RESEARCH ARTICLE

Grounding-IQA: Multimodal Language Grounding Model for Image Quality Assessment

Zheng Chen[1], Yulun Zhang[1]

1 Shanghai Jiao Tong University, Shanghai, China

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

The development of multimodal large language models (MLLMs) enables the evaluation of image quality through natural language descriptions. This advancement allows for more detailed assessments. However, these MLLM-based IQA methods primarily rely on general contextual descriptions, sometimes limiting fine-grained quality assessment. To address this limitation, we introduce a new image quality assessment (IQA) task paradigm, **grounding-IQA**. This paradigm integrates multimodal referring and grounding with IQA to realize more fine-grained quality perception. Specifically, grounding-IQA comprises two subtasks: grounding-IQA-description (GIQA-DES) and visual question answering (GIQA-VQA). GIQA-DES involves detailed descriptions with precise locations (*e.g.*, bounding boxes), while GIQA-VQA focuses on quality QA for local regions. To realize grounding-IQA, we construct a corresponding dataset, GIQA-160K, through our proposed automated annotation pipeline. Furthermore, we develop a well-designed benchmark, GIQA-Bench. The benchmark comprehensively evaluates the model grounding-IQA performance from three perspectives: description quality, VQA accuracy, and grounding precision. Experiments demonstrate that our proposed task paradigm, dataset, and benchmark facilitate the more fine-grained IQA application.
Code: https://github.com/zhengchen1999/Grounding-IQA.

Corresponding author: Yulun Zhang, yulun100@gmail.com

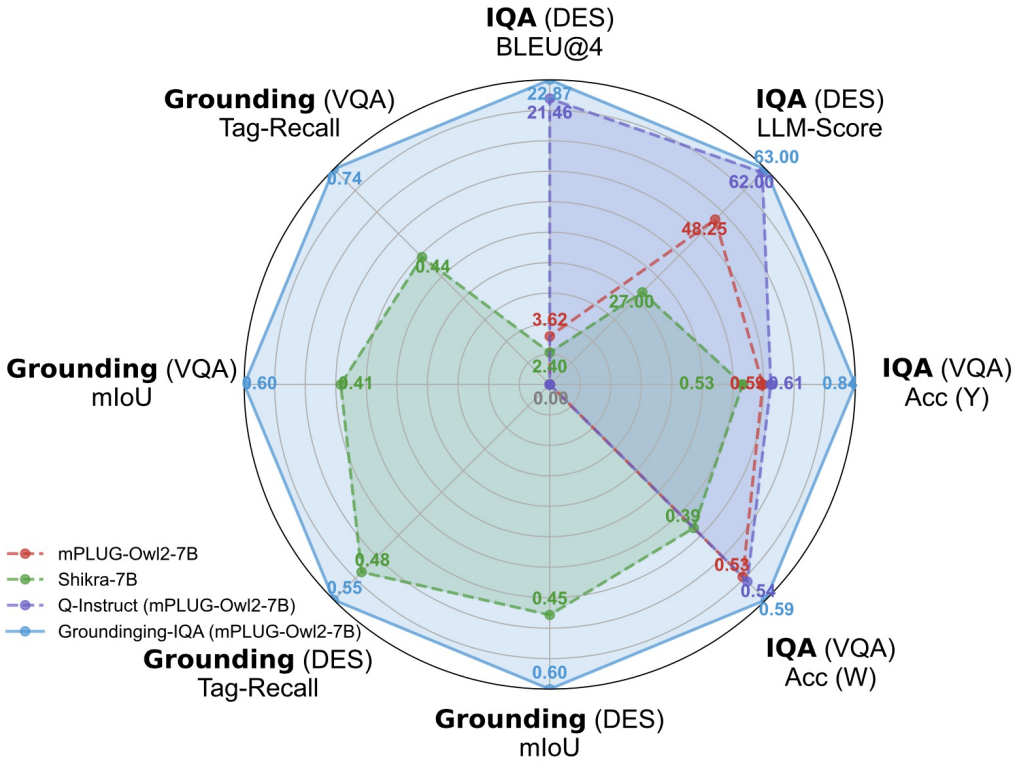

Figure 1. Performance comparisons on GIQA-Bench. Our proposed grounding-GPT effectively combines grounding and IQA.

1. Introduction

Image quality assessment (IQA) seeks to evaluate image quality in alignment with human perception. As a fundamental task in low-level vision, IQA is critical across multiple fields, *e.g.*, image processing[1][2], media transmission[3], and generative artificial intelligence[4]. However, this task is challenging since the human visual system is inherently subjective and complex to model[5]. To enhance evaluation precision, substantial research efforts continue to be dedicated to this area[6][7][8][9].

