Refine-IQA: Multi-Stage Reinforcement Finetuning for Perceptual Image Quality Assessment

Ziheng Jia[1], Jiaying Qian[1], Zicheng Zhang[2], Zijian Chen[1], Xiongkuo Min[1]*

[1]Shanghai Jiao Tong University
[2]Shanghai Artificial Intelligence Laboratory
jzhws1@sjtu.edu.cn

Abstract

Reinforcement fine-tuning (RFT) is an advancing paradigm for LMM training. Analogous to high-level tasks, RFT is also applicable to low-level vision domains, including image quality assessment (IQA). Existing RFT-based IQA methods typically use rule-based rewards to verify the model's rollouts but provide no specific supervision for the "think" process. Furthermore, these methods typically fine-tune directly on downstream IQA tasks without explicitly enhancing the model's native low-level quality perception, which may constrain its performance upper bound. In response to these gaps, we propose a multi-stage RFT framework for IQA (**Refine-IQA**). In *Stage-1*, we build the **Refine-Perception-20K** dataset (with 12 main distortions, $20,907$ locally-distorted images, and over $55K$ RFT samples) and design multi-task reward functions to strengthen the model's visual quality perception. In *Stage-2*, targeting the quality scoring task, we introduce a **probability difference reward involved strategy** for "think" process supervision. The resulting **Refine-IQA Series Models** achieve outstanding performance on both perception and scoring tasks—and, notably, our paradigm activates a robust "think" (quality-interpretating) capability that also attains remarkable results on the quality interpreting task.

Code —
https://github.com/jzhws/VisualQualityAssessment-RL

Introduction

Critic-model-free reinforcement learning (RL) algorithms such as *REINFORCE Leave-One-Out* (*RLOO*) (Ahmadian et al. 2024) and *Group Relative Policy Optimization* (*GRPO*) (Shao et al. 2024) have given rise to more efficient reinforcement finetuning (RFT) paradigms. These approaches utilize the model's reward across a group of sampled outputs (rollouts) to construct intra-group advantages, thereby directly guiding the policy gradient and minimizing the dependence on label-intensive offline instruction data. Consequently, this paradigm has seen widespread use in high-level large language model (LLM) and large multimodal model (LMM) reasoning tasks such as mathematical reasoning (Shao et al. 2024; Ren et al. 2025; Wang et al. 2025a), code generation (Wang et al. 2025b), and image/video understanding (Yu et al. 2025a; Huang et al. 2025; Li et al. 2025b; Zhang et al. 2025a,b). Likewise, it exhibits promise in the low-level vision domain, with one prominent application being perceptual image quality assessment (IQA). Most existing LMM-IQA works focus on supervised fine-tuning (SFT), explicitly **teaching** the model to enhance its visual-quality assessment capability using offline-annotated data. A major drawback of SFT is its tendency to cause **overfitting**, compromising the model's versatility and its adherence to complex instructions. One example is depicted in the upper left of Fig. 1. Introducing RFT substantially mitigates overfitting, enables unrestrained policy exploration, and raises the upper bound of IQA performance. These merits make it a new training paradigm for this field.

The prevalent "think-answer" output paradigm of LLMs in RFT is also applicable to the IQA domain. According to the classic "perception-decision" mechanism (Mazurek et al. 2003) in the human visual system (HVS)—where the process of "perceiving visual quality features (perception)" followed by "producing a quantitative quality score (decision)" can be viewed as a close analogue to the "think-answer" process. Specifically, the LMM generates the **interpretation** of the input image's visual quality along with its reasoning when "thinking"; subsequently, it outputs the quantitative score prediction, aligned with the image's subjective mean opinion score (MOS) when "answering". At this juncture, a natural question arises: *How can we ensure that the "think" process in RFT for IQA is genuinely reliable and effective?*

Unlike high-level reasoning tasks, low-level image quality perception is an **implicit**, **intuition-based** process. Firstly, humans do not follow a pre-defined "think" pathway when judging image visual quality. Moreover, we have surprisingly observed a "**think collapse**" phenomenon in standard *GRPO* training on the quality scoring task with rule-based outcome reward: **as the training progresses, the length of the "think" process rapidly collapses, while the model's quality-scoring performance continues to improve (visualized in Supplementary Material (*Supp.*)).** The above findings and analysis underscore two primary technical challenges: (1) In IQA tasks, supervising the "think" process through a predefined, rule-based process reward model (PRM) is challenging. (2) Without proper reward supervision, the "think" process in the quality scoring task deterio-

*Corresponding author.

