Q-REAL: Towards Naturalness and Distortion Evaluation for AI-Generated Content

Shushi Wang[1], Zicheng Zhang[3], Chunyi Li[1], Wei Wang[2], Liya Ma[2], Fengjiao Chen[2], Xiaoyu Li[2], Xuezhi Cao[2], Guangtao Zhai[1], and Xiaohong Liu[1,4†]

[1] Shanghai Jiao Tong University, China
[2] Meituan, China
[3] Shanghai AI Lab, China
[4] Shanghai Innovation Institute, China

Abstract. Quality assessment of AI-generated content is vital for model optimization, yet most existing evaluation datasets and models rely on coarse-grained single scores, failing to provide targeted guidance. To bridge this gap, we introduce **Q-Real**, a fine-grained quality assessment dataset specifically designed for AI-generated images. This dataset comprises 10K images generated by multiple models, annotated along two critical dimensions, **naturalness** and **distortion** which are widely regarded as the most significant aspects of AI-generated image quality. For each image, we localize major entities and provide a set of judgment questions and attribution descriptions along these dimensions to facilitate comprehensive evaluation. Based on this dataset, we establish the **Q-Real Bench** to rigorously evaluate models on the challenging tasks of judgment and grounding with reasoning. And to tackle these challenges, we further propose a fine-grained training pipeline for Multimodal Large Language Models (MLLMs), empowering them to judge, localize problematic entities with detailed analysis, and predict quality score. Experimental results demonstrate the high quality and significance of our dataset, as well as the effectiveness of our proposed training pipeline. The dataset is available at `https://huggingface.co/datasets/AGI-Eval/Q-Real`.

Keywords: AI-generated image · Quality assessment · Multimodal Large Language Model

1 Introduction

With the rapid advancement of generative AI, massive amounts of text-to-image content are being produced and applied across various domains [14,29]. However, current generative content sometimes exhibits issues, which necessitate accurate quality assessment models to assess the generated content before practical use. Such evaluation models are also critical for guiding the iterative optimization of generative models. Existing studies have established a comprehensive evaluation

† Corresponding author.

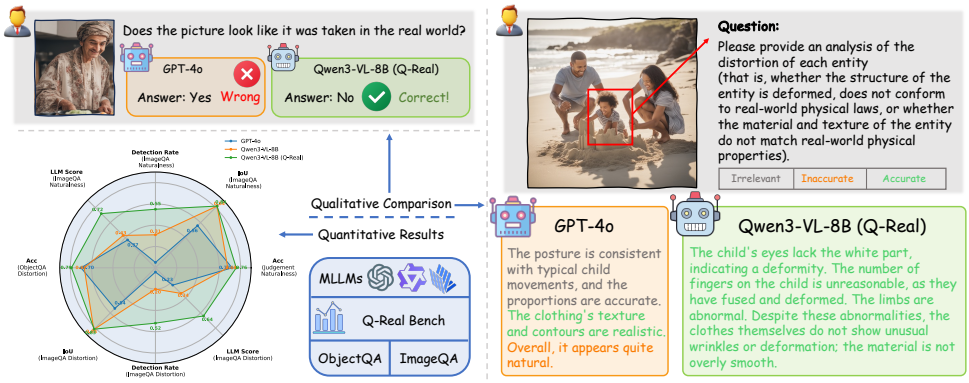

Fig. 1: Abilities of Q-Real-tuned Qwen3-VL-8B-Instruct [6] on fine-grained quality assessment tasks about AI-generated images, in comparison with the baseline version.

paradigm for AI-generated images, encompassing text–image alignment, image quality, and visual aesthetics, with effective models available for each evaluation dimension [5, 13, 15, 17, 38, 46]. However, most existing evaluation approaches remain relatively coarse-grained, typically compressing complex generation results into a single score, which makes it difficult to precisely identify the sources of errors. Due to the lack of interpretable evaluation signals, such methods provide limited guidance for the targeted optimization of generative models. Consequently, developing fine-grained evaluation models is of critical importance.

Given that mainstream AI-generated image applications predominantly target photorealism, our work focuses on constructing a fine-grained quality evaluation framework tailored for this scenario. Within the context of photorealistic generation, we evaluate image quality along two complementary dimensions: **naturalness**, which measures perceptible artificiality or AI-related artifacts (often referred to as the "AI look"), and **distortion**, which focuses on visual degradations such as blur, deformation, and structural corruption.

To support the construction of the proposed fine-grained quality evaluation framework, we construct a dataset of 10K AI-generated images with annotations on the two dimensions of naturalness and distortion, named **Q-Real**. Specifically, 7K images are sourced from the existing Q-Eval-100K dataset [46], while the remaining 3K are generated by more recent state-of-the-art models [3, 4, 9, 28]. For fine-grained annotation along these two dimensions, we localize major entities within each image using bounding boxes. For each entity, we design corresponding objective judgment questions for both naturalness and distortion. Furthermore, for entities identified as problematic, we provide detailed descriptive analyses of the specific issues. Given the high labor cost and time-consuming nature of such fine-grained annotation, drawing inspiration from prior works [7, 22], we design an automated annotation pipeline combined with manual refinement. This approach ensures both the efficiency of dataset construction and the high quality of the annotations.

Based on this dataset, we construct the benchmark **Q-Real Bench**, featuring two complementary tasks: **ObjectQA** and **ImageQA**. ObjectQA requires models to answer judgment questions about naturalness and distortion of entities in images, which assesses the fundamental ability to discriminate naturalness and distortion in AI-generated images, enabling efficient filtering of low-quality entities. In contrast, ImageQA requires models to localize problematic entities and provide detailed description, which demands a comprehensive understanding to localize and explain these issues, providing detailed feedback crucial for optimizing generative models.

To tackle the challenges posed by Q-Real Bench, we leverage Multimodal Large Language Models (MLLMs) [1, 2, 6, 18, 24], which excel at general visual understanding [40] but remain limited in fine-grained quality assessment (Figure 1). We therefore propose a three-stage progressive training pipeline. First, object-level judgment QA pairs are used to build a basic understanding of naturalness and distortions. Second, image-level grounding and reasoning descriptions enable fine-grained evaluation. Finally, the generated assessments are incorporated as Chain-of-Thought (CoT) [33] reasoning context, together with a rating token mapping strategy, to predict quality scores. The resulting model can simultaneously perform judgment, localization, analysis, and scoring.
The main contributions of this paper are summarized as follows:

- We present **Q-Real**, the first fine-grained quality evaluation dataset for AI-generated images on **Naturalness** and **Distortion**, built via an automated pipeline with manual refinement.
- We design **Q-Real Bench**, comprising **ObjectQA** and **ImageQA**, to offer a comprehensive evaluation framework that systematically examines the capacity of MLLMs in fine-grained quality assessment.
- We propose a **three-stage training pipeline** and conduct experiments on multiple MLLMs, demonstrating the high quality of the dataset, the comprehensiveness of the benchmark, and the effectiveness of our approach in enhancing fine-grained evaluation.