Traditional IQA methods employ handcrafted metrics to estimate quality scores[5][10]. With advancements in deep neural networks, learning specific priors from large datasets enables more accurate score predictions[11][12][1][13][14]. Nevertheless, score-based IQA methods face challenges in complex scenarios. In such cases, image quality is influenced by multiple factors that a single score cannot effectively express[15]. Recently, the emergence of multimodal large language models (MLLMs)[16][17][18] with strong visual and linguistic perception capabilities provides an alternative to score-based IQA[19][20]. These models achieve more detailed and accurate image assessments through description and reasoning. However, current MLLM-based IQA methods[9][21] primarily rely on general contextual descriptions, which sometimes limits fine-grained quality assessments. For instance, in Fig. 2a, the existing method (*i.e.*, Q-Instruct[9]) describes the objects/areas affecting image quality through language, but cannot provide precise location information. Moreover, in Fig. 2b, for local perception, the language referring may not accurately pinpoint the target. These limitations restrict the application of MLLMs in comprehensive low-level perception and understanding, especially for fine-grained

cases.

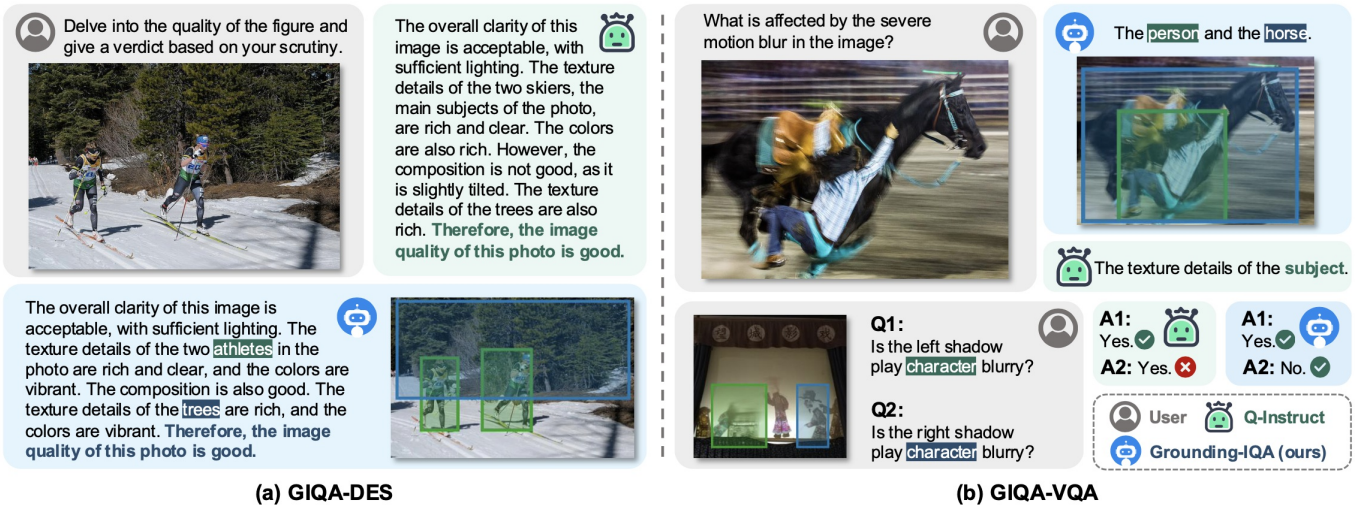

Figure 2. Grounding-IQA combines referring and grounding with IQA. (a) GIQA-DES: Quality description include precise locations (*i.e.*, bounding boxes). (b) GIQA-VQA: The question (referring, bottom instance) or answer (grounding, top instance) contains locations.

To address these challenges and unleash the potential of MLLMs in fine-grained image quality understanding, we introduce grounding-IQA. This is a novel IQA task paradigm that integrates multimodal referring (*position in*) and grounding (*position out*)[22][23][17] with image quality assessment. Specifically, we categorize grounding-IQA into two sub-tasks: (1) **Grounding-IQA-Description (GIQA-DES)**. As illustrated in Fig. 2a, this task requires generating descriptive assessments of image quality while providing precise locations (*i.e.*, bounding boxes) for important objects/regions impacting quality. (2) **Grounding-IQA-Visual Question Answering (GIQA-VQA)**. As shown in Fig. 2b, this task involves QA about low-level attributes of images, especially regarding local objects. It includes addressing questions with specific coordinates (*referring*) or providing answers with precise positions (*grounding*).