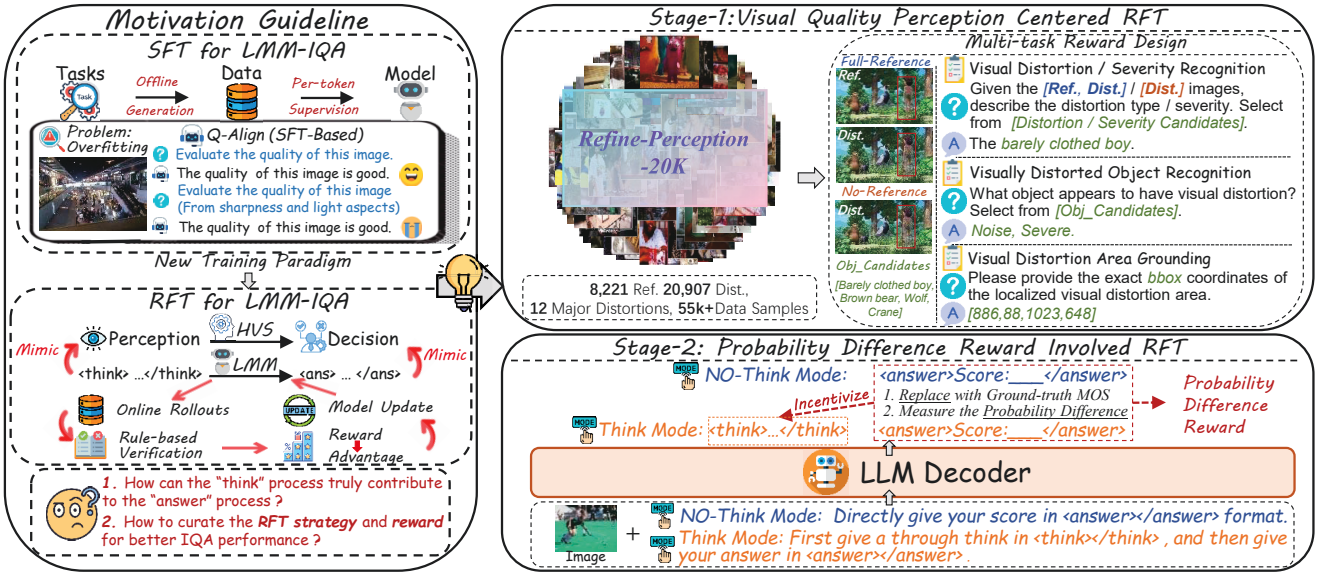

Figure 1: Overview of the *Refine-IQA*. Our primary motivation is to **curate an RFT framework that incorporates an effective "think" process, meanwhile optimizing the IQA performance**. To address these issues, we curate the *Refine-IQA*. *Stage-1* enhances the model's inherent visual quality perception by training on the *Refine-Perception-20K* dataset with the multi-task reward system. *Stage-2* training incentivizes the "think" process through the involvement of the probability difference reward.

rates over time, contributing minimally to the final decision. Moreover, we hypothesize that the "think" process serves as a **visual quality perception recalibration**, facilitating fine-grained **visual quality interpretation**. Therefore, effectively leveraging this side-effect remains a critical focus of our research.

Motivated by these challenges, we propose a **multi-stage reinforcement fine-tuning** paradigm for constructing LMM with *refined* IQA expertise (***Refine-IQA***). The overview is shown in Fig. 1. Our core contributions are as follows:

1. We construct the **Refine–Perception–20K** dataset, the **first** RFT dataset meticulously for enhancing LMM's **native visual quality perception**. It spans 12 primary distortion categories and comprises over $20,000$ images from diverse scene contexts, each containing different *synthetic* and *localized* visual distortions. To guarantee data quality, we implement a *human-in-the-loop data scrutiny* that verifies both the semantic consistency and the perceptual clarity of the data.
2. We propose an efficient, multi-stage RFT strategy for IQA-expert LMM. Building on the *Refine-Perception-20K* dataset, we implement a multi-task reward scheme that comprehensively enhances the model's sensitivity to low-level visual distortions (*Stage-1*). Subsequently, for the quality-decision (scoring) task, we introduce a **probability difference reward involved** strategy that implicitly supervises the "think" process by measuring the difference in ground-truth output probabilities between the "think" and "no think" modes (*Stage-2*).
3. Leveraging the curated datasets and training strategy, we develop the *Refine-IQA Series Models*, which demonstrate robust performance on quality scoring tasks across 6 IQA datasets with varied scenarios. Furthermore, these models excel in qualitative quality interpretation: with just about $13K$ images for quality-scoring RFT, they perform competitively with large-scale SFT-based LMMs on the quality interpretation task.

Related Works

LMM for IQA

Studies have already leveraged LMMs for IQA tasks. *Q-Align* (Wu et al. 2024c) lays the foundation of the LMM-based quality scoring using the log-probability estimation strategy. *Compare2Score* (Zhu et al. 2024) tackles subjective label scarcity using pairwise preference relationships as pseudo-labels. *Q-Instruct* (Wu et al. 2024b) and *Aes-expert* (Huang et al. 2024) pioneer in training LMMs with qualitative image quality interpretation capabilities in image technical and aesthetics quality assessment, respectively. *Co-instruct* (Wu et al. 2024d) and *DepictQA* (You et al. 2024b,a) focus on image pair quality analysis tasks.

While these approaches effectively address downstream IQA tasks, they share a common limitation: being entirely trained on SFT, they experience substantial degradation in multi-task versatility and adherence to complex instructions.

RFT for IQA

Recently, research has begun to involve RL strategies for LMM-IQA. *Q-Insight* (Li et al. 2025a) employs the standard rule-based outcome reward for multi-task RFT. *Q-Ponder* (Cai et al. 2025) follows the "Cold-start to RL" workflow to construct a comprehensive training pipeline.

Although these works represent valuable advances, they share some limitations. First, they do not explicitly enhance the LMM's visual quality perception; instead, they directly fine-tune the model on downstream tasks — potentially constraining its performance ceiling. More critically, these mod-

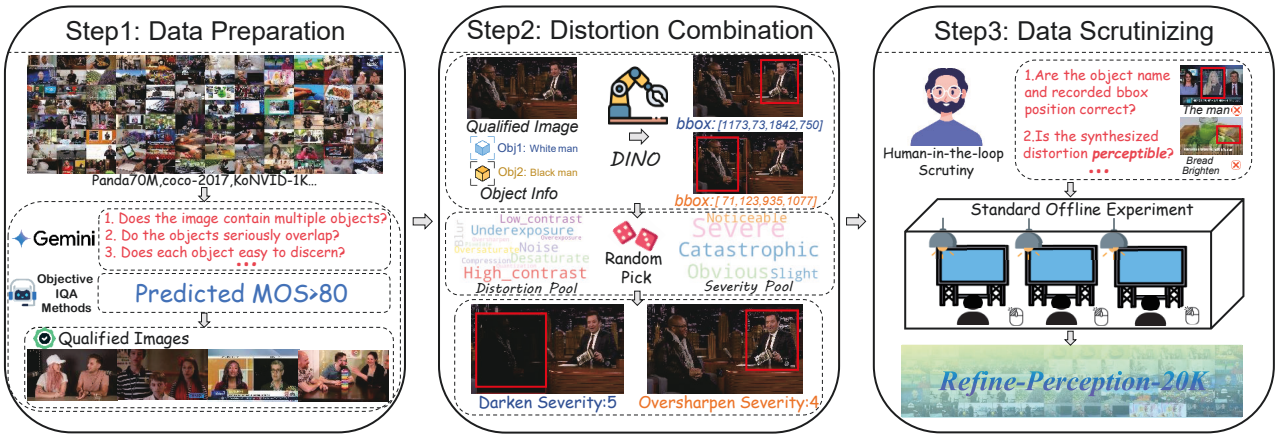

Figure 2: The construction pipeline of the *Refine-Perception-20K* dataset.

els fail to incorporate reward supervision for the "think" process, thereby reducing it to an ancillary output. These shortcomings present insights for our work.