2 Related Work

2.1 Coarse-grained Image Quality Assessment

Early image quality assessment (IQA) methods, including both full-reference and no-reference approaches, were primarily developed for user-generated images and aimed to predict overall perceptual quality by modeling human visual perception. Traditional metrics and early CNN-based models such as NIQE [23] and DBCNN [43], as well as transformer-based methods like LIQE [44] and CLIP-IQA [32], typically provide a single quality score. More recently, MLLM-based approaches such as Q-Align [35] and DEQA [39] have been introduced for IQA, but most still follow a coarse-grained evaluation paradigm.

With the development of generative models like GANs [10], diffusion models [11], and autoregressive models [31], the evaluation of generated images has

increased. Early evaluation of AI-generated images mainly relied on objective metrics such as PSNR [27] and LPIPS [41], which focus on pixel-wise or perceptual feature differences. To better reflect human perception, several subjectively annotated datasets have been introduced, including AGIQA-3K [17], AIGIQA-20K [16], and Q-Eval-100K [46], which provide mean opinion scores (MOS), as well as Pick-a-Pic [15] and HPS [21, 36], which use side-by-side annotations. These datasets have enabled the adaptation of conventional IQA models for assessing AI-generated images. However, most existing methods and datasets remain coarse-grained,as they primarily predict a single overall quality score and lack the ability to provide analyses of quality issues in AI-generated images.

2.2 Fine-grained Image Quality Assessment

To overcome the limitation that score-based evaluation fails to reveal the underlying causes of image quality degradation, fine-grained image quality assessment has been increasingly studied.

Among existing fine-grained image quality assessment efforts for user-generated content, Q-Instruct [34] extends quality evaluation beyond scalar scores by formulating fine-grained image quality assessment as quality-related questions, including multiple-choice, true–false, overall quality judgment, and explanatory reasoning. It also introduces quality-aware reasoning tasks, such as suggesting shooting-condition adjustments to improve image quality. Grounding-IQA [7] further enriches the expression of fine-grained quality assessment by introducing spatial grounding. Instead of only providing text evaluations, it requires models to localize regions with quality issues by outputting bounding boxes, thereby explicitly indicating where visual degradations occur within an image.

For AI-generated images, RichHF-18K [19] pioneered fine-grained quality assessment by linking quality with plausibility and providing scores with pixel-level heatmaps. FakeXplain [12] and X-AIGD [37] further focus on distortion regions: FakeXplain combines grounding with descriptions but lacks structured question-based annotations, while X-AIGD performs region-level distortion classification without descriptive analysis. However, they still lack a unified treatment of distortion localization, description, spatial grounding, and naturalness assessment for photorealistic images, motivating our Q-Real.

3 Q-Real Dataset Construction

3.1 Source Collection

Data Collection. To construct a high-quality dataset for our task, we collected images from two sources. First, we sampled a subset from Q-Eval-100K [46], a dataset featuring diverse images generated by multiple models across various categories and quality levels. Second, considering the continuous iteration of generative models, and to ensure the quality and timeliness of the dataset, we used the latest models including flux.1.1pro [4], Kwai-Kolors [28], sd3.5 [3] and dreamina3.0 [9] to generate additional images.

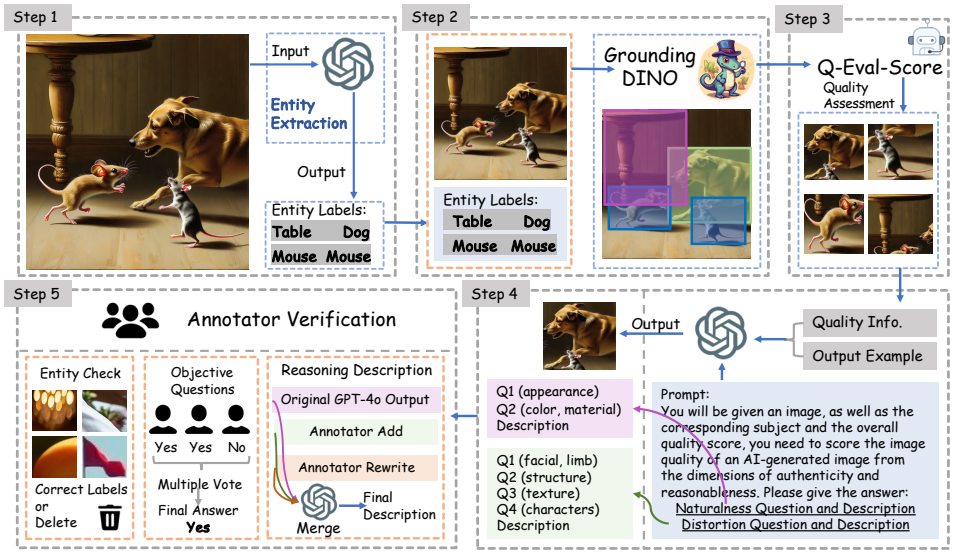

Fig. 2: The pipeline of our dataset construction.

Prompt Selection. Whether sampling images from the Q-Eval-100K dataset or generating new ones using models, selecting appropriate prompts is crucial to ensuring the comprehensiveness and diversity of the dataset. When sampling from Q-Eval-100K, we designed four types of prompts (portrait generation, character-animal generation, entity generation, and scene generation) and filtered the images based on these prompts to cover a variety of generative tasks. For the generated portion, we created relevant prompts for these four categories using GPT, and then used them to produce images, thereby guaranteeing the dataset's representativeness and diversity.

3.2 Q-Real Annotation Pipeline

A high-quality dataset is essential for training accurate evaluation models and establishing a reliable evaluation benchmark. However, annotating such a dataset with subjective description and bounding boxes is labor-intensive and costly. To address this challenge, we have developed a hybrid annotation strategy that combines automated annotation with manual verification and supplement, shown in Figure 2. The process is as follows:

Entity Name Extraction. Initially, we provide both the generated images and their corresponding prompts to GPT-4o [24] to extract the names of entities appearing in the images, ensuring comprehensive entity identification. By jointly analyzing visual content and textual context, GPT-4o can more accurately capture entities present in the generated images.