Based on the task designed above, we construct a high-quality dataset,**GIQA-160K**, to enable existing MLLMs with grounding-IQA capabilities. This dataset comprises 160K instruction-tuning data with 40K images from diverse domains (*e.g.*, artificial distortion and in-the-wild scenes). Specifically, the dataset corresponds to two sub-tasks: GIQA-DES includes 60K corresponding data, and GIQA-VQA contains 100K related data. To circumvent the time-consuming and costly process of manual annotation, we design an **automated annotation pipeline**. This system utilizes the public IQA dataset[9][21] (with human-annotated description), to construct the corresponding dataset. (1)**For GIQA-DES.** The task includes detailed descriptions with coordinates. We generate the data through advanced vision[24] and language[25] models. Through these models, we extract and filter objects and corresponding coordinates from existing descriptions and images. Meanwhile, coordinates are expressed in natural language and attached to text. This avoids extra specialized tokens and ensures data compatibility. (2) **For GIQA-VQA.** Inspired by previous work[9][26][27], we construct the required data from the detailed descriptions in GIQA-DES via the LLM. We use specific QA templates (*i.e.*, “Yes/No’, abbreviated as Y; “What/How/Why”, abbreviated as W) and emphasize location-specific objects to generate appropriate data. The coordinates are also combined with the generated QA.

Fine-tuning on the GIQA-160K dataset enables existing pre-trained MLLMs to achieve impressive grounding-IQA capabilities. As shown in Fig. 2, the fine-tuned model can ground key objects affecting image quality, and perform more fine-grained assessments based on reference coordinates. Moreover, to comprehensively evaluate the model performance on the grounding-IQA task, we propose a well-designed benchmark, **GIQA-Bench**. This benchmark includes 100 varying types and quality images, corresponding to 100 GIQA-Des and 150 GIQA-VQA test samples. Each sample is annotated over multiple rounds by at least three experts. We quantitatively assess grounding-IQA performance in three aspects: (1)assessment description quality (*i.e.*, BLEU@4, LLM-Score); (2) VQA accuracy (*i.e.*, Accuracy); and (3) grounding precision (*i.e.*, mIoU, Tag-Recall). We test recent MLLMs, with results shown in Fig. 1. Observations indicate significant improvement in grounding-IQA after fine-tuning with GIQA-160K.

Overall, our contributions are threefold:

- We introduce multimodal referring and grounding into IQA, establishing a new IQA paradigm, grounding-IQA, for fine-grained quality perception and assessment.
- We construct a high-quality dataset, GIQA-160K, with an automated annotation pipeline. The dataset is versatile and suitable for fine-tuning existing MLLMs.
- We propose a high-quality benchmark, GIQA-Bench, to comprehensively evaluate the model performance on grounding-IQA from three aspects.

2. Related work

2.1. Image Quality Assessment

Score-based Methods.

Most current IQA methods are score-based. Early IQA approaches compute scores through handcrafted image data metrics[5][28][29][6][30]. For instance, PSNR calculates the ratio of signal to noise. NIQE[10] relies on the statistical characteristics of natural images. However, these methods show a gap in quality perception compared to human judgment and are unsuitable for complex scenarios. With the development of the deep neural network, learning-based IQA methods have gradually become mainstream[11][31][12][13][32][33]. These methods leverage data-driven training to achieve more accurate quality assessments. For example, LPIPS[1] applies the convolutional neural network to compute scores. MUSIQ[14] employs the Transformer to extract multi-scale features for score prediction. Moreover, meta-learning[34], multimodal models[8][35], and graph neural networks[36] have been adopted to further improve IQA performance. However, score-based IQA methods face limitations in complex scenarios. The simple score cannot effectively represent the multiple aspects affecting image quality.

MLLM-based Methods.

Multimodal large language models (MLLMs) exhibit remarkable multimodal (language/vision) understanding by integrating visual modules into LLMs[16][37][38][39][18]. MLLMs achieve outstanding performance in various multimodal tasks, including visual question answering and image captioning. Recently, several studies have also demonstrated the potential of MLLMs in low-level visual perception and assessment[9][21][21][40][41]. For instance, Q-Instruct[9] constructs a multimodal dataset to enhance. Q-Align[42] guides MLLMs in scoring by defining discrete text-based levels. DepictQA[15] enables quality comparison and reasoning based on reference images. These approaches advance the application of MLLMs in IQA, achieving more accurate assessments. Nevertheless, these models primarily rely on contextual descriptions, and face limitations in fine-grained applications, *e.g.*, local perception.

2.2. Multimodal Referring and Grounding

Multimodal spatial perception involves referring and grounding. **Referring** requires the model to understand the specific region based on position input, *e.g.*, region-level captioning[43][44]. **Grounding**, on the other hand, involves the model describing the region by outputting position, *e.g.*, referring expression comprehension[45][46]. Currently, MLLMs perform impressively in spatial perception, further advancing these tasks. Some methods focus on grounding, achieving complex reasoning[47] or multi-object[48] segmentation. Meanwhile, other approaches, *e.g.*, GPT4RoI[49], emphasize understanding specific regions (referring). Furthermore, some works unify referring and grounding[23][27][50]. Kosmos-2[17] utilizes bounding box coordinates to integrate both aspects. Ferret[26] extends to referring to arbitrary shapes. Additionally, in IQA, Q-Ground[41] achieves degradation region grounding but lacks referring capabilities. Overall, compared to previous work, our Grounding-IQA integrates multimodal referring and grounding with IQA. This new IQA paradigm enables more fine-grained and flexible quality perception.

