theta_{\text{old}}}(a_t^{(i,j)} | s_t^{(i,j)}, \mathcal{I}_i, \mathcal{P}_i)} \right) \quad (16)$$

where $a_t^{(i,j)}$ and $s_t^{(i,j)}$ denote the generated token and the corresponding contextual state at time step t within the trajectory $\mathcal{C}^{(i,j)}$, respectively. Furthermore, to address the “zero-reward group” problem in early training, we adopt asymmetric gradient clipping (setting $\epsilon_{\text{upper}} = 0.25$) and dynamic sampling [49]. The final objective is:

$$J_i(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{j=1}^G \min(s_{i,j}(\theta) \hat{A}_{i,j}, \text{clip}_h(s_{i,j}(\theta)) \hat{A}_{i,j}) \right] - \beta KL[\pi_{\theta} || \pi_{\text{old}}] \quad (17)$$

Table 1: Explanatory Quality Description Task ability comparison between baseline models and our method. We directly utilize released version of other methods for evaluation. GPT-4o [15] was utilized as the judge model. The best is in **bold**, while the second best is underlined. Human evaluation result is shown at Appendix.

Model	Q-Bench						ViDA-UGC-Test*					
	complete- ness	precision grounding	type	severe	rele- vance	overall	complete- ness	precision grounding	type	severe	reason- ing	overall
<i>No Fine-tune</i>												
Qwen2-VL-7B [35]	1.36	0.75	0.54	1.11	1.95	5.71	0.78	0.61	0.77	0.79	0.91	3.86
Qwen2.5-VL-7B [2]	1.44	1.16	0.89	1.43	1.95	6.87	0.91	0.71	0.83	0.85	1.15	4.45
Qwen3-VL-8B [43]	1.46	1.19	0.91	1.42	1.98	6.96	0.94	0.73	0.81	0.90	1.62	5.00
<i>Supervised Fine-tuned</i>												
DepictQA-wild [47]	1.45	1.10	1.03	1.29	1.97	6.84	0.79	0.96	1.05	1.11	1.75	5.66
Q-instruct [39]	1.44	1.13	1.15	1.24	<u>1.99</u>	6.95	0.93	0.99	0.99	1.04	1.93	5.88
Qwen2.5-VL-7B-SFT	1.52	1.27	1.29	1.33	2.00	7.41	1.26	1.13	1.13	1.19	1.77	6.48
Qwen3-VL-8B-SFT	1.56	1.29	1.30	1.32	2.00	7.47	1.29	1.13	1.15	1.21	1.76	6.54
<i>Reinforcement Fine-tuned</i>												
Q-insight [19]	1.52	1.34	1.29	1.30	2.00	7.45	1.13	1.07	1.07	1.22	1.93	6.42
VisualQuality-R1 [41]	1.56	1.23	1.33	<u>1.35</u>	2.00	7.47	1.15	1.05	1.11	1.23	<u>1.95</u>	6.53
Qwen2.5-VL-7B-RGPR	<u>1.61</u>	<u>1.38</u>	<u>1.45</u>	1.39	2.00	<u>7.83</u>	<u>1.33</u>	<u>1.24</u>	<u>1.21</u>	<u>1.27</u>	1.96	<u>6.99</u>
Qwen3-VL-8B-RGPR (Ours)	1.64	1.39	1.52	1.39	2.00	7.93	1.35	1.25	1.26	1.28	1.96	7.10

Table 2: Perception ability comparison between baseline models and our method. Q-Bench refers to the LLVisionQA-*dev*. The evaluation is under the same settings as described in Table 1.