Object Detection. For each entity, we employ Grounding DINO [20] for precise localization, which offers superior reliability compared to direct MLLM extraction. To ensure annotation quality, we filter out low-confidence detections and extremely small, visually ambiguous bounding boxes.

Quality Evaluation. Following detection, we use the AI-generated image quality assessment model Q-Eval-Score [46] to assess the quality of each detected object. The quality score serves as an important conditioning signal for GPT-4o or other MLLMs to generate accurate analyses of naturalness and distortion. By incorporating quality scores, the model is guided to produce more reliable and well-aligned descriptions of the visual quality of detected entities.

Generate Question-Answer Pairs and Reasoning Descriptions. Lastly, each extracted object, together with its corresponding quality score, is fed into GPT-4o to generate objective judgments and reasoning descriptions of its naturalness and distortion. Through this automated annotation pipeline comprising the first four steps, we obtain the initial version of the Q-Real dataset.

Subjective Experiment. After the automated annotation pipeline, a manual verification step is conducted to ensure annotation quality.

First, a trained annotator verifies whether each extracted object matches its assigned label, correcting any mismatches. Objects with bounding boxes that are too small or visually unclear are removed.

After this initial verification, each retained object is independently reviewed by three annotators. They first determine whether the GPT-4o-generated description contains errors related to naturalness or distortion. If errors are identified, all annotators independently write human descriptions to replace the generated one; if not, they may supplement missing details. The final subjective description is synthesized by merging annotator inputs with an LLM, ensuring that multiple perspectives are considered and the description is as comprehensive as possible. For objects with erroneous descriptions, annotators also answer objective judgment questions, and the majority vote is used as the final label. This multi-step protocol ensures accurate fine-grained annotation.

Additionally, considering that human portrait generation is comparatively more challenging and involves diverse application scenarios, we provide finer-grained evaluation annotations for the distortion dimension of 400 human-centric images, shown in Figure 3. Specifically, each human entity is analyzed across multiple body and semantic dimensions, including face, hands, limbs, torso, feet, clothing, and interactions with the surrounding environment. For each of these dimensions, three annotators independently examine the entity and mark the specific issues observed, ensuring both granularity and consistency of the evaluation.

Fig. 3: An example of finer-grained annotation for the distortion of human portrait.

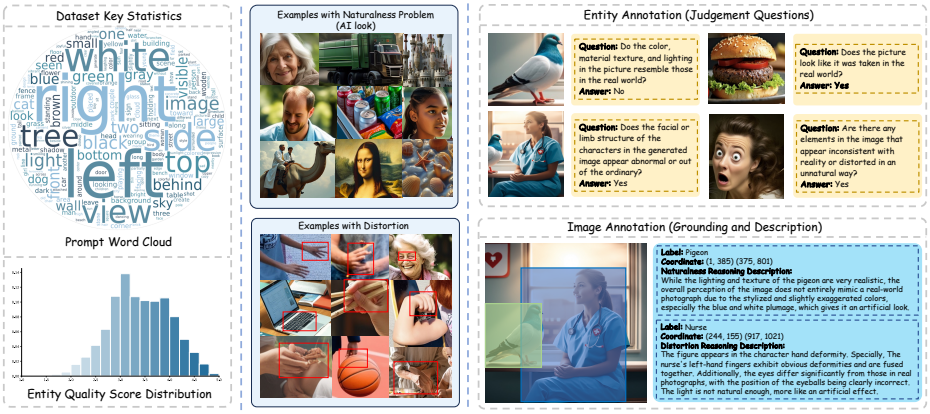

Fig. 4: Overview of the Q-Real dataset. Left: Word clouds and quality score distributions showcasing diversity. Middle: Representative examples of naturalness and distortion issues. Right: Sample images with annotations.

3.3 Dataset Format

Our dataset consists of 10K AI-generated images, with 7K sampled from the Q-Eval-100K dataset and 3K generated using the latest models. All images were generated using a total of 3,049 prompts. For each image, we marked all the main entities using bounding boxes. In total, we annotated 32,815 entities, covering various categories such as people, animals, and objects. Figure 4 provides an overview of the Q-Real, illustrating its diversity and representative examples.

For all annotated entities, we provide fine-grained quality annotations under the two dimensions, **naturalness** and **distortion**. Each dimension includes both objective judgment questions and a subjective attribution description, enabling structured evaluation as well as explanatory analysis.

For each entity, two judgment questions are constructed to evaluate its naturalness:

1. *Does the entity appear as if it exists in the real world?*
2. *Do the color, material, texture, and lighting of the entity resemble those observed in the real world?*

In addition, four judgment questions are formulated to assess the distortion of the entity:

1. *Does the structural configuration (e.g., facial features or limbs) of the entity appear abnormal or anatomically implausible?*
2. *Are there any unnatural deformations or inconsistencies in the shape of the entity?*
3. *Does the material or texture of the entity deviate from realistic physical properties?*
4. *If text is present within the entity, does it contain errors or illegible characters?*

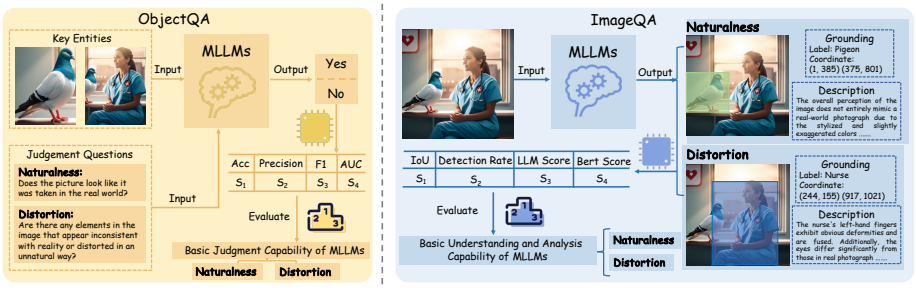

Fig. 5: Overview of Q-Real Bench, illustrating its two tasks: ObjectQA and ImageQA, including evaluation procedures and metrics.

For the subjective attribution descriptions, the naturalness dimension reflects overall visual naturalness and whether the entity appears AI-generated ("AI look"), focusing on aspects such as color, lighting, and material realism. In contrast, the distortion dimension describes whether there are regions with structural abnormalities, blurriness, or physically implausible features. Compared to RichHF-18K [36], which provides coarse heatmap annotations for distortion regions and lacks descriptive analysis, our dataset is more fine-grained and includes more accurate localization and detailed attribution descriptions. More detailed information about Q-Real can be found in the Supplementary Material.