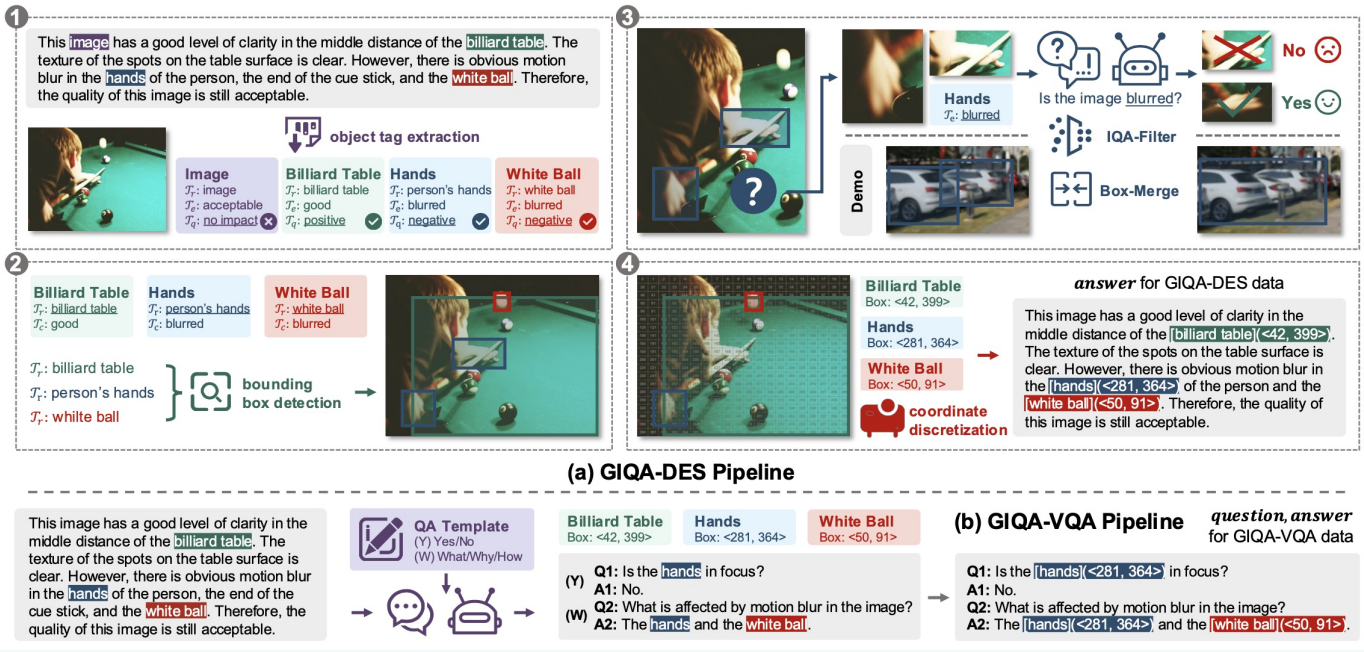

Figure 3. The illustration of the automated annotation pipeline. (a) GIQA-DES Pipeline: Constructs the *answer* from the given image and corresponding description through a four-stage process, while the *question* comes from a predefined question pool. (b) GIQA-VQA Pipeline: Generates the corresponding QA data utilizing descriptions from GIQA-DES and the advanced LLM, Llama3[25].

3. Method

In this section, we introduce the newly defined IQA paradigm, grounding-IQA. The content includes**(1)** definition of paradigm and two subtasks, Sec. 3.1; **(2)** data construction pipeline, Sec. 3.2;**(3)** details of GIQA-160K, Sec. 3.3;**(4)** benchmark for grounding-IQA, Sec. 3.4.

3.1. Grounding-IQA

As analyzed above, existing MLLM-based IQA methods leverage descriptions to enable more accurate and detailed quality assessments. However, these methods remain limited in performing fine-grained evaluations, as in Fig. 2. Inspired by work on multimodal referring and grounding, we believe that spatial perception is key to achieving more fine-grained assessments. Therefore, to further unlock the potential of MLLMs, we introduce a new IQA paradigm, grounding-IQA. This paradigm combines referring and grounding with IQA to enable more precise and flexible quality assessments. Specifically, grounding-IQA should include the two sub-tasks/capabilities: grounding-IQA-description (GIQA-DES) and grounding-IQA-visual question answering (GIQA-VQA).

GIQA-DES.