Model	Q-Bench									ViDA-UGC-Test*			
	Yes-or-No ↑	What ↑	How ↑	Distortion ↑	Other ↑	I-C Dist ↑	I-C Other ↑	Overall ↑		Type ↑	Position ↑	Severe ↑	Overall ↑
<i>No Fine-tune</i>													
Qwen2.5-VL-7B [2]	79.75%	82.21%	73.27%	76.40%	78.41%	77.95%	82.91%	78.36%	63.95%	68.36%	73.11%	67.77%	
Qwen3-VL-8B [43]	80.31%	82.19%	73.82%	77.21%	78.12%	78.83%	82.97%	78.74%	64.57%	68.52 %	74.37%	68.37%	
<i>Supervised Fine-tuned</i>													
Q-instruct [39]	81.18%	80.29%	71.47%	75.83%	77.53%	76.91%	82.99%	77.71%	67.63%	72.62%	83.16%	73.19%	
Qwen2.5-VL-7B-SFT	82.20%	83.17%	74.69%	78.60%	80.12%	79.93%	82.99%	80.02%	68.40%	72.95%	84.67%	73.97%	
Qwen3-VL-8B-SFT	82.87%	83.15%	74.66%	79.15%	80.09%	80.15%	83.01%	80.25%	73.92%	71.96%	77.38%	74.03%	
<i>Reinforcement Fine-tuned</i>													
Q-insight [19]	81.81%	82.39%	73.93%	77.07%	80.33%	79.01%	83.11%	79.39%	73.62%	75.41%	78.14%	75.36%	
VisualQuality-R1 [41]	82.81%	83.15%	75.19%	78.40%	81.41%	80.29%	83.01%	80.40%	72.23%	76.22%	90.45%	78.07%	
Qwen2.5-VL-7B-RGPR (Ours)	83.30%	83.19%	75.03%	79.15%	80.78%	80.43%	83.21%	80.54%	81.28%	80.32%	90.95%	83.25%	
Qwen3-VL-8B-RGPR (Ours)	83.81%	<u>83.18%</u>	75.80%	79.15%	<u>81.31%</u>	81.25%	83.27%	80.87%	82.82%	81.80%	91.20%	84.45%	

5 Experiment

5.1 Experimental Setup

Dataset. We construct our training set using VIDA-UGC [21], a comprehensive IQA dataset providing detailed annotations for distortion regions, categories, and severity. The dataset comprises 11,534 images, including 7,138 samples aggregated from KonIQ-10K [14], SPAQ [8], and other public sources, alongside 4,396 newly collected images. We developed an automated pipeline to transform these annotations into a structured multi-step reasoning format (details in Appendix). Due to the lack of an official test split, we employ an 8 : 2 ratio to partition the data into a training set and an internal test set (VIDA-test*). Crucially, our model is trained exclusively on the explanatory quality description task; it is evaluated in a **zero-shot manner** on multiple-choice questions (MCQs) and precise quality-scoring tasks without any task-specific fine-tuning.

Tasks and Metrics. To comprehensively evaluate our method on IQA tasks, four tasks are utilized: **(i) Explanatory Quality Description Task.** This task assesses the quality of the generated explanations using GPT-based evaluation against ground-truth descriptions across Precision, Completeness, and Relevance (each scored on a $[0, 2]$ scale). For fine-grained analysis, Precision is further decomposed into Grounding, Type Discrimination, and Severe Discrimination. The evaluation sets comprise 100 images from Q-Pathway [39] with annotated descriptions and the 2,262 images from VIDA-test*. **(ii) Perception Task.** This evaluates the model’s ability to answer MCQs regarding low-level image attributes (e.g., clarity, brightness). We use LLVisionQA-dev [38] (1,495 MCQs) and VIDA-test* (2,262 MCQs) as evaluation sets. The metric is answering accuracy. **(iii) Quality Assessment Task.** This measures the model’s ability to accurately predict image visual quality based on output logits. Performance is validated on five popular benchmarks: KonIQ [14], SPAQ [8], CLIVE [11], BID [6], and SRIQA-Bench [3]. Metrics used are the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-Order Correlation Coefficient (SRCC).

Furthermore, we also evaluate models on **Grounding Task**. We specifically test the model’s ability to ground distortion regions, measuring performance through matching rate between outputs and labels as well as IoU metrics.