The Refine-IQA

To design an RFT framework that supports effective "think" process and empowers the LMM with refined IQA performance, a key focus is enhancing the model's native ability to perceive low-level visual quality features - specifically, its ability to **identify fundamental visual distortion types, severity, and location**. Simultaneously, it is also crucial to establish a reward scheme that **bridges the gap between the "think" and "answer" processes**. In response to this, we develop *Refine-IQA*, a comprehensive RFT paradigm.

Visual Quality Perception Centered RL

In *Stage-1*, we have constructed the *Refine-Perception-20K*, a dataset of $20,907$ images, encompassing 12 distortion types and 5 levels of distortion severity. The construction pipeline is demonstrated in Fig. 2. The detailed statistical information is recorded in *Supp.*.

Dataset Construction Pipeline The dataset construction commences with data preparation. To ensure that the source image pool enables the identification of multiple objects, the data source needs to contain relatively complex semantic content. Therefore, we extract keyframes from the video datasets *LSVQ* (Ying et al. 2021) and *Panda-70M* (Chen et al. 2024) and select images from the *COCO* object-detection dataset (Lin et al. 2014), resulting in over $100,000$ images across varying resolutions and contexts. To mitigate the effects of inherent visual distortions, three state-of-the-art objective IQA models—*Q-Align* (Wu et al. 2024c), *DBCNN* (Zhang et al. 2020), and *Hyper-IQA* (Su et al. 2020)—are employed to compute the quality score for each image, with only those with all three scores above 80 retained. Next, we utilize *Gemini-2.5-PRO* (Team et al. 2024) to filter images containing at least two distinct semantic objects, recording a concise phrase-level description for each identified object. This process yields $8,221$ qualified images along with their semantic object information.

Subsequently, we employ *DINO* (Ren et al. 2024) to perform object detection on the qualified images using the recorded information. For each source image, a single bounding box (bbox)—specified by its top-left $(x_{\mathrm{TL}}, y_{\mathrm{TL}})$ and bottom-right $(x_{\mathrm{BR}}, y_{\mathrm{BR}})$ coordinates—is generated for each annotated object. This procedure results in over $30K$ marked images, each containing exactly one bbox. We then randomly apply distortions of varying severity to the bbox regions. The **distortion pool** comprise *blur*, *noise*, *compression*, *overexposure*, *underexposure*, *high contrast*, *low contrast*, *oversaturate*, *desaturate*, *oversharpen*, *pixelate*, and *quantization*, and the **severity candidates** encompass *slight*, *just noticeable*, *relatively obvious*, *severe*, and *very severe*. We partially adopt the distortion combination method from *DepictQA-V2* (You et al. 2024b), which is detailed in *Supp.*.

As the perceptibility of synthetic distortions depends on both content and visual characteristics, human-in-the-loop scrutiny is essential. Therefore, we conduct a subjective experiment in a standard laboratory environment. Human experts are instructed to exclude any images where the bbox content fails to match the description or where the added distortion is not perceptible. The subjective experiment settings are given in *Supp.*. Finally, we extract $1,500$ images for test (denoted as *Refine-Perception-20K-test*), with the remaining part used for training (*Refine-Perception-20K-train*). The distortion distribution in both parts is kept consistent.

Multi-Task Reward Design We employ *Qwen2.5-VL-7B* (Bai et al. 2025) as the base model. Considering that it has already obtained inherent capability for low-level visual perception (i.e., the ability to recognize commonly-seen distortion types), our objective is to **calibrate** and **refine** such ability through RFT. Thus, we employ a multi-task, rule-based reward schema in conjunction with standard *GRPO* for RFT. Here we define the fundamental sub-tasks:

1. **Visual Distortion Type / Severity Recognition** To ensure verifiability, this task is formulated as a multi-choice (single-answer) problem. The distortion and severity pool serve as candidate choice sets; the model receives a reward 1 only if its output matches the correct answer.

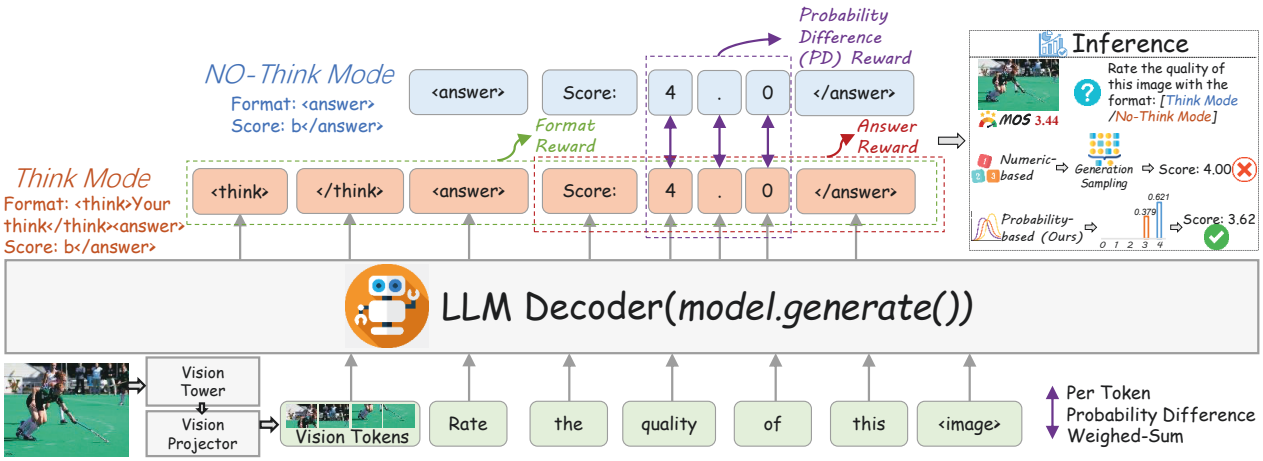

Figure 3: The model structure, our inference method, and the illustration of the PD reward involved RL strategy.

2. **Visually Distorted (Semantic) Object Recognition** Also arranged as a multi-choice task. Candidate answers are drawn from the recorded semantic objects of the original image. The verification rule is identical to the above.
3. **Visual Distortion Area Grounding** The model outputs the coordinates of the bbox enclosing the distortion area. The reward is computed as the Intersection-over-Union (*IoU*) between the predicted and ground-truth bboxs.