The task requires the model to provide a detailed description of image quality. Additionally, it needs accurate location information (*e.g.*, bounding box) for key objects/regions that impact image quality, as shown in Fig. 5a. This corresponds to the fact that humans consider not only the overall quality (*e.g.*, image clarity) but also the quality of specific objects or locations when assessing image quality. Meanwhile, accurate location information also enables targeted information for downstream tasks (*e.g.*, image editing). This task is similar to grounded image captioning[51], but places greater emphasis on low-level attributes. While some MLLMs[23][17][27] perform well in grounded image captioning, they still struggle with quality perception. We demonstrated it in Sec. 4.3.

GIQA-VQA.

The second task focuses on the question-answering ability in low-level perception, particularly for local objects. Corresponding to multimodal referring and grounding, this task can be divided into two scenarios. ***Referring*:** querying low-level attributes in the specified region (*input position*), as shown in Fig. 5b. ***Grounding*:** providing answers that include specific locations (*output position*) based on the question, as depicted in Fig. 5b. These two scenarios are related to region captioning[51] and phrase grounding[51], respectively. However, like GIQA-DES, GIQA-VQA involves quality perception, which is challenging for current MLLMs. We evaluate in Sec. 4.3.

3.2. Automated Annotation Pipeline

Data is essential for achieving Grounding-IQA. Therefore, we construct an automated annotation pipeline to generate data

(*i.e.*, GIQA-160K). This pipeline leverages public IQA datasets[9][21] that contain human-annotated descriptions. Following previous schemes[52][18], the data format is {**image**, **question**, **answer**}. The **image** is the evaluation target. Depending on the sub-task, the **question** and **answer** fields may include precise coordinates (*i.e.*, **bounding box**), in addition to text. We introduce the pipeline below; refer to the supplement for more details.

For GIQA-DES.

In this task, the **question** is relatively fixed, as in Fig. 5a. To enhance data diversity, we construct a pool of 15 similar questions. For each data, the **question** is randomly selected from the pool. For the **answer**, it is a detailed description with coordinates. We construct it via a four-stage process from existing images and associated description, as illustrated in Fig. 3: (1) **Stage-1:** object tag extraction; (2) **Stage-2:** bounding box detection; (3) **Stage-3:** box refinement (filter and merge); and (4) **Stage-4:** transformation and fusion. Each stage is detailed below.

Figure 4. Compared with applying object name ("man"), utilizing the description phrase T_q ("the man wearing a white t-shirt") can achieve more accurate detection results.

Stage-1: object tag extraction. Firstly, we apply the advanced LLM, Llama3[25], to extract key objects (*e.g.*, "billiard table" in Fig. 3a) from the given descriptions. Each object is assigned a three-tuple form tag: {T_r, T_q, T_e}. The T_r is the object description phrase (sometimes same as name); T_q denotes the quality of object (*e.g.*, "clear"); T_e represents the object effect on image quality (*i.e.*, "no impact", "positive", or "negative"). All tag items are inferred from the description, with T_r and T_q used in later stages. The T_e item enables us to filter out non-critical objects (*e.g.*, "image", which refers to the whole). This explicit effect classification, similar to chain-of-thought (CoT), can reduce hallucinations. Overall, by extracting and filtering, we can obtain the target object and information for subsequent processing.

Stage-2: bounding box detection. Then, we detect bounding boxes for the extracted objects from the image. To accomplish this, we utilize the state-of-the-art object detection model, Grounding DINO[24]. Since multiple same-category objects may appear in one image, we utilize the T*r* generated **Stage-1** rather than the object name for detection. For instance, in Fig. 4, the object name is "man", and T*r* is "the man wearing a white t-shirt". Leveraging "man" detects two objects (left case), while using T*r* can achieve the more precise result (right case).

Stage-3: box refinement. We further refine the detected boxes. Although **Stage-2** adopts Tr to limit the detection range, multiple boxes may still exist. In some cases, multiple boxes may contain the wrong target. Through observations, most detection errors arise from the detection model inability to distinguish objects of same class with different quality. For instance, in Fig. 3a, for "hands", the key (reduce image quality) is the blurry one, and the other is irrelevant. To address this problem, we design the IQA-Filter algorithm (Alg. 1). We use the MLLM-based IQA method, Q-Instruct, to verify detected bounding boxes by inputting each box patch and asking: "Is the image quality is <T$_q$>?", with T$_q$ from **Stage-1**. We check all boxes in single-object-multiple-targets, and remove those with a "No" response.

Furthermore, in some cases, multiple small or overlapping targets correspond to the same object. While these detections are accurate, an excess of targets may increase the learning difficulty for MLLMs. To address this issue, we propose the Box-Merge algorithm (Alg. 1). We merge boxes that satisfy the normalized area threshold T_a (set to 0.256), and the overlap threshold T_o (set to 95%). Overall, the IQA-Filter and Box-Merge algorithms effectively refine the quality of bounding boxes.