Baseline models. For explanatory and perception tasks, we compare against: (i) **Vanilla VLMs**: Qwen2-VL-7B [35], Qwen2.5-VL-7B [2], and Qwen2.5-VL-7B-Instruct. (ii) **SFT-based VLMs**: DepictQA-wild [47] and Q-Instruct [39]. (iii) **RL-based VLMs**: Q-Instruct [19] and VisualQuality-R1 [41]. For the Quality Assessment task, we include traditional methods (NIQE [27] and BRISQUE [26]), discriminative deep models (UNIQUE [55], MUSIQ [18], and MANIQA [45]), and VLM-based methods (LIQE [56], DeQA-Score [46], Q-Instruct [19], and VisualQuality-R1 [41]).

Implementation Details. We adopt Qwen3-VL-8B-Instruct [43] as our base model. For the RL phase, the group size is set to $G = 8$, and the KL penalty weight β is 1×10^{-3} . The matching threshold τ in Eq. 12 is 0.4. For dynamic sampling, the maximum resampling count is 3. We use the AdamW optimizer with a learning rate of 1×10^{-6} , decaying linearly to 1×10^{-9} . Training is conducted over 3 epochs with a global batch size of 32 across $8 \times$ NVIDIA A100 GPUs using the Unsloth [7] and Swift [59] frameworks.

Table 3: Quality Assessment ability comparison. Note that some baseline results are sourced from VisualQuality-R1 [41]. To prevent data leakage, data collected from KonIQ [14] and SPAQ [8] were excluded from our training set. The best is in **bold**, while the second best is underlined.

Model	KonIQ		SPAQ		CLIVE		BID		SRIQA		Avg	
	<i>SRCC</i>	<i>PLCC</i>	<i>SRCC</i>	<i>PLCC</i>	<i>SRCC</i>	<i>PLCC</i>	<i>SRCC</i>	<i>PLCC</i>	<i>SRCC</i>	<i>PLCC</i>	<i>SRCC</i>	<i>PLCC</i>
<i>Handcrafted</i>												
NIQE [27]	0.421	0.439	0.676	0.683	0.450	0.494	0.515	0.527	0.557	0.587	0.524	0.546
BRISQUE [26]	0.385	0.400	0.614	0.624	0.314	0.362	0.522	0.528	0.482	0.556	0.463	0.494
<i>Discriminative Deep-Learning-based</i>												
UNIQUE [55]	0.649	0.590	0.751	0.708	0.470	0.472	0.412	0.385	0.649	0.668	0.586	0.564
MUSIQ [18]	0.473	0.435	0.720	0.666	0.284	0.325	0.327	0.280	0.404	0.441	0.441	0.429
MANIQA [45]	0.213	0.257	0.745	0.753	0.487	0.571	0.420	0.512	0.263	0.243	0.426	0.467
<i>VLM-based</i>												
LIQE [56]	0.684	0.652	0.815	0.814	0.719	0.726	0.677	0.680	0.743	0.775	0.728	0.729
Q-Align [40]	0.573	0.612	0.767	0.779	0.554	0.643	0.576	0.651	0.684	0.713	0.631	0.680
DeQA-Score [46]	0.677	0.703	0.852	0.858	0.743	0.795	0.702	0.743	0.710	0.763	0.737	0.772
Qwen2.5-VL-7B [2]	0.754	0.810	0.848	0.854	0.733	0.760	0.711	0.725	0.603	0.653	0.730	0.760
Q-Instruct [19]	0.812	0.809	0.907	0.912	0.804	0.837	0.806	0.818	0.700	0.779	0.806	0.831
VisualQuality-R1 [41]	<u>0.855</u>	<u>0.870</u>	<u>0.913</u>	<u>0.917</u>	0.811	0.844	<u>0.811</u>	<u>0.820</u>	0.752	0.824	<u>0.828</u>	<u>0.855</u>
Qwen3-VL-8B-RGPR (Ours)	0.857	0.881	0.919	0.921	<u>0.810</u>	<u>0.839</u>	0.819	0.823	<u>0.750</u>	<u>0.823</u>	0.831	0.857

5.2 Main Result

(i) **Explanatory Quality Description (EQD) Task.** As summarized in Table 1, our RGPR framework yields a marked improvement in the average EQD score, surpassing the base model by 1.53 points. Notably, even when compared to the

Table 4: Comparison of the **Grounding** ability between baseline VLMs and fine-tuned VLMs. The best is in **bold**, while the best improvement is in **red**.