All three tasks include the "full-reference" (FR) (with the original and distorted images) and "no-reference" (NR) (only with the distorted image) sub-tasks (examples are shown in the upper right of Fig. 1 and detailed in *Supp.*). Finally, we obtain over $55K$ RFT data samples. During training, data from all three tasks are randomly mixed.

Probability Difference Reward Involved RL

For the quality-decision (scoring) task, the primary focus is to ensure that the "think" process is genuinely effective while optimizing its alignment with the ground-truth MOS. Accordingly, we propose the *probability difference reward involved* RFT strategy (depicted in detail in Fig. 3).

The "Think Collapse" Phenomenon We use the outcome answer reward r_{ans} and standard *GRPO* setting for the quality scoring task. The r_{ans} is denoted as:

$$r_{\text{ans}} = 1 \quad \text{if} \quad |\hat{s} - s| < \epsilon, \text{ otherwise } 0, \tag{1}$$

where $\hat{s}$ and s are the predicted and ground-truth scores on a $[0, 5)$ scale, and ϵ is set to 0.5. During training, we observe the above-mentioned "**think collapse**" phenomenon. We provide further explanation in *Supp.*.

The Probability Difference Reward To address this issue, we introduce the *probability difference (PD) reward*, which provides implicit supervision on the "think" process.

LLMs inherently model text generation as **next-token probability prediction**. In the "think-answer" paradigm, the increase in the probability of the ground-truth—relative to the "no-think" mode—signals a greater contribution from the "think" process. This motivates our use of the difference in ground-truth token probabilities (likelihoods) between the "think" and "no-think" modes as the reward signal. Given the ground-truth MOS with $M-1$ decimal places (we assign $M = 2$), we set $\boldsymbol{z}^{\text{think}} \in \mathbb{R}^M$ and $\boldsymbol{z}^{\text{no-think}} \in \mathbb{R}^M$ as the predicted probabilities (*softmax normalization to the vocab logits*) corresponding to each digit of the ground-truth MOS under the "think" and "no-think" modes. We derive the reward for the two modes through the weighted sum of the token probability of each ground-truth MOS digit:

$$r_{\text{MODE}} = w_0 \boldsymbol{z}_0^{\text{MODE}} + w_1 \boldsymbol{z}_1^{\text{MODE}} + \cdots + w_{M-1} \boldsymbol{z}_{M-1}^{\text{MODE}}, \tag{2}$$

where the placeholder MODE can be replaced by think or no−think. The weights can be formulated as $\boldsymbol{w} = [w_0, w_2, \cdots, w_{M-1}] = [1, 0.1, \cdots, 10^{1-M}]$. For the i-th rollout in one group, its probablity difference reward r_{pd}^i is:

$$r_{\text{pd}}^i = \text{clip}\left(r_{\text{think}}^i - r_{\text{no-think}}^i, 0, 1\right). \tag{3}$$

Specifically, we employ two distinct prompts corresponding to the two modes to elicit rollouts. Thereafter, we replace the model's predicted score with the ground-truth MOS (with the same digit numbers) and calculate the probability of each corresponding ground-truth token. It is important to note that the output format of the "no-think" mode is fixed, resulting in the invariant likelihoods for the ground-truth MOS tokens within each group; hence, the $r_{\text{no-think}}^i$ can be denoted as r_{ref} and r_{pd}^i can be reformulated as:

$$r_{\text{pd}}^i = \text{clip}\left(r_{\text{think}}^i - r_{\text{ref}}, 0, 1\right). \tag{4}$$

The final reward r_{final}^i is defined as the weighted sum of the answer reward r_{ans}^i (same as the setting in Eq. 1), the format reward r_{fmt}^i (the reward is 1 only if the output format follows the requirement and otherwise 0) and the r_{pd}^i:

$$\begin{aligned} r_{\text{final}}^i = \lambda_1 r_{\text{ans}}^i + \lambda_2 r_{\text{fmt}}^i + \lambda_3 r_{\text{pd}}^i \ \ if \ \ r_{\text{ans}}^i = r_{\text{fmt}}^i = 1, \\ otherwise \ \ \lambda_1 r_{\text{ans}}^i + \lambda_2 r_{\text{fmt}}^i, \end{aligned} \tag{5}$$

we set $\lambda_1 = \lambda_2 = \lambda_3 = 1$. The rationale for our reward design can be summarized as follows: Overall, the r_{pd}

functions as an **incremental incentive**. When the model's scoring accuracy for one training sample (quantified by the proportion of qualified rollouts with $r_{\text{ans}} = r_{\text{fmt}} = 1$ in one group) is low, the training concentrates on improving the accuracy. Only when accuracy is sufficiently high do rollouts with higher PD rewards provide additional advantage. This ensures correct predictions are effectively utilized, while preventing training instability.

Modifying the Gradient Policy for Quality Scoring In *GRPO*, the intra-group advantage calculation is prone to assigning negative advantages to incorrect rollouts, leading to decreased probabilities. For the quality scoring task, this may result in overly definite outputs, which impairs the generalization ability. To address this, we draw inspiration from *Q-Align* and reformulate the scoring task as a **quality level classification (distribution prediction)** task. We define the modified group-relative (with G rollouts) advantage as:

$$\begin{aligned}
&\hat{A}_{i,t} = \max\left[\frac{r^{(i)} - \text{mean}\left(\left\{r^{(1)}, r^{(2)} \ldots, r^{(G)}\right\}\right)}{\text{std}\left(\left\{r^{(1)}, r^{(2)} \ldots, r^{(G)}\right\}\right)}, 0\right] \\
&if\ \ 0 \leq |\lambda_1 r_{\text{ans}} + \lambda_2 r_{\text{fmt}} < \lambda_1 + \lambda_2| < G, \\
&otherwise\ \hat{A}_{i,t} = \frac{r^{(i)} - \text{mean}\left(\left\{r^{(1)}, r^{(2)} \ldots, r^{(G)}\right\}\right) - \epsilon}{\text{std}\left(\left\{r^{(1)}, r^{(2)} \ldots, r^{(G)}\right\}\right)},
\end{aligned} \tag{6}$$

where ϵ is set to 0.02. Following techniques proposed in *DAPO* (Yu et al. 2025b), we remove the KL penalty and shift from *sample-level* averaging to *token-level* averaging in the loss processing to promote a longer "think" process. Finally, let V be an input image and q be its scoring prompt. The modified *GRPO* optimization objective is denoted as:

$$\begin{aligned}
&\mathcal{J}(\theta) = \mathbb{E}\left[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O \mid q, V)\right] \\
&\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left\{\min\left[\rho_{i,t}\hat{A}_{i,t}, \text{clip}\left(\rho_{i,t}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right]\right\},
\end{aligned} \tag{7}$$

where Q denotes the input prompt set and $\rho_{i,t}$ represents the importance sampling coefficient $\frac{\pi_\theta\left(o_{i,t}|q,V,o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t}|q,V,o_{i,<t}\right)}$. In an on-policy setting without importance sampling, this configuration mirrors a **cross-entropy optimization target** with **dynamic gradient weights** (the detailed proof is presented in *Supp.*). Correct rollouts of training samples that are difficult to score accurately receive higher advantages, driving the model to prioritize their optimization. If no rollout in a group is correct ($|\lambda_1 r_{\text{ans}} + \lambda_2 r_{\text{fmt}} < \lambda_1 + \lambda_2| = G$), a negative advantage is employed to prevent training data waste and encourage exploration of alternative predictions.

Inference Method We adopt the following **probablity-based expectation estimation** (shown in Fig. 3) to obtain the score $\mathcal{Q}$ from output vocab logits in the inference stage:

$$\mathcal{Q} = \sum_{j=0}^{4} j \frac{e^{\mathcal{P}_j}}{\sum_{i=0}^{4} e^{\mathcal{P}_i}} + 0.5, \tag{8}$$

where $\mathcal{P}$ denotes the model's **logits** of the corresponding digit values on the output **integer part**.

The Refine-IQA Series Models

Building upon the *Refine-IQA*, we derive the *Refine-IQA Series Models*. In *Stage-1*, a single epoch training in "no-think" mode on *Refine-Perception-20K-train* produces *Refine-IQA-S1*. Subsequently, we advance to *Stage-2*, performing 3 training epochs in "think" mode on the combined dataset of *KonIQ-10K* (Hosu et al. 2020) and *SPAQ* (Fang et al. 2020) training sets (approximately $13K$ samples, the ground-truth MOS of all training data is uniformly scaled to the range $[0, 5)$) to produce *Refine-IQA-S2*. The training prompts design is shown in *Supp.*. All experiments are conducted on 8 NVIDIA A100 (80 GB) GPUs, with G set to 8.

Experiments

To rigorously assess the performance of the *Refine-IQA Series Models*, we perform extensive comparison experiments. Additionally, we carry out ablation studies on the critical configurations to enable more in-depth analysis.

Performance on Image Quality Perception

We evaluate the models' visual quality perception capability using the *Refine-Perception-20K-test*. First, we partition the test data by distortion type based on their prevalence in natural scenes into two categories: **Easy** and **Hard**. We then assess the two sub-tasks: (1) **Description**: For each image, the model identifies the distortion category, the distorted semantic object, and the distortion severity from the candidate pools. A test case is deemed correct only if all three attributes are correctly chosen. We record the overall accuracy as the experimental result. (2) **Grounding**: We compute the *IoU* between the model's predicted and the ground-truth region. Since no open-source quality perception LMM is currently available, we only compare against the base model (all models use the "no-think" mode). The results for the two tasks are shown in the Tabs. 1 and 2. They indicate that after RFT, the model exhibits significant improvements in performance on both tasks, particularly in **Hard** cases and the **Grounding** task. This suggests that, although RFT does not introduce new knowledge, it effectively guides the model to recalibrate its native visual quality perception capabilities.

Performance on Image Quality Scoring

We select in-domain (in-the-wild) IQA datasets *KonIQ* (*test set*), *SPAQ* (*test set*), and *LIVE-C* (Ghadiyaram and Bovik 2015), alongside out-of-domain datasets *AGIQA-3K* (Li et al. 2023) (AIGC-images), *KADID-10K* (Lin, Hosu, and Saupe 2019), and *CSIQ* (Larson and Chandler 2010) (synthetic distortions) for evaluation. We carefully search the comparison models **with open-source training code** for reproduction. Except for the *Q-Align series* and *Compare2Score*, in which we use pre-trained LMMs, all comparison models are retrained on the same training dataset as ours (for *Q-Insight*, we adopt the training setup with only the quality scoring task). The evaluation metrics are the commonly used *Pearson Linear Correlation Coefficient* (PLCC) and *Spearman Rank Correlation Coefficient* (SRCC). The results are shown in Tab. 3. The *Refine-IQA-S2* achieves excellent performance across all six datasets under both

Categories	Easy					Hard							Overall
Models	*Blur*	*Noise*	*Comp.*	*OE*	*UE*	*OSat.*	*DSat.*	*OSharp.*	*HC*	*LC*	*Pixel.*	*Quant.*	
Qwen2.5-VL-*7B* (base)	75.47%	72.13%	78.23%	74.18%	69.32%	31.25%	23.08%	33.39%	29.18%	24.24%	15.16%	9.32%	51.25%
Refine-IQA-S1	**91.20%**	**87.81%**	**88.83%**	**96.97%**	**85.42%**	**70.39%**	**65.32%**	**58.23%**	**70.25%**	**72.18%**	48.45%	**37.12%**	**76.37%**
Refine-IQA-S2	85.73%	86.49%	88.10%	94.58%	83.25%	65.48%	65.32%	57.45%	64.36%	68.73%	**50.42%**	34.23%	73.18%
Advantage	*15.73%*	*15.68%*	*10.60%*	*22.79%*	*16.10%*	*39.14%*	*42.24%*	*24.84%*	*41.07%*	*47.94%*	*33.29%*	*27.80%*	*25.12%*

Table 1: *Description* performance on image quality perception task on the *Refine-Perception-20K-test* (where *Comp.*, *OE*, *UE*, *OSat.*, *DSat.*, *OSharp.*, *HC*, *LC*, *Pixel.*, and *Quant.* represent *compression*, *overexposure*, *underexposure oversaturation*, *desaturation*, *oversharpening*, *high contrast*, *low contrast*, *pixelation*, and *quantization*, respectively). *Advantage* denotes the performance improvement of the *Refine-IQA-S1* compared to the *base model*. [Per column: highest in **bold**.]