3.4 Q-Real Bench

For evaluating model performance in fine-grained assessment of AI-generated images along the naturalness and distortion, we construct a benchmark, Q-Real Bench, as shown in Figure 5. The Q-Real Bench includes 1,000 images of various types and quality, which are selected from Q-Real Dataset. Based on this benchmark, we evaluate the performance of models from two perspectives:

ObjectQA. This task requires the model to answer the objective judgment questions annotated in Q-Real for each entity in an AI-generated image, evaluating its basic ability to assess naturalness and distortion at the entity level. Since ObjectQA involves only binary (Yes/No) responses, we adopt four standard metrics—Accuracy (*Acc*), *Precision*, *F1*, and the Area Under the ROC Curve (*AUC*)—to comprehensively evaluate performance.

ImageQA. The ImageQA task requires the model to localize problematic entity regions in the image under the naturalness and distortion dimensions, and provide corresponding descriptions of the issues.

To evaluate the model's grounding performance, we adopt two standard metrics: Intersection over Union (*IoU*) and *Detection Rate*. For computing IoU, we use the Hungarian matching algorithm to establish an optimal one-to-one correspondence between predicted and ground-truth entities. Given a set of predicted bounding boxes $\mathcal{P} = \{p_i\}_{i=1}^{N}$ and ground-truth boxes $\mathcal{G} = \{g_j\}_{j=1}^{M}$, we construct a cost matrix $C \in \mathbb{R}^{N \times M}$ based solely on coordinate overlap, and solve for the

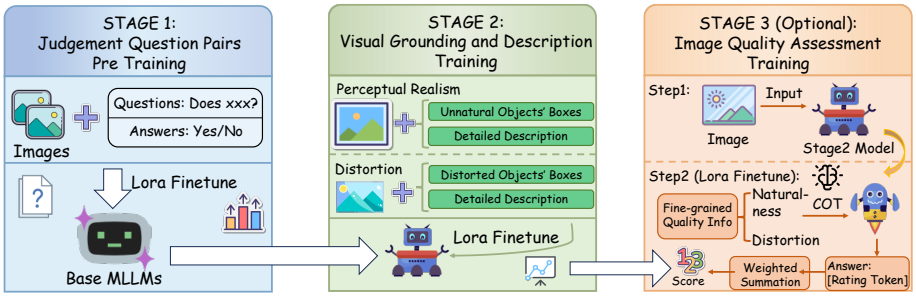

Fig. 6: Our proposed fine-grained quality assessment training pipeline.

optimal matching permutation π^* as follows:

$$C_{ij} = 1 - \text{IoU}(p_i, g_j), \quad \pi^* = \arg\min_{\pi} \sum_{i=1}^{K} C_{i\pi(i)} \tag{1}$$

where $K = \min(N, M)$, and $C_{i\pi(i)}$ denotes the cost between the i-th predicted and matched ground-truth box. The final matched pairs are then used to compute the IoU. Detection Rate is defined as the fraction of predicted entities that both match a ground-truth box with IoU greater than 0.5 and provide a description that is consistent with the annotated issue (LLM Score is greater than 0.5). *LLM Score* is introduced later.

To evaluate reasoning descriptions, we adopt two complementary metrics: *LLM-Score* and *Bert-Score*. LLM-Score [30, 45] feeds the predicted and ground-truth descriptions into GPT-4o to measure semantic consistency, enabling robust evaluation beyond surface wording differences. We additionally report Bert-Score [42] as a traditional lexical overlap metric. Together, they provide complementary views of semantic correctness and textual similarity.

4 Fine-grained Quality Assessment Training Pipeline

To empower MLLMs with comprehensive quality assessment capabilities, we propose a three-stage progressive training pipeline. This pipeline systematically enhances the model's ability to judge, localize, reason, and score AI-generated images, as illustrated in Figure 6.

4.1 Stage 1: Object-Level Judgment Training

We fine-tune MLLMs using object-level judgment QA pairs from Q-Real to establish fundamental discriminative abilities for naturalness and distortion. For each key entity, we utilize six question-answer pairs, formatted as follows:
"query": "<image>Does the entity appear as if it exists in the real world?"
"response": "Yes/No"

4.2 Stage 2: Fine-Grained Evaluation Capability Training

We construct a training dataset where each instance pairs a complete AI-generated image with the coordinates (bounding boxes) and detailed reasoning descriptions of problematic entities. To focus the model on discerning defects, we exclusively include entities exhibiting naturalness or distortion issues. This targeted approach enhances the model's capability to localize and analyze specific quality degradations. Taking the naturalness dimension as an example, the training instance is formatted as follows:
"query": "<image>Analyze this AI-generated image and identify all entities that exhibit naturalness issues. For each such entity, provide its type label, its coordinates within the image and a brief explanation describing why the entity appears unrealistic."
"response": "<label><coordinate><description>... (repeated for each problematic entity)"

4.3 Stage 3: Quality Score Prediction Training

Context Prompt. Prior work on LMM-based quality assessment has demonstrated that employing a structured Chain-of-Thought (CoT) prompt can effectively guide models to yield more accurate quality evaluations [46]. Inspired by this, we leverage the model fine-tuned in Stage 2 to first identify issues. This approach is grounded in two key observations: image quality is determined by the level of naturalness and the severity of distortion, and it is also heavily dependent on the quality of major entities within the image. Then, we incorporate the detailed fine-grained evaluation results about naturalness and distortion (<detected_issues>) into the prompt as a structured COT.
<image>Please evaluate the overall quality of this image and assign it one of the following ratings: [Excellent, Good, Fair, Poor, Bad]. Reference Information: The following list details specific regions in the image identified as problematic. Each entry includes the bounding box coordinates of the defective area and a brief description of the issue: <detected_issues>. Based on the visual content and the identified defects above, output only one of the ratings: [Excellent, Good, Fair, Poor, Bad]. Do not provide any reasoning.

To ensure consistency between training and inference, we do not directly utilize ground-truth annotations to construct the prompt during this stage. Instead, we employ the model fine-tuned in Stage 2 to perform inference, dynamically generating localized defects and reasoning descriptions to form the prompt.
Scoring Method. Following established practices [35, 47], we map numerical MOS values to discrete adjective ratings for training, as LLMs interpret these more effectively. Specifically, we discretize the MOS range $[1, 5]$ into five intervals corresponding to $\{Bad, Poor, Fair, Good, Excellent\}$. The mapping function $R(s)$ is defined as:

$$R(s) = r_i, \quad \text{if } m + \frac{i-1}{5}(M-m) < s \leq m + \frac{i}{5}(M-m), \tag{2}$$

where $\{r_i\}_{i=1}^{5} = \{Bad, Poor, Fair, Good, Excellent\}$, $m = 1$, and $M = 5$.