Method	Resp.	Acc _{0.5}	mIoU
Qwen2.5-VL-7B	0.80	18.3	30.1
+SFT	0.91	31.3(+13.0)	35.4(+5.3)
+RGPR	1	38.1(+19.8)	41.2(+11.1)
Qwen3-VL-8B	0.89	19.2	32.1
+SFT	0.98	31.2(+12)	39.6(+7.4)
+RGPR	1	40.0(+20.8)	43.9(+11.8)

Table 5: Performance of models trained with two different reward mechanisms across training steps, evaluated on Q-bench.

Reward	200	400	600	800	1000	Final
OR	7.48	7.49	7.53	7.55	7.57	7.60
GPR	7.47	7.48	7.52	7.67	7.81	7.93

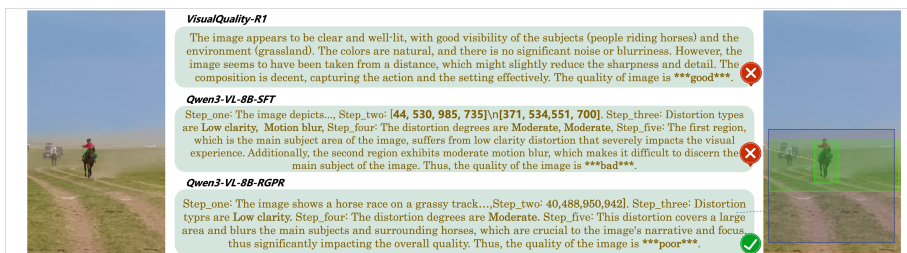


Fig. 3: Case Study. We qualitatively compare RGPR with VisualQuality-R1 [41] and an SFT-only baseline. VisualQuality-R1 tends to yield overly optimistic assessments with coarse-grained explanations. The SFT-only model, while predicting correct scores, often suffers from imprecise grounding and generates redundant distortion regions. In contrast, RGPR achieves superior alignment between visual grounding and textual analysis.

SFT baseline trained on the identical dataset, RGPR achieves a significant average performance gain of over 0.5 points. Notably, RGPR consistently outperforms all state-of-the-art methods across every metric. These performance gains are primarily driven by enhanced precision in *Grounding* and *Type Discrimination*, confirming that our approach significantly bolsters the descriptive fidelity and factual accuracy of the generated explanations. Furthermore, to mitigate potential biases inherent in model-based preferences (e.g., GPT-4 bias), we conducted additional human evaluations to verify the quality of our explanations; these results are detailed in Appendix.

(ii) Perception Task. Tab. 2 demonstrates that the model trained with RGPR achieves superior results on both benchmarks, outperforming the vanilla VLM baseline and SFT-only model by 13% and 10% on ViDA-UGC-Test*. This validates that our grounding-based reinforcement strategy effectively sharpens the model’s sensitivity to low-level visual artifacts, which is fundamental for accurate IQA.

(iii) Quality Assessment Task. Despite the absence of explicit score-regression training, RGPR achieves SOTA performance in a zero-shot manner, even surpassing models fine-tuned on dedicated scoring datasets. This underscores the rectifying power of our GPR: by enforcing a multi-step reasoning trajectory, the model is compelled to base its final rating on verifiable visual evidence rather