Stage-4: transformation and fusion. Finally, we integrate the extracted and filtered boxes into the original descriptions to construct the **answer**. To avoid introducing extra specialized tokens for box representation, we treat box coordinates as regular text tokens, attaching them to the text in the interleaved format: "[object/region](bounding box)".

Moreover, bounding boxes are typically represented by normalized corner coordinates: $\langle x_1, y_1, x_2, y_2 \rangle$. When the coordinate values are rounded to two decimal places (*e.g.*, $\langle 0.01, 0.02, 0.03, 0.04 \rangle$), representing box requires **21** tokens ($4 \times 4 + 5$). Inspired by previous work[26][17], we discretize the coordinates to simplify the representation. We divide the image into $n \times m$ grids and numbering grids from top-left to bottom-right from top-left to bottom-right: $\{0,1,..,nm-1\}$. Patch numbers then represent the top-left and bottom-right coordinates of the box:

$$\begin{aligned} \text{idx}_l &= y_1 \cdot m \cdot n + x_1 \cdot n, \\ \text{idx}_r &= y_2 \cdot m \cdot n + x_2 \cdot n, \end{aligned}$$

where idx_l and idx_r denotes the discretized coordinates. The box can be represented as $\langle \text{idx}_l, \text{idx}_r \rangle$, *e.g.*, $\langle 10, 110 \rangle$. Accordingly, we remap the discrete coordinates back to continuous format using the centre coordinates of each grid:

$$\begin{aligned} x_1^{'} &= (\text{idx}_l \% n + 0.5)/n, \quad & y_1^{'} &= (\text{idx}_l / n + 0.5)/m, \\ x_2^{'} &= (\text{idx}_r \% n + 0.5)/n, \quad & y_2^{'} &= (\text{idx}_r / n + 0.5)/m, \end{aligned}$$

where new coordinates is $\langle x_1^{'}, y_1^{'}, x_2^{'}, y_2^{'} \rangle$. Though the discretization reduces coordinate precision, it effectively decreases the representation complexity. In our dataset, we set $n = m = 20$, requiring at most **9** tokens ($2 \times 3 + 3$) for a box.

Algorithm 1 IQA-Filter & Box-Merge

Input: target image I, object bounding boxes $\mathcal{B}$, object quality $\mathcal{T}_q$, area threshold T_a, overlap threshold T_o
Output: the refined bounding boxes $\mathcal{R}$
Init: $\mathcal{R} \leftarrow \emptyset$
// IQA-Filter
for $b \in \mathcal{B}$ **do**
  $p \leftarrow \text{patch}(I, b)$; $q \leftarrow$"Is the image quality <$\mathcal{T}_q$>?"
  if Q-Instruct$(p, q) =$ 'Yes" **then**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{b\}$
  end if
end for
// Box-Merge
for $i = 0; i < |\mathcal{R}|; i \leftarrow i + 1$ **do**
  $j \leftarrow i + 1$
  while $j < |\mathcal{R}|$ **do**
    if area$(\mathcal{R}[i]) < T_a$ **and** is-touch$(\mathcal{R}[i], \mathcal{R}[j])$ **or** coverage-ratio$(\mathcal{R}[i], \mathcal{R}[j]) > T_o$ **then**
      $\mathcal{R}[i] \leftarrow \text{merge}(\mathcal{R}[i], \mathcal{R}[j]); \mathcal{R} \leftarrow \mathcal{R} \setminus \{\mathcal{R}[j]\}$
    else
      $j \leftarrow j + 1$
    end if
  end while
end for
return $\mathcal{R}$

Algorithm 1. IQA-Filter & Box-Merge

In conclusion, the final **answer** is a natural language description with precise coordinates, as shown in Fig. 3a.

For GIQA-VQA.

The task requires that the **question** or **answer** relate to low-level attributes and include explicit spatial information (*i.e.*, bounding boxes). Inspired by previous work[9][26][27], we apply the LLM (*i.e.*, Llama3[25]) to generate the corresponding QA pairs from the descriptions in GIQA-DES (depicted in Fig. 3b). We use specific templates to generate diverse QA, as follows:

1. ***Binary questions ("Yes/No"):*** Answers are limited to "Yes" or "No". The "Yes" answer corresponds to questions inferred directly from the description. Conversely, quality questions that cannot be inferred are answered "No".
2. ***Open-ended questions ("What/Why/How"):*** These questions address low-level attributes or related context (e.g., "What distortion?"); cause analysis (e.g., "Why is quality poor?"); perceptual degree (e.g., "How is clarity?"). All answers are inferred from the description and given as short phrases (e.g., "Noise" and "Medium").