During inference, we compute the continuous quality score $\hat{r}$ by extracting logits of the five rating tokens. We apply softmax to obtain probabilities p_j and calculate the expectation:

$$p_j = \frac{\exp(z_j)}{\sum_{k=1}^{5} \exp(z_k)}, \quad \hat{r} = \sum_{j=1}^{5} p_j \cdot w_j, \tag{3}$$

where $w_j = \{0, 0.25, 0.5, 0.75, 1\}$ are weights for ratings from Bad to Excellent.

Loss Function. To optimize the model for both general question-answering capabilities and precise score prediction, we adopt a hybrid loss function combining Cross-Entropy (CE) Loss and Mean Squared Error (MSE) Loss. The CE Loss, $\mathcal{L}_{CE}$, facilitates learning the QA format and necessary knowledge, while the MSE Loss, $\mathcal{L}_{MSE}$, refines the accuracy of the predicted quality score. The total loss $\mathcal{L}$ is defined as a weighted sum:

$$\mathcal{L} = \alpha \mathcal{L}_{CE} + \beta \mathcal{L}_{MSE}, \tag{4}$$

where $\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(p_i)$ represents the standard cross-entropy loss for answer tokens, and $\mathcal{L}_{MSE} = (\hat{r} - r_{MOS})^2$ measures the squared difference between the predicted score $\hat{r}$ and the ground-truth MOS label r_{MOS}. We set the weights α and β to 1 by default.

5 Experiments

5.1 Experiment Setup

Q-Real Bench Evaluation. We evaluate three model groups: closed-source (GPT-5.4 and Gemini3.1-Pro) [8,25], quality assessment-specific (Q-Eval-Score) [46], and open-source MLLMs (including Qwen3-VL-8B [1], Qwen2.5-VL-7B [2], InternVL3-8B [48], and InternVL2.5-8B [6]). For the open-source models (Qwen and InternVL series), we compare their performance in both zero-shot settings and after fine-tuning with our proposed Stage 1 and Stage 2 training.

Quality Assessment Evaluation. We employ Qwen3-VL-8B as the backbone model, fine-tuned using our proposed three stages training pipeline. We evaluate its performance on both Q-Real Bench and the existing AGIQA-3K dataset [17], using SRCC and PLCC as metrics and further test on ImageReward [38] with accuracy as the metric.

Training Settings. We fine-tune the models with LoRA using a rank of 8 and alpha of 32. The models are trained for 3, 5, and 3 epochs in Stages 1, 2, and 3, respectively. Other hyperparameters follow the base-model defaults. Experiments are conducted on 8 NVIDIA H100 (80GB) GPUs using PyTorch.

5.2 Discussion & General Findings

1)For Q-Real Bench Evaluation. The general performance on the ObjectQA and ImageQA is exhibited in Table 1 and Table 2. In the ObjectQA task,

Table 1: Performance comparison on the **ObjectQA** task. Evaluation metrics include Acc ↑, Precision ↑, F1 score ↑, and AUC ↑. The models marked with * indicate models fine-tuned using Q-Real. And questions refer to the judgment questions in the naturalness and distortion dimensions, as defined in Section 3.3.

Model	Naturalness Question 1				Naturalness Question 2				Distortion Question 1			
	Acc	*Precision*	*F1 Score*	*AUC*	*Acc*	*Precision*	*F1 Score*	*AUC*	*Acc*	*Precision*	*F1 Score*	*AUC*
GPT-5.4	0.661	0.574	0.617	-	0.611	0.519	0.661	-	0.640	0.666	0.261	-
Gemini3.1-Pro	0.629	0.590	0.407	-	0.654	0.576	0.607	-	0.691	0.575	0.632	-
Q-Eval-Score	0.675	0.723	0.613	0.741	0.645	0.636	0.634	0.688	0.683	0.555	0.552	0.711
Qwen2.5-VL-7B-Instruct	0.605	0.769	0.388	0.776	0.630	0.734	0.497	0.750	0.652	0.714	0.082	0.711
Qwen2.5-VL-7B-Instruct*	0.751 +14.6%	0.716 −5.3%	0.757 +36.9%	0.828 +5.2%	0.743 +11.3%	0.700 −3.4%	0.757 +26.0%	0.826 +7.6%	0.770 +12.2%	0.685 −2.9%	0.669 +57.3%	0.828 +11.7%
InternVL2.5-8B	0.700	0.692	0.687	0.285	0.651	0.602	0.699	0.341	0.656	0.574	0.221	0.683
InternVL2.5-8B*	0.762 +6.2%	0.740 +4.8%	0.758 +7.1%	0.807 +52.2%	0.752 +10.1%	0.730 +12.8%	0.753 +5.4%	0.814 +47.3%	0.767 +11.1%	0.691 +11.7%	0.658 +43.7%	0.842 +15.9%
Qwen3-VL-8B-Instruct	0.702	0.694	0.686	0.776	0.661	0.651	0.652	0.724	0.562	0.283	0.195	0.398
Qwen3-VL-8B-Instruct*	0.762 +6.0%	0.728 +3.4%	0.765 +7.9%	0.843 +6.7%	0.755 +9.4%	0.724 +7.3%	0.761 +10.9%	0.840 +11.6%	0.777 +21.5%	0.718 +43.5%	0.663 +46.8%	0.843 +44.5%
InternVL3-8B	0.682	0.690	0.653	0.776	0.643	0.591	0.702	0.748	0.659	0.596	0.227	0.714
InternVL3-8B*	0.737 +5.5%	0.692 +0.2%	0.655 +0.2%	0.810 +3.4%	0.734 +9.1%	0.700 +10.9%	0.646 +6.4%	0.805 +5.7%	0.794 +13.5%	0.752 +15.6%	0.767 +54.0%	0.874 +16.0%