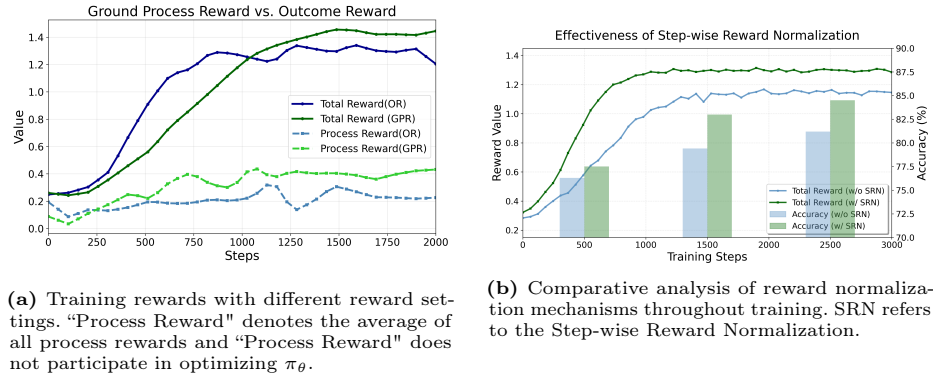


Fig. 4: Further analysis of different mechanisms. More results are shown in Appendix

than spurious correlations. As shown in Fig. 3, while baseline models often yield **inflated ratings** for OOD images, RGPR maintains robust evaluative integrity through its evidence-driven reasoning.

Collectively, the aforementioned experiments substantiate both the effectiveness and the robust generalization of our methodology across diverse IQA tasks. To further investigate the underlying drivers of these performance gains, we conducted **Grounding Task**. We compare our method against VLMs specifically fine-tuned. Empirical evidence shows that RGPR provides a substantial boost in distortion grounding accuracy, far exceeding the SFT-only baseline. This performance gap underscores the critical role of our grounding process rewards in aligning the model’s internal visual perception with its linguistic reasoning paths, explaining the significant overall gains over conventional supervised fine-tuning.

5.3 Further Analysis

Grounding Process Rewards vs. Outcome Reward To validate the effectiveness of the grounding process rewards, we compare models trained with RGPR against a baseline reinforced with only the outcome reward Eq. (4). As shown in Fig. 4a, the total reward for the baseline (OR) converges faster but plateaus below 1.4, a value significantly lower than that achieved by our method. Furthermore, we examine the average process rewards during training: for baseline, these rewards are not used for policy optimization and remain low, hovering around 0.25. Conversely, our method successfully elevates the average process reward above 0.4, demonstrating that our GPR effectively guides the policy. Evaluation across different training steps Tab. 5) confirms that although the baseline shows initial performance gains, RGPR rapidly surpasses it and maintains superior performance.

Effectiveness of Step-wise Reward Normalization We investigate the impact of Step-wise Reward Normalization (SRN) on training dynamics. Figure 4b shows that incorporating SRN leads to significantly faster convergence and higher final

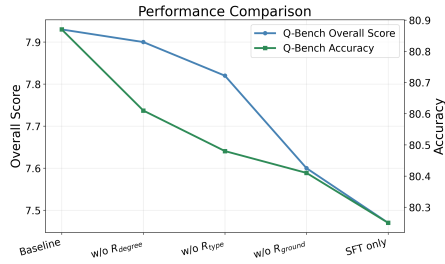


Fig. 5: Ablation study of grounding process reward. w/o R_{ground} means training with outcome reward.

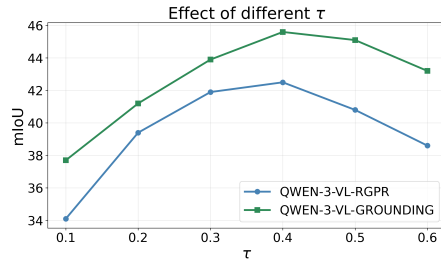


Fig. 6: Effect of different τ . We choose 0.4 as the τ .

convergence reward levels. By independently scaling rewards from heterogeneous reasoning steps, SRN mitigates optimization conflicts and prevents reward aliasing. This is further corroborated by the model’s performance on the LLVisionQA-dev benchmark, underscoring the necessity of SRN in maintaining training stability and enhancing the discriminative precision of the learning signals.

Ablation Study We performed a progressive ablation study by sequentially removing individual process rewards, as summarized in Fig. 5. The results indicate that excluding the (R_{ground}) leads to a drastic degradation in explanation quality, as the model loses the ability to accurately localize distortions. Furthermore, ablating the *type* and *severe* rewards significantly impairs performance on low-level perception tasks. These findings validate that each component of our Grounding Process Reward contributes uniquely to the model’s overall evaluative prowess. Further ablation studies can be found in Appendix.