Categories	Easy					Hard							Overall
Models	*Blur*	*Noise*	*Comp.*	*OE*	*UE*	*OSat.*	*DSat.*	*OSharp.*	*HC*	*LC*	*Pixel.*	*Quant.*	
Qwen2.5-VL-*7B* (base)	0.542	0.513	0.489	0.403	0.398	0.298	0.245	0.283	0.192	0.332	0.352	0.143	0.343
Refine-IQA-S1	**0.952**	**0.943**	**0.965**	**0.982**	**0.885**	**0.752**	**0.788**	**0.679**	**0.737**	**0.685**	**0.483**	**0.315**	**0.772**
Refine-IQA-S2	0.898	0.903	0.869	0.937	0.832	0.705	0.723	0.632	0.658	0.617	0.392	0.252	0.738
Advantage	*0.410*	*0.430*	*0.476*	*0.579*	*0.487*	*0.454*	*0.543*	*0.396*	*0.545*	*0.353*	*0.131*	*0.172*	*0.429*

Table 2: *Grounding* performance on image quality perception task. [Per column: highest in **bold**.]

Datasets	KonIQ		SPAQ		LIVE-C		AGIQA-3K		KADID-10K		CSIQ	
Models	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*
Performance of Deep Neural Network (DNN)-Based Models												
NIMA (Talebi and Milanfar 2018)	0.859	0.896	0.856	0.838	0.771	0.814	0.654	0.715	0.535	0.532	0.662	0.683
DBCNN (Zhang et al. 2020)	0.875	0.884	0.806	0.812	0.755	0.773	0.641	0.730	0.484	0.497	0.552	0.589
HyperIQA (Su et al. 2020)	0.906	0.917	0.788	0.791	0.749	0.772	0.640	0.702	0.468	0.506	0.631	0.685
MUSIQ (Ke et al. 2021)	0.919	0.914	0.863	0.868	0.830	0.789	0.630	0.722	0.556	0.575	0.642	0.698
CLIP-IQA+ (Wang, Chan, and Loy 2022)	0.895	0.909	0.864	0.866	0.805	0.832	0.685	0.736	0.654	0.653	0.695	0.710
LIQE (Zhang et al. 2023)	0.928	0.912	0.833	0.846	*0.870*	0.830	0.708	0.772	0.662	0.667	0.703	0.715
Performance of SFT-Based Domain-Specific LMMs												
Q-Align-*IQA-7B* (Wu et al. 2023)	0.920	0.918	0.897	0.896	0.860	0.853	0.735	0.772	0.684	0.674	0.668	0.701
Q-Align-*Onealign-7B*	*0.933*	**0.930**	0.915	0.908	0.868	0.872	0.758	0.801	**0.712**	**0.725**	0.683	0.715
Compare2Score-*7B* (Zhu et al. 2024)	0.915	0.905	0.924	0.914	0.808	0.787	0.765	0.703	0.532	0.585	0.686	0.700
Performance of RL-Based Domain-Specific LMMs												
Q-Insight-*7B* (Li et al. 2025a) (w *think*)	0.860	0.873	0.873	0.881	0.785	0.824	0.758	0.785	0.573	0.591	0.665	0.692
Q-Insight-*7B* (w/o *think*)	0.871	0.894	0.899	0.902	0.801	0.833	0.777	0.832	0.580	0.584	0.683	0.714
Performance of Refine-IQA Series Models												
Refine-IQA-S2 (*w think*)	0.920	0.916	**0.930**	*0.918*	**0.870**	**0.892**	*0.789*	*0.835*	*0.703*	*0.715*	*0.711*	*0.739*
Refine-IQA-S2 (*w/o think*)	**0.938**	*0.924*	*0.927*	**0.921**	0.860	*0.885*	**0.798**	**0.841**	0.698	0.702	**0.724**	**0.758**

Table 3: Evaluation results on image quality scoring task (*w think* and *w/o think* denote that the model turns on / off the "think" mode during evaluation. Except where specifically indicated, all other models are tested using the "no think" mode). [Per column: highest in **bold**, second in *italic*, third in underlined.]

Sub-categories	Question Types			Quality Concerns				Overall
Models	*Binary*	*What*	*How*	*Technical*	*Other*	*In-context Technical*	*In-context Other*	
Performance of Open-sourced General LMMs								
mPLUG-Owl3-*7B* (Ye et al. 2024)	78.72%	79.77%	67.45%	73.44%	71.74%	71.19%	84.89%	74.21%
InternVL3-*8B* (Zhu et al. 2025)	78.28%	81.56%	69.95%	70.82%	79.23%	73.97%	*86.69%*	76.58%
LLaVA-Onevision-*7B* (Li et al. 2024)	79.12%	78.19%	69.73%	70.06%	76.54%	73.11%	83.01%	74.68%
Qwen2-VL-*7B* (Wang et al. 2024)	81.56%	79.60%	72.63%	73.89%	79.95%	75.00%	**86.69%**	78.06%
Qwen2.5-VL-*7B* (Bai et al. 2025) (*w/o think*)	80.47%	84.81%	69.95%	76.19%	79.47%	77.39%	82.12%	78.39%
Qwen2.5-VL-*7B* (*w think*)	80.58%	84.37%	71.35%	78.00%	79.58%	78.25%	81.93%	79.51%
Performance of Proprietary General LMMs								
GPT-4o (Achiam et al. 2023)	82.48%	83.94%	70.16%	76.00%	**80.19%**	79.45%	82.12%	78.92%
Claude-3.7-Sonnet (Anthropic 2025)	74.08%	78.95%	66.46%	70.05%	75.65%	68.83%	79.84%	73.11%
Performance of Open-sourced Domain Specific LMMs								
Q-Align-*Onealign-7B*	67.51%	58.83%	56.94%	60.85%	70.13%	55.28%	69.35%	62.22%
Q-Instruct (*LLaVA-1.5*)-*13B* (Wu et al. 2024b)	80.66%	67.25%	61.93%	66.03%	70.41%	69.86%	79.85%	70.43%
Q-Instruct (*Qwen-2.5-VL*)-*7B* (*SFT-based reference model*)	*83.94%*	*85.29%*	**73.01%**	*81.71%*	79.89%	**83.07%**	84.04%	**81.52%**
Q-Insight-*7B* (*w think*)	80.80%	83.18%	72.60%	77.52%	79.21%	78.13%	82.23%	78.86%
Q-Insight-*7B* (*w/o think*)	81.20%	83.24%	71.60%	78.13%	78.75%	78.42%	82.50%	79.01%
Performance of Refine-IQA Series Models								
Refine-IQA-S1	81.75%	85.24%	72.63%	78.50%	79.71%	80.13%	82.50%	79.86%
Refine-IQA-S2 (*w/o think*)	**84.11%**	84.90%	72.57%	81.27%	79.80%	81.10%	83.26%	80.78%
Refine-IQA-S2 (*w think*)	83.38%	**86.16%**	*72.92%*	**82.35%**	*80.10%*	*82.13%*	82.87%	*81.48%*

Table 4: Evaluation results on the *Q-Bench-test*. [Per column: highest in **bold**, second in *italic*, third underlined.]