	Distortion Question 2				Distortion Question 3				Distortion Question 4			
	Acc	*Precision*	*F1 Score*	*AUC*	*Acc*	*Precision*	*F1 Score*	*AUC*	*Acc*	*Precision*	*F1 Score*	*AUC*
GPT-5.4	0.638	0.700	0.706	-	0.572	0.787	0.511	-	0.893	0.452	0.523	-
Gemini3.1-Pro	0.576	0.602	0.720	-	0.555	0.614	0.636	-	0.893	0.463	0.601	-
Q-Eval-Score	0.609	0.584	0.706	0.666	0.589	0.558	0.698	0.692	0.852	0.406	0.520	0.873
Qwen2.5-VL-7B-Instruct	0.563	0.810	0.351	0.690	0.673	0.828	0.587	0.754	0.892	0.736	0.074	0.881
Qwen2.5-VL-7B-Instruct*	0.702 +23.9%	0.719 −9.1%	0.717 +36.6%	0.772 +3.2%	0.735 +6.2%	0.744 −8.4%	0.738 +15.1%	0.812 +5.8%	0.933 +4.1%	0.689 −4.7%	0.701 +62.7%	0.929 +4.8%
InternVL2.5-8B	0.608	0.604	0.668	0.554	0.554	0.536	0.680	0.402	0.896	0.648	0.223	0.886
InternVL2.5-8B*	0.707 +9.9%	0.707 +10.3%	0.731 +6.3%	0.760 +20.6%	0.744 +19.0%	0.744 +20.8%	0.752 +7.2%	0.819 +41.7%	0.945 +4.9%	0.798 +15.0%	0.727 +50.4%	0.935 +4.9%
Qwen3-VL-7B-Instruct	0.600	0.590	0.676	0.665	0.622	0.642	0.612	0.667	0.928	0.874	0.559	0.915
Qwen3-VL-7B-Instruct*	0.709 +10.9%	0.719 +12.9%	0.727 +5.1%	0.787 +12.2%	0.736 +11.4%	0.747 +10.5%	0.739 +12.7%	0.821 +11.4%	0.946 +1.8%	0.801 −7.3%	0.737 +17.8%	0.953 +3.8%
InternVL3-8B	0.589	0.689	0.507	0.677	0.624	0.718	0.434	0.748	0.898	0.804	0.185	0.933
InternVL3-8B*	0.701 +11.2%	0.706 +1.7%	0.723 +21.6%	0.784 +10.7%	0.727 +10.3%	0.724 +0.6%	0.737 +30.3%	0.822 +7.4%	0.939 +4.1%	0.729 −7.5%	0.715 +53.0%	0.958 +2.5%

Table 2: Performance comparison of different models on the **ImageQA** task. Evaluation metrics include IoU ↑, Detection Rate↑, LLM Score↑, and Bert Score↑. The models marked with * indicate fine-tuned models.

Model	Naturalness				Distortion			
	IoU	*Detection Rate*	*LLM Score*	*Bert Score*	*IoU*	*Detection Rate*	*LLM Score*	*Bert Score*
GPT-5.4	0.700	0.069	0.439	0.373	0.661	0.063	0.197	0.290
Gemini3.1-Pro	0.752	0.481	0.376	0.357	0.820	0.136	0.431	0.296
Q-Eval-Score	0.835	0.165	0.353	0.208	0.838	0.038	0.249	0.328
Qwen2.5-VL-7B-Instruct	0.875	0.271	0.409	0.344	0.829	0.059	0.230	0.387
Qwen2.5-VL-7B-Instruct*	0.875 +0.0%	0.596 +32.5%	0.713 +30.4%	0.482 +13.8%	0.878 +4.9%	0.548 +48.9%	0.633 +40.3%	0.394 +0.7%
InternVL2.5-8B	0.730	0.095	0.390	0.362	0.763	0.031	0.255	0.278
InternVL2.5-8B*	0.825 +9.5%	0.491 +39.6%	0.675 +28.5%	0.441 +7.9%	0.817 +5.4%	0.448 +41.7%	0.593 +34.8%	0.366 +8.8%
Qwen3-VL-8B-Instruct	0.848	0.313	0.435	0.322	0.848	0.206	0.341	0.263
Qwen3-VL-8B-Instruct*	0.834 −1.4%	0.548 +23.5%	0.722 +28.7%	0.483 +16.1%	0.832 −1.6%	0.523 +31.7%	0.639 +29.8%	0.385 +12.2%
InternVL3-8B	0.888	0.060	0.230	0.354	0.878	0.274	0.413	0.274
InternVL3-8B*	0.829 −5.9%	0.527 +46.7%	0.737 +50.7%	0.499 +14.5%	0.825 −5.3%	0.481 +20.7%	0.645 +23.2%	0.391 +11.7%

existing MLLMs demonstrate a certain capability in judging naturalness and distortion in AI-generated images, generally performing better on naturalness than on distortion. However, their performance in the distortion dimension is notably weak, particularly in identifying implausible regions and illegible characters, where they show limited discriminative ability (characterized by high precision but low recall). After fine-tuning with Q-Real, the models exhibit a significant improvement in both dimensions. Most evaluation metrics across various questions exceed 0.7, substantially outperforming the zero-shot performance of MLLMs and even surpassing closed-source models like GPT.

In the ImageQA task, existing MLLMs perform poorly, often indiscriminately detecting all entities in the image without accurately determining whether they exhibit naturalness or distortion issues. As a result, while IoU scores are high, both detection rate and description similarity remain low. After fine-tuning, models retain object detection ability and gain naturalness and distortion analysis, enabling accurate localization and detailed issue descriptions. Both detection rate and description similarity are significantly improved. Specifically, for the naturalness dimension, the detection rate reaches above 0.5 and the description similarity exceeds 0.7 (LLM Score); for the distortion dimension, the detection rate increases to 0.5 and the description similarity reaches 0.6 (LLM Score). Detailed visualization results are provided in the Supplementary Material.

2)For Quality Assessment Evaluation. The general performance is exhibited in Table 3. Our model trained with the Stage 3, named Q-Real-Score, achieves the best performance on all Q-Real Bench, AGIQA-3K, and ImageReward. Compared to the second-best competitor (Q-Eval-Score), while it employs a structured CoT prompt to guide the model on which dimensions to consider for quality assessment, it only provides general instructions without details. In contrast, our approach incorporates detailed issues about naturalness and distortion into the prompt, offering more explicit guidance thereby improving performance.

Table 3: Performance comparison on quality assessment datasets. All models are trained on the subset of Q-Real derived from Q-Eval-100K. The best results are highlighted in **bold**, and the second-best are underlined.