Effect of different τ We investigate the sensitivity of the RL process to the grounding matching threshold τ , a critical hyperparameter governing the strictness of process supervision. As shown in Fig. 6, excessively high thresholds (e.g., $\tau > 0.5$) induce **reward sparsity**, preventing sufficient positive signals during early training and stalling optimization. Conversely, overly low thresholds (e.g., $\tau < 0.3$) risk **reward hacking**, where the model generates oversized bounding boxes to superficially satisfy IoU criteria without genuine localization. Grounding capability follows a bell-shaped curve, peaking at $\tau = 0.4$, which optimally balances precision and exploration. We adopt $\tau = 0.4$ for all subsequent experiments.

Effectiveness of Trajectory-Level Policy Optimization We conducted comparative experiments against a token-level optimization strategy to verify the stability of our trajectory-level approach. As illustrated in Fig. 7(a), the token-level policy optimization exhibits high fluctuations in the reward curves and fails to ascend stably. Correspondingly, models trained with token-level policy optimization show only marginal performance gains over progressive training steps. Evaluation results on the LLVisionQA-dev benchmark confirm that the trajectory-level weighting is crucial for stable and effective training.

Effect of Different Normalization Function Here, we investigate the effect of different reward normalization functions (F) in Eq. (14), such as $Min(\cdot)$, $Max(\cdot)$, and $Average(\cdot)$. We test the models trained with different F on the quality description task. We find that the maximum function setting leads to a degradation in model capabilities, particularly on the description task, as it allows maximum reward to be obtained simply by producing correct answers—this violates our original intention of establishing process rewards.

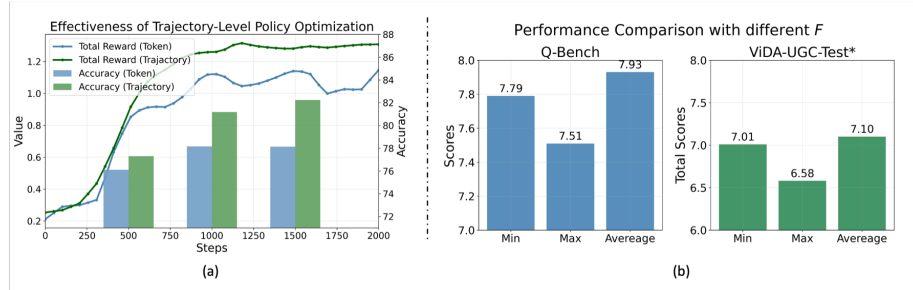


Fig. 7: (a) Rewards using two optimization strategies and model performance comparison across different training steps. (b) Effect of different F . The average function achieves the best results.

6 Conclusions and Limitations

Conclusions In conclusion, this paper presents RGPR, a novel reinforcement fine-tuning framework that aligns VLMs with human reasoning processes for IQA through GPR. By decomposing human assessment into a structured five-step reasoning chain and implementing trajectory-level policy optimization, RGPR achieves fine-grained alignment between model and human reasoning while mitigating training instability. Comprehensive evaluations demonstrate that RGPR substantially improves explanation quality and low-level visual perception, attaining SOTA performance on three benchmarks and competitive zero-shot results. This work establishes a principled approach toward interpretable and reliable IQA systems, bridging the gap between perceptual accuracy and human-understandable reasoning.

Limitations Despite its efficacy, our proposed RGPR is not without inherent constraints. First, while the predefined five-step reasoning chain guarantees reasoning fidelity and interpretability, such structural rigidity inevitably constrains the model’s linguistic flexibility. While highly effective for specialized IQA tasks, this fixed format may lack the stylistic variety required for broader, open-ended human-AI interactions. Second, our framework is primarily optimized for discrete quality categorization rather than continuous scalar regression. Although more granular scores can be derived through post-processing, these quantized

predictions still lack the resolution offered by specialized scoring models. This gap in precision limits our model’s direct utility as a dense reward model for fine-grained ranking tasks. Furthermore, the grounding process rewards rely heavily on comprehensive, fine-grained annotations—including spatial bounding boxes and step-wise reasoning trajectories. This annotation-intensive nature implies significant labeling overhead when migrating the framework to novel domains or specialized industrial scenarios, where such dense supervision may not be readily available.