Train	Scoring				Interpret	
	KonIQ	*AGIQA*	*SPAQ*	*KADID*	*Tech.*	*Overall*
w/o *Stage-1*	0.916	0.815	0.920	0.671	80.10%	79.85%
w *Stage-1*	**0.931**	**0.817**	**0.924**	**0.709**	**81.27%**	**80.78%**

Table 5: Ablation study of the effects of *Stage-1* (using the *Refine-IQA-S2 (w/o think)*. **Scoring** performance is represented as the average of *SRCC* and *PLCC* on the individual dataset. The **Interpreting** performance is reported as the performance on the *Technical (Tech.)* and *Overall* dimensions. [Per column: highest in **bold**.]

"think" and "no-think" modes. Notably, the negligible performance gap between the two modes (compared to *Q-Insight*) underscores the effectiveness of the "think" process.

Performance on Image Quality Interpreting

As previously discussed, we believe that developing an effective "think" process in the quality scoring task RFT can improve the LMM's performance in quality interpretation (further explained in *Supp.*). To validate this, we choose the *Q-bench-test* (Wu et al. 2024a) (with $1,495$ multi-choice (single answer) questions). Here, we select the latest open-sourced and proprietary general LMMs, along with some high-performing IQA-LMMs for comparison. All models are evaluated using the *model.generate()* mode with *greedy search* to ensure reproducibility. The experiment results are shown in Tab.4, from which we have some critical observations: (1) After the perception-centered RFT *Stage-1*, the *Refine-IQA-S1* already demonstrates improved performance. (2) The *Refine-IQA-S2* achieves substantial gains under both "think" and "no-think" modes (particularly in the *Technical* dimension)—significantly outperforming general LMMs and nearly matching the fine-tuned base model (with SFT) using the *Q-Pathway*-$200K$ *(Q-Instruct)* (Wu et al. 2024b). This indicates that the "think" process in *Stage-2* RFT further refines the model's visual quality interpreting ability.

Ablation Study

The Effects of *Stage-1* Training We alternatively remove the *Stage-1* training, with all other training and evaluation settings kept consistent. The performance of this ablation experiment is shown in Tab. 5.

The results demonstrate that the *Stage-1* training positively optimizes the model's performance on both scoring and interpreting tasks, highlighting the importance of enhancing the model's inherent quality perception capabilities.

***Stage-2* Attributes Ablation Study** We ablate key attributes in *Stage-2*, while keeping other settings the same. The ablation results are presented in Tab. 6. Here, we present some key findings: (1) After removing the *PD* reward, the model exhibits the "think collapse", and performance on the quality interpreting task significantly decreases, highlighting that reward supervision for the "think" process is essential for ensuring its effectiveness. (2) The *PM* strategy is necessary for successfully applying the probability-based inference method for scoring. Furthermore, our experimental setup maintains the optimal performance in all ablation settings, validating its rationality.

Train		Inference		Scoring				Interpret	
PD	*PM*	*Num.*	*Prob.*	*KonIQ*	*AGIQA*	*SPAQ*	*KADID*	*Tech.*	*Overall*
×	✓	×	✓	0.928	0.815	**0.930**	0.702	79.35%	78.51%
✓	×	×	✓	0.735	0.629	0.762	0.487	/	/
✓	✓	✓	×	0.912	0.808	0.907	0.691	/	/
✓	✓	×	✓	**0.931**	**0.817**	0.924	**0.709**	**81.27%**	**80.78%**

Table 6: Ablation study of *Stage-2* attributes (using the *Refine-IQA-S2 (w/o think)*). *PD* denotes the probability difference reward and *PM* represents the policy gradient modification. *Num.* and *Prob.* represent the numeric-based and probability-based inference strategy **in the scoring task**. The inference method for the interpreting task is the same as the above-mentioned. [Per column: highest in **bold**.]

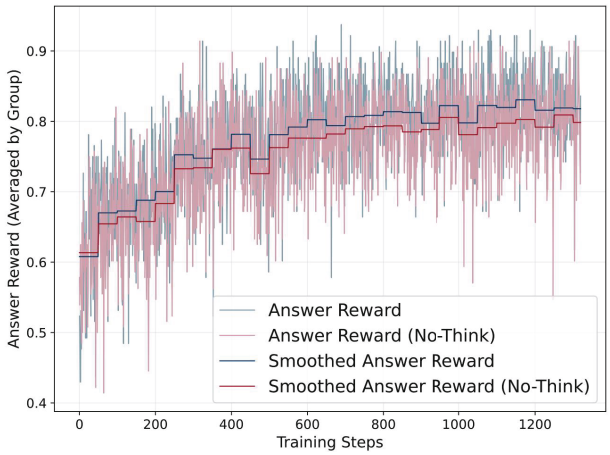

Figure 4: The *Stage-2* group averaged answer reward of "think" and "no-think" modes during training.

In addition, we present visualizations of some training ablation details in Fig. 4. The figure demonstrates that, in the absence of specific optimization for the "no-think" mode output, its output accuracy still increases simultaneously with the "think" mode. This observation justifies our decision to exclude dedicated model optimization for the "no-think" mode when calculating the PD reward.