Model	Q-Real		AGIQA-3K		ImageReward
	SRCC	*PLCC*	*SRCC*	*PLCC*	*Acc*
CLIP-IQA [32]	0.264	0.262	0.074	0.088	0.493
IPCE [26]	0.631	0.636	0.074	0.057	0.495
Q-Align [35]	0.651	0.644	0.572	0.558	0.556
Q-Eval-Score [46]	0.690	0.693	0.587	0.582	0.601
Q-Real-Score	**0.710**	**0.720**	**0.608**	**0.611**	**0.607**

3)Cross-dataset Validation. In the RichHF-18K [19], the plausibility dimension is defined similarly to the distortion dimension in our work. To demonstrate the value of Q-Real, we conduct cross-dataset validation on the test data of RichHF-18K in the distortion dimension, with performance results shown in Table 4.

Table 4: Cross-dataset validation on RichHF-18K. The models marked with * indicate fine-tuned models.

Model	ImageQA		ObjectQA	
	Recall	*Precision*	*Acc*	*F1 Score*
Qwen2.5-VL-7B-Instruct	0.102	0.122	0.493	0.492
Qwen2.5-VL-7B-Instruct*	$0.401_{+29.9\%}$	$0.531_{+40.9\%}$	$0.551_{+5.8\%}$	$0.518_{+2.6\%}$
Intern2.5VL-8B	0.221	0.175	0.464	0.441
Intern2.5VL-8B*	$0.280_{+5.9\%}$	$0.285_{+11.0\%}$	$0.564_{+10.0\%}$	$0.500_{+5.9\%}$
Qwen3-VL-8B-Instruct	0.505	0.273	0.532	0.404
Qwen3-VL-8B-Instruct*	$0.537_{+3.2\%}$	$0.615_{+34.2\%}$	$0.556_{+2.4\%}$	$0.492_{+8.8\%}$
Intern3VL-8B	0.172	0.175	0.442	0.465
Intern3VL-8B*	$0.542_{+37.0\%}$	$0.635_{+46.0\%}$	$0.552_{+11.0\%}$	$0.516_{+5.1\%}$

For ImageQA, we compute recall and precision between the annotated regions in RichHF-18K and the predicted boxes from our model. For ObjectQA, since RichHF-18K provides plausibility scores for each image, we consider images with scores above 0.6 as having no plausibility issues. We then use our model's judgments on the distortion dimension and calculate the accuracy and F1 score between the two. Experimental results show that the model fine-tuned on Q-Real achieves accurate and stable performance on RichHF-18K, with clear improvements over the zero-shot baseline.

Table 5: Ablation study on Stage 1 training. For each model, the [1] denotes training without Stage 1, while [2] indicates training with Stage 1 followed by Stage 2.

Model	Naturalness				Distortion			
	IoU	*Detection Rate*	*LLM Score*	*Bert Score*	*IoU*	*Detection Rate*	*LLM Score*	*Bert Score*
Qwen2.5-VL-7B-Instruct[1]	0.874	0.595	0.715	0.481	0.879	0.547	0.623	0.388
Qwen2.5-VL-7B-Instruct[2]	0.875$_{+0.1\%}$	0.596$_{+0.1\%}$	0.723$_{+0.8\%}$	0.482$_{+0.1\%}$	0.878$_{-0.1\%}$	0.548$_{+0.1\%}$	0.633$_{+1.0\%}$	0.394$_{+0.6\%}$
InternVL2.5-8B[1]	0.820	0.482	0.667	0.435	0.808	0.418	0.577	0.356
InternVL2.5-8B[2]	0.825$_{+0.5\%}$	0.491$_{+0.9\%}$	0.675$_{+0.8\%}$	0.441$_{+0.6\%}$	0.817$_{+0.9\%}$	0.448$_{+3.0\%}$	0.593$_{+1.6\%}$	0.366$_{+1.0\%}$
Qwen3-VL-8B-Instruct[1]	0.832	0.540	0.721	0.479	0.831	0.518	0.627	0.385
Qwen3-VL-8B-Instruct[2]	0.834$_{+0.2\%}$	0.548$_{+0.8\%}$	0.722$_{+0.1\%}$	0.483$_{+0.4\%}$	0.832$_{+0.1\%}$	0.523$_{+0.5\%}$	0.639$_{+1.2\%}$	0.385$_{+0.0\%}$
InternVL3-8B[1]	0.830	0.522	0.728	0.488	0.821	0.477	0.629	0.385
InternVL3-8B[2]	0.829$_{-0.1\%}$	0.527$_{+0.5\%}$	0.737$_{+0.9\%}$	0.499$_{+1.1\%}$	0.825$_{+0.4\%}$	0.481$_{+0.4\%}$	0.645$_{+1.6\%}$	0.391$_{+0.6\%}$

Table 6: Ablation study of different prompts on quality assessment. **N-CoT**: Naturalness CoT, **D-CoT**: Distortion CoT, **TIC**: Training-Inference Consistency.

Strategy			Q-Real		AGIQA-3K		ImageReward
N-CoT	*D-CoT*	*TIC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*Acc*
-	✓	✓	<u>0.700</u>	0.694	0.570	0.598	0.558
✓	-	✓	0.696	<u>0.700</u>	<u>0.576</u>	<u>0.601</u>	<u>0.560</u>
✓	✓	-	0.692	0.686	0.550	0.574	0.557
✓	✓	✓	**0.710**	**0.720**	**0.608**	**0.611**	**0.607**

5.3 Ablation Study

1)For Q-Real Bench Evaluation. We conduct an ablation study to assess the contribution of Stage 1 training on the model's performance in the ImageQA task. The result is presented in Table 5. It is clear that performing Stage 1 training first better prepares the model to handle the more complex training task in Stage 2.

2)For Quality Assessment Evaluation. We conduct an ablation study to validate the effectiveness of our prompt design in Stage 3. Specifically, we evaluate the contributions of including naturalness issues, distortion issues, and training-inference consistency in the prompt design. The results are presented in Table 6, which show that including both naturalness and distortion issues in the prompt improves quality prediction, with distortion issues having a stronger correlation with quality and training-inference consistency has the most significant impact—models without this strategy perform the worst.

6 Conclusion

We present **Q-Real**, a fine-grained quality assessment dataset of 10K AI-generated images, focusing on naturalness and distortion. Moving beyond traditional score-based metrics, Q-Real integrates entity localization, objective judgments, and subjective analyses. Based on this, we introduce **Q-Real Bench** for accurate model evaluation about the fine-grained assessment performance and design a dedicated **training pipeline** that effectively unifies fine-grained assessment with score prediction. Our experimental results demonstrate the effectiveness of both the dataset and our proposed method. We hope our work inspires future advancements in MLLM-based quality assessment for AI-generated content.