Meanwhile, we supply the LLM with the names of key objects/regions (with bounding boxes), constraining the QA to relate

to relevant entities. We also use keyword detection to filter out any unrelated QA pairs. Finally, we incorporate bounding box information into the generated QA pairs, forming the corresponding **question** and **answer**. This is achieved by applying the same process from Stage-4 of the GIQA-DES construction pipeline.

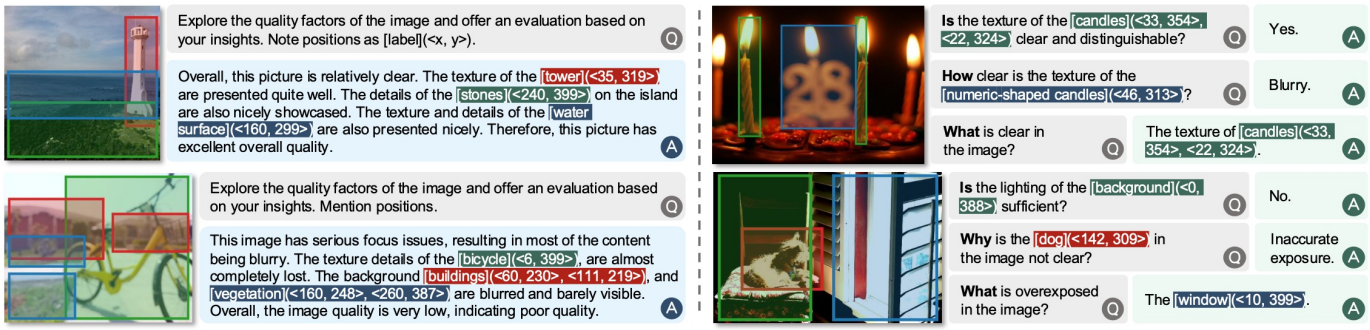

Figure 5. Some instances from the GIQA-160K dataset, involving two subtasks: GIQA-DES and GIQA-VQA.

Table 1. Statistics information of GIQA-160K and GIQA-Bench. DES: GIQA-DES; VQA: GIQA-VQA.

Dataset	Image	Total	DES	VQA (Y)	VQA (W)
GIQA-160K	42,960	167,657	66,689	50,484	50,484
GIQA-Bench	100	250	100	90	60

3.3. GIQA-160K

We construct our grounding-IQA dataset, GIQA-160K, utilizing the automated annotation pipeline, from existing public datasets[9][21]. Figure 5 contains some instances.

Data Source.

To build our dataset, we require two types of data: diverse images and their corresponding detailed quality descriptions. Currently, two public datasets, Q-Pathway[9] and DQ-495K[21], meet our requirements. For Q-Pathway, we select in-the-wild images (KonIQ-10K[53], SPAQ[54], LIVE-FB[3], and LIVE-itw[55]) and AI-generated images (AGIQA-3K[4] and ImageRewardDB[56]), along with their professionally human-annotated texts. The total image-text pairs is 53K. For DQ-495K, the descriptive texts with human annotations correspond to the images in KADIS-700K[57]. These are various types of artificial-degraded images, 27K in total.

Dataset Statistic.

Utilizing the above raw data (80K image-text pairs), we construct a dataset with**167,657** instruction-tuning samples and **42,960** images. Dataset statistics are shown in Tab. 1. For GIQA-DES, we generate 66,689 detailed quality descriptions with coordinates. The GIQA-VQA contains 100,968 question-answer pairs. For GIQA-VQA, to balance question types, we

randomly filter to maintain an equal amount of “Yes/No” and “What/Which/How” questions (50,484 each). Additionally, we ensured a balanced distribution between “Yes” and “No” responses, with 25,242 samples in each category.

3.4. GIQA-Bench

We construct a high-quality benchmark, GIQA-Bench, to evaluate the model grounding-IQA performance, detailing its data statistics and evaluation criteria.

Bench Statistic.

The GIQA-Bench includes 100 images of various types and quality, which are not included in GIQA-160K. We create 100 GIQA-DES and 150 GIQA-VQA test samples based on these images. Among the 150 GIQA-VQA data, 90 are of the “Yes/No” questions (“Yes”: 35; “No”: 55), and 60 are “What/Which/How” questions (“What”: 30; “Why”: 18; “How”: 12). Specifically, the descriptions for GIQA-DES are from Q-Pathway and adjusted, with key objects and corresponding bounding boxes manually determined. GIQA-VQA questions are generated by the annotation pipeline and further refined and answered by humans. Each sample is annotated in multiple rounds by at least three experts with relevant expertise in a controlled laboratory environment to ensure data accuracy.

Evaluation Criteria.

We evaluate the grounding-IQA capabilities from three perspectives: (1) Description quality; (2)VQA accuracy; and (3) Grounding precision. For all metrics, higher values indicate better performance.