Conclusion

In this paper, we propose the *Refine-IQA*, an RFT framework that focuses on enhancing the LMM's perception of low-level visual quality and incentivizing the model's effective "think" capability for IQA. In *Stage-1*, we construct the *Refine-Perception-20K* dataset and a multi-task reward scheme. In *Stage-2*, we introduce the *PD reward* along with *gradient policy modification*. After training, we obtain the *Refine-IQA Series Models*, which achieve excellent performance on both quality perception and scoring tasks. Additionally, we validate the model's "think" capability in quality interpretating tasks. With only approximately $13K$ scoring labels for RFT, we can trigger the model's "think" capability. Our work provides compelling insights for developing **efficient visual quality assessment agents** that are both **labeling friendly** and **functionally robust**.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62522116, Grant 62271312 and Grant 62132006, and in part by STCSM under Grant 22DZ2229005. We also sincerely appreciate the Shanghai Artificial Intelligence Laboratory for computation resources supporting.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Pietquin, O.; Üstün, A.; and Hooker, S. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Anthropic, T. 2025. Claude 3.7 sonnet and claude code.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Cai, Z.; Zhang, J.; Yuan, X.; Jiang, P.-T.; Chen, W.; Tang, B.; Yao, L.; Wang, Q.; Chen, J.; and Li, B. 2025. Q-Ponder: A Unified Training Pipeline for Reasoning-based Visual Quality Assessment. *arXiv preprint arXiv:2506.05384*.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.-w.; Jeon, B. E.; Fang, Y.; Lee, H.-Y.; Ren, J.; Yang, M.-H.; et al. 2024. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 13320–13331.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3677–3686.

Ghadiyaram, D.; and Bovik, A. C. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE TIP*, 25(1): 372–387.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP*, 29: 4041–4056.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Huang, Y.; Sheng, X.; Yang, Z.; Yuan, Q.; Duan, Z.; Chen, P.; Li, L.; Lin, W.; and Shi, G. 2024. AesExpert: Towards Multi-modality Foundation Model for Image Aesthetics Perception. *arXiv preprint arXiv:2404.09624*.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-Scale Image Quality Transformer. In *ICCV*, 5148–5157.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1): 011006–011006.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; and Lin, W. 2023. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. arXiv:2306.04717.

Li, W.; Zhang, X.; Zhao, S.; Zhang, Y.; Li, J.; Zhang, L.; and Zhang, J. 2025a. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*.

Li, X.; Yan, Z.; Meng, D.; Dong, L.; Zeng, X.; He, Y.; Wang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2025b. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *QoMEX*, 1–3.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Mazurek, M. E.; Roitman, J. D.; Ditterich, J.; and Shadlen, M. N. 2003. A role for neural integrators in perceptual decision making. *Cerebral cortex*, 13(11): 1257–1269.

Ren, T.; Chen, Y.; Jiang, Q.; Zeng, Z.; Xiong, Y.; Liu, W.; Ma, Z.; Shen, J.; Gao, Y.; Jiang, X.; et al. 2024. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*.

Ren, Z.; Shao, Z.; Song, J.; Xin, H.; Wang, H.; Zhao, W.; Zhang, L.; Fu, Z.; Zhu, Q.; Yang, D.; et al. 2025. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *CVPR*.

Talebi, H.; and Milanfar, P. 2018. NIMA: Neural Image Assessment. *IEEE TIP*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wang, J.; Chan, K. C. K.; and Loy, C. C. 2022. Exploring CLIP for Assessing the Look and Feel of Images.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, W.; Gao, Z.; Chen, L.; Chen, Z.; Zhu, J.; Zhao, X.; Liu, Y.; Cao, Y.; Ye, S.; Zhu, X.; et al. 2025a. Visualprm: An

effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Wang, Y.; Yang, L.; Tian, Y.; Shen, K.; and Wang, M. 2025b. Co-evolving llm coder and unit tester via reinforcement learning. *arXiv preprint arXiv:2506.03136*.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; and Lin, W. 2024a. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision. In *ICLR*.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. 2024b. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *CVPR*, 25490–25500.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2024c. Q-ALIGN: teaching LMMs for visual scoring via discrete text-defined levels. In *ICML*, 54015–54029.

Wu, H.; Zhu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Li, C.; Wang, A.; Sun, W.; Yan, Q.; et al. 2024d. Towards open-ended visual quality comparison. In *ECCV*, 360–377.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 13040–13051.

Ying, Z.; Mandal, M.; Ghadiyaram, D.; and Bovik, A. 2021. Patch-vq:'patching up'the video quality problem. In *CVPR*, 14019–14029.

You, Z.; Gu, J.; Li, Z.; Cai, X.; Zhu, K.; Dong, C.; and Xue, T. 2024a. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*.

You, Z.; Li, Z.; Gu, J.; Yin, Z.; Xue, T.; and Dong, C. 2024b. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, 259–276. Springer.

Yu, E.; Lin, K.; Zhao, L.; Yin, J.; Wei, Y.; Peng, Y.; Wei, H.; Sun, J.; Han, C.; Ge, Z.; et al. 2025a. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2020. Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network. *IEEE TCSVT*, 30(1): 36–47.

Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, Z.; Wang, J.; Guo, Y.; Wen, F.; Chen, Z.; Wang, H.; Li, W.; Sun, L.; Zhou, Y.; Zhang, J.; Yan, B.; Jia, Z.; Xiao, J.; Tian, Y.; Zhu, X.; Zhang, K.; Li, C.; Liu, X.; Min, X.; Jia, Q.; and Zhai, G. 2025a. AIBench: Towards trustworthy evaluation under the 45° law. *Displays*, 103255.

Zhang, Z.; Wang, J.; Wen, F.; Guo, Y.; Zhao, X.; Fang, X.; Ding, S.; Jia, Z.; Xiao, J.; Shen, Y.; Zheng, Y.; Zhu, X.; Wu, Y.; Jiao, Z.; Sun, W.; Chen, Z.; Zhang, K.; Fu, K.; Cao, Y.; Hu, M.; Zhou, Y.; Zhou, X.; Cao, J.; Zhou, W.; Cao, J.; Li, R.; Zhou, D.; Tian, Y.; Zhu, X.; Li, C.; Wu, H.; Liu, X.; He, J.; Zhou, Y.; Liu, H.; Zhang, L.; Wang, Z.; Duan, H.; Zhou, Y.; Min, X.; Jia, Q.; Zhou, D.; Zhang, W.; Cao, J.; Yang, X.; Yu, J.; Zhang, S.; Duan, H.; and Zhai, G. 2025b. Large Multimodal Models Evaluation: A Survey. *SCIS*.

Zhu, H.; Wu, H.; Li, Y.; Zhang, Z.; Chen, B.; Zhu, L.; Fang, Y.; Zhai, G.; Lin, W.; and Wang, S. 2024. Adaptive image quality assessment via teaching large multimodal model to compare. *arXiv preprint arXiv:2405.19298*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.