Acknowledgements

The work was supported in part by the National Natural Science Foundation of China under Grant 62301310 and 62572317.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. In: arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. In: arXiv preprint arXiv:2502.13923 (2025)
3. Bandyopadhyay, H., Entezari, R., Scott, J., Adithyan, R., Song, Y.Z., Jampani, V.: Sd3.5-flash: Distribution-guided distillation of generative flows. In: arXiv preprint arXiv:2509.21318 (2025)
4. Black-Forest-Labs: https://bfl.ai/models/flux-pro (2025)
5. Cao, S., Ma, N., Li, J., Li, X., Shao, L., Zhu, K., Zhou, Y., Pu, Y., Wu, J., Wang, J., Qu, B., Wang, W., Qiao, Y., Yao, D., Liu, Y.: Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In: arXiv preprint arXiv:2507.14533 (2025)
6. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. In: arXiv preprint arXiv:2412.05271 (2024)
7. Chen, Z., Zhang, X., Li, W., Pei, R., Song, F., Min, X., Liu, X., Yuan, X., Guo, Y., Zhang, Y.: Grounding-iqa: Multimodal language grounding model for image quality assessment. In: arXiv preprint arXiv:2411.17237 (2024)
8. DeepMind, G.: https://deepmind.google/models/gemini/pro/ (2026)
9. DreaminaAI: https://dreamina.capcut.com/ (2025)
10. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Advances in Neural Information Processing Systems (2014)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (2020)
12. Ji, Y., Hong, Y., Fan, Q., Lan, J., Zhu, H., Wang, W., Zhang, L., Zhang, J.: Fakexplain: Ai-generated image detection via human-aligned grounded reasoning. In: International Conference on Learning Representations (2026)
13. Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., Chen, W.: Genai arena: An open evaluation platform for generative models. In: arXiv preprint arXiv:2406.04485 (2024)
14. Journal, E.: Ai image statistics for 2024: How much content was created by ai. In: Everypixel (2024)
15. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: Advances in Neural Information Processing Systems (2020)
16. Li, C., Kou, T., Gao, Y., Cao, Y., Sun, W., Zhang, Z., Zhou, Y., Zhang, Z., Zhang, W., Wu, H., et al.: Aigiqa-20k: A large database for ai-generated image quality assessment. In: arXiv preprint arXiv:2404.03407 (2024)

17. Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., Lin, W.: Agiqa-3k: An open database for ai-generated image quality assessment. In: IEEE Transactions on Circuits and Systems for Video Technology (2023)
18. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. In: arXiv preprint arXiv:2407.07895 (2024)
19. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., Ke, J., Dj Dvijotham, K., Collins, K.M., Luo, Y., Li, Y., Kohlhoff, K.J., Ramachandran, D., Navalpakkam, V.: Rich human feedback for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: arXiv preprint arXiv:2303.05499 (2023)
21. Ma, Y., Wu, X., Sun, K., Li, H.: Hpsv3: Towards wide-spectrum human preference score. In: International Conference on Computer Vision (2025)
22. Meng, X., Zhang, Z., Zhang, Z., Liao, J., Qin, L., Wang, W.: Identity-grpo: Optimizing multi-human identity-preserving video generation via reinforcement learning. In: arXiv preprint arXiv:2510.14256 (2025)
23. Mittal, A., Soundararajan, A.K., Bovik, A.C.: Making a 'completely blind' image quality analyzer. In: IEEE Signal Processing Letters (2013)
24. OpenAI: https://openai.com/index/gpt-4o (2024)
25. OpenAI: https://developers.openai.com/api/docs/models/gpt-5.4 (2026)
26. Peng, F., Fu, H., Ming, A., Wang, C., Ma, H., He, S., Dou, Z., Chen, S.: Aigc image quality assessment via image-prompt correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2024)
27. Rafael C, G., Wintz, P.: Digital image processing. In: Addison-Wesley Longman Publishing (1987)
28. Team, K.: Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis (2025)
29. Techreport: Ai image generator market statistics in 2024. In: Techreport (2024)
30. Thakur, A.S., Choudhary, K., Ramayapally, V.S.: Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In: arXiv preprint arXiv:2406.12624 (2024)
31. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: arXiv preprint arXiv:2404.02905 (2024)
32. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI Conference on Artificial Intelligence (2023)
33. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: arXiv preprint arXiv:2201.11903 (2023)
34. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., Xue, G., Sun, W., Yan, Q., Lin, W.: 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 (2024)
35. Wu, H., Zhang, Z., Zhang, W., Chen, C., Li, C., Liao, L., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In: arXiv preprint arXiv:2312.17090 (2023)

36. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. In: arXiv preprint arXiv:2306.09341 (2023)
37. Xiao, Y., Chen, W., Chen, J., Cao, Z., Deng, W., Yang, B., Dong, Z., Ji, X., Ke, W., Wei, P., Lin, L.: Unveiling perceptual artifacts: A fine-grained benchmark for interpretable ai-generated image detection. In: The Fourteenth International Conference on Learning Representations (2026)
38. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: learning and evaluating human preferences for text-to-image generation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (2023)
39. You, Z., Cai, X., Gu, J., Xue, T., Dong, C.: Teaching large language models to regress accurate image quality scores using score distribution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
40. Zhang, P., Dong, X., Wang, B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Duan, H., Zhang, S., Ding, S.: Internlmxcomposer: A vision-language large model for advanced text-image comprehension and composition. In: arXiv preprint arXiv:2309.15112 (2023)
41. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
42. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020)
43. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. In: IEEE Transactions on Circuits and Systems for Video Technology (2020)
44. 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 Conference on Computer Vision and Pattern Recognition (2023)
45. Zhang, Z., Jia, Z., Wu, H., Li, C., Chen, Z., Zhou, Y., Sun, W., Liu, X., Min, X., Lin, W., Zhai, G.: Q-bench-video: Benchmarking the video quality understanding of lmms. In: arXiv preprint arXiv:2409.20063 (2024)
46. Zhang, Z., Kou, T., Wang, S., Li, C., Sun, W., Wang, W., Li, X., Wang, Z., Cao, X., Min, X., Liu, X., Zhai, G.: Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In: arXiv preprint arXiv:2503.02357 (2025)
47. Zhang, Z., Wu, H., Ji, Z., Li, C., Zhang, E., Sun, W., Liu, X., Min, X., Sun, F., Jui, S., Lin, W., Zhai, G.: Q-boost: On visual quality assessment ability of low-level multi-modality foundation models. In: arXiv preprint arXiv:2312.15300 (2023)
48. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. In: arXiv preprint arXiv:2504.10479 (2025)