References

1. Agarwal, V., Shetty, R., Fritz, M.: Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9690–9698 (2020)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Chen, D., Wu, T., Ma, K., Zhang, L.: Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12742–12752 (2025)
4. Chen, Z., Li, Y., Wang, K.: Optimizing reasoning abilities in large language models: A step-by-step approach. Authorea Preprints (2024)
5. Chrisman, L.: Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In: AAAI. pp. 183–188 (1992)
6. Ciancio, A., Da Silva, E.A., Said, A., Samadani, R., Obrador, P., et al.: No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on image processing* **20**(1), 64–75 (2010)
7. Daniel Han, M.H., team, U.: Unsloth (2023), <http://github.com/unslothai/unsloth>
8. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3677–3686 (2020)
9. Fu, Y., Peng, H., Ou, L., Sabharwal, A., Khot, T.: Specializing smaller language models towards multi-step reasoning. In: International Conference on Machine Learning. pp. 10421–10430. PMLR (2023)
10. Gao, F., Tao, D., Gao, X., Li, X.: Learning to rank for blind image quality assessment. *IEEE transactions on neural networks and learning systems* **26**(10), 2275–2290 (2015)
11. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. *IEEE transactions on image processing* **25**(1), 372–387 (2015)
12. Golestaneh, S.A., Dadsetan, S., Kitani, K.M.: No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1220–1230 (2022)
13. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

14. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* **29**, 4041–4056 (2020)
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
16. Jamshidi Avanaki, N., Ghildyal, A., Barman, N., Zadtootaghaj, S.: Lar-iqa: A lightweight, accurate, and robust no-reference image quality assessment model. In: *European Conference on Computer Vision*. pp. 328–345. Springer (2024)
17. Ji, M., Wu, Y., Wu, Z., Wang, S., Yang, J., Dras, M., Naseem, U.: A survey on progress in llm alignment from the perspective of reward design. *arXiv preprint arXiv:2505.02666* (2025)
18. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5148–5157 (2021)
19. Li, W., Zhang, X., Zhao, S., Zhang, Y., Li, J., Zhang, L., Zhang, J.: Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679* (2025)
20. Li, Z., Luo, R., Zhang, J., Qiu, M., Huang, X.J., Wei, Z.: Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. pp. 3769–3798 (2025)
21. Liao, W., Yuan, J., Xu, Y., Guo, C., Zhang, Z., Li, J., Fu, J., Fan, H., Li, T., Cui, J., et al.: Vida-ugc: Detailed image quality analysis via visual distortion assessment for ugc images. *arXiv preprint arXiv:2508.12605* (2025)
22. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let’s verify step by step. In: *The Twelfth International Conference on Learning Representations* (2023)
23. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al.: Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918* (2025)
24. Ma, K., Liu, W., Liu, T., Wang, Z., Tao, D.: dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on image processing* **26**(8), 3951–3964 (2017)
25. Mao, Q., Liu, S., Li, Q., Jeon, G., Kim, H., Camacho, D.: No-reference image quality assessment: Past, present, and future. *Expert Systems* **42**(3), e13842 (2025)
26. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* **21**(12), 4695–4708 (2012)
27. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012)
28. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters* **17**(5), 513–516 (2010)
29. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing* **20**(12), 3350–3364 (2011)
30. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 26374–26383 (2024)

31. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
32. Tu, Z., Chen, C.J., Chen, L.H., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: Regression or classification? new methods to evaluate no-reference picture and video quality models. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2085–2089. IEEE (2021)
33. Wang, B., Lin, R., Lu, K., Yu, L., Zhang, Z., Huang, F., Zheng, C., Dang, K., Fan, Y., Ren, X., et al.: Worldpm: Scaling human preference modeling. arXiv preprint arXiv:2505.10527 (2025)
34. Wang, H., Tu, Y., Liu, X., Tan, H., Liu, H.: Deep ordinal regression framework for no-reference image quality assessment. IEEE Signal Processing Letters **30**, 428–432 (2023)
35. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
36. Wang, W., Gao, Z., Chen, L., Chen, Z., Zhu, J., Zhao, X., Liu, Y., Cao, Y., Ye, S., Zhu, X., et al.: Visualprm: An effective process reward model for multimodal reasoning. arXiv preprint arXiv:2503.10291 (2025)
37. Wang, Y., Li, Z., Zang, Y., Wang, C., Lu, Q., Jin, C., Wang, J.: Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. arXiv preprint arXiv:2505.03318 (2025)
38. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:2309.14181 (2023)
39. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., et al.: Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25490–25500 (2024)
40. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: teaching llms for visual scoring via discrete text-defined levels. In: Proceedings of the 41st International Conference on Machine Learning. pp. 54015–54029 (2024)
41. Wu, T., Zou, J., Liang, J., Zhang, L., Ma, K.: Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. arXiv preprint arXiv:2505.14460 (2025)
42. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2024)
43. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
44. Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., et al.: Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122 (2024)
45. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1191–1200 (2022)
46. You, Z., Cai, X., Gu, J., Xue, T., Dong, C.: Teaching large language models to regress accurate image quality scores using score distribution. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14483–14494 (2025)

47. You, Z., Gu, J., Li, Z., Cai, X., Zhu, K., Dong, C., Xue, T.: Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842* (2024)
48. You, Z., Li, Z., Gu, J., Yin, Z., Xue, T., Dong, C.: Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In: *European Conference on Computer Vision*. pp. 259–276. Springer (2024)
49. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Fan, T., Liu, G., Liu, L., Liu, X., et al.: Dapo: An open-source llm reinforcement learning system at scale. *arXiv e-prints* pp. arXiv–2503 (2025)
50. Zaval, L., Gureckis, T.M.: The impact of perceptual aliasing on exploration and learning in a dynamic decision making task. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. vol. 32 (2010)
51. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Leizhang, Li, C., et al.: Llava-grounding: Grounded visual chat with large multimodal models. In: *European Conference on Computer Vision*. pp. 19–35. Springer (2024)
52. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* **24**(8), 2579–2591 (2015)
53. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Qiao, Y., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In: *European Conference on Computer Vision*. pp. 169–186. Springer (2024)
54. Zhang, R., Tong, C., Zhao, Z., Guo, Z., Zhang, H., Zhang, M., Liu, J., Gao, P., Li, H.: Let’s verify and reinforce image generation step by step. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 28662–28672 (2025)
55. Zhang, W., Ma, K., Zhai, G., Yang, X.: Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing* **30**, 3474–3486 (2021)
56. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: 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*. pp. 14071–14081 (2023)
57. Zhang, Z., Zheng, C., Wu, Y., Zhang, B., Lin, R., Yu, B., Liu, D., Zhou, J., Lin, J.: The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301* (2025)
58. Zhang, Z., Kou, T., Wang, S., Li, C., Sun, W., Wang, W., Li, X., Wang, Z., Cao, X., Min, X., et al.: Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 10621–10631 (2025)
59. Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., Zhou, W., Chen, Y.: Swift:a scalable lightweight infrastructure for fine-tuning (2024), <https://arxiv.org/abs/2408.05517>
60. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al.: Group sequence policy optimization. *arXiv preprint arXiv:2507.18071* (2025)
61. Zhong, Y., Wu, X., Zhang, L., Yang, C., Jiang, T.: Causal-iqa: Towards the generalization of image quality assessment based on causal inference. In: *Forty-first International Conference on Machine Learning* (2024)
62. Zhu, H., Wu, H., Li, Y., Zhang, Z., Chen, B., Zhu, L., Fang, Y., Zhai, G., Lin, W., Wang, S.: Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems* **37**, 32611–32629 (2024)