1. ***Description quality.*** Assess GIQA-DES performance in quality descriptions. We compare the generated description to the ground truth, excluding coordinates for accuracy. We apply some image captioning metrics: ***BLEU@4***. Moreover, we employ the LLM (i.e., Llama3[25]) to provide a score from 0 to 4 (higher is better), based on the relevance between the description and the ground truth. For clarity, the final score is scaled proportionally from 0 to 100. We denote the score as the ***LLM-Score***.
2. ***VQA accuracy.*** Evaluate GIQA-VQA performance in image quality VQA. For “Yes/No” questions, accuracy is determined by matching with the word “Yes” or “No”. For “What/Which/How”, we use LLM to calculate accuracy. The LLM scores the model response from 0 to 4 (higher is better) based on the question and correct answer. The score is then normalized to 0~1. For clarity, we denote the accuracy of “Yes/No” as ***Acc (Y)***, “What/Which/How” as ***Acc (W)***, and overall accuracy as ***Acc (Total)***.
3. ***Grounding precision.*** Measure the grounding performance for both GIQA-DES and GIQA-VQA. We use category-agnostic Intersection over Union (***IoU***) to evaluate box quality. We also define ***Tag-Recall*** to assess category-specific grounding capabilities. In Tag-Recall, a result is true positive (TP) only if both the IoU and object name similarity exceeds a 0.5 threshold. For fairness, the bounding box is represented by the normalized corner coordinate.

Table 2. Ablation study on box optimization in the automated annotation pipeline. We conduct experiments on the GIQA-

DES task.

Table 2a. Box refinement.

Method	mIoU	Tag-Recall	BLEU@4	LLM-Score
Baseline	N/A	N/A	3.62	48.25
Raw-Box	0.5624	0.5045	20.97	61.00
Ref-Box	**0.5851**	**0.5497**	**23.67**	**61.75**

Table 2b. Box representation

Method	mIoU	Tag-Recall	BLEU@4	LLM-Score
Baseline	N/A	N/A	3.62	48.25
Norm-Coord	**0.6046**	0.5490	22.03	61.00
Disc-Coord	0.5851	**0.5497**	**23.67**	**61.75**

Table 3. Ablation study on multi-task training.

	GIQA-DES		GIQA-VQA	
Method	**Tag-Recall**	**LLM-Score**	**Tag-Recall**	**Acc (Total)**
Baseline	N/A	48.25	N/A	0.5633
Only-DES	**0.5497**	61.75	0.5577	0.5900
Only-VQA	0.3283	38.50	0.4872	0.7217
GIQA-160K	0.5474	**63.00**	**0.7372**	**0.7417**

Table 4. Ablation study on different baselines (data compatibility).

Method	SFT	GIQA-DES		GIQA-VQA	
		Tag-Recall	**LLM-Score**	**Tag-Recall**	**Acc (Total)**
LLaVA-1.5-7B		N/A	47.00	N/A	0.4733
	✓	**0.5283**	**60.00**	**0.5961**	**0.6850**
LLaVA-1.5-13B		N/A	49.00	N/A	0.4433
	✓	**0.5548**	**60.50**	**0.7564**	**0.6950**
LLaVA-1.6-7B		N/A	50.50	N/A	0.5067
	✓	**0.5981**	**60.00**	**0.6538**	**0.7250**
mPLUG-Owl-2-7B		N/A	48.25	N/A	0.5633
	✓	**0.5474**	**63.00**	**0.7372**	**0.7417**

4. Experiments

4.1. Experimental Settings

Implementation Details.

We conduct experiments on four pre-trained MLLM models: LLaVA-v1.5-7B[58], LLaVA-v1.5-13B[58], LLaVA-v1.6-7B[59], and mPLUG-Owl2-7B[18]. These models involve different versions, sizes, and architectures. The models are fine-tuned on our proposed GIQA-160K dataset using supervised fine-tuning. We evaluate their performance on grounding-IQA using the GIQA-Bench. Details about the training/testing datasets and evaluation criteria are provided in Secs. 3.3 and 3.4.

Training Settings.

We adopt cross-entropy loss for full fine-tuning, following previous methods[9][16][18]. The optimizer is AdamW[60], with β_1 = 0.9 and β_2 = 0.999. We apply the cosine decay scheduler with an initial learning rate of 2×10^{-5}, and a warmup ratio of 0.03. The batch size is set to 64, and the epoch is 2. Other hyper-parameters follow the default settings of each model. Experiments are implemented with PyTorch[61] on four Nvidia A100-80G GPUs.

4.2. Ablation Study

We conduct experiments in this section, analyzing data design and data properties. The training settings are consistent with Sec. 4.1. We apply mPLUG-Owl2-7B[18] as the baseline in all experiments (except in Tab. 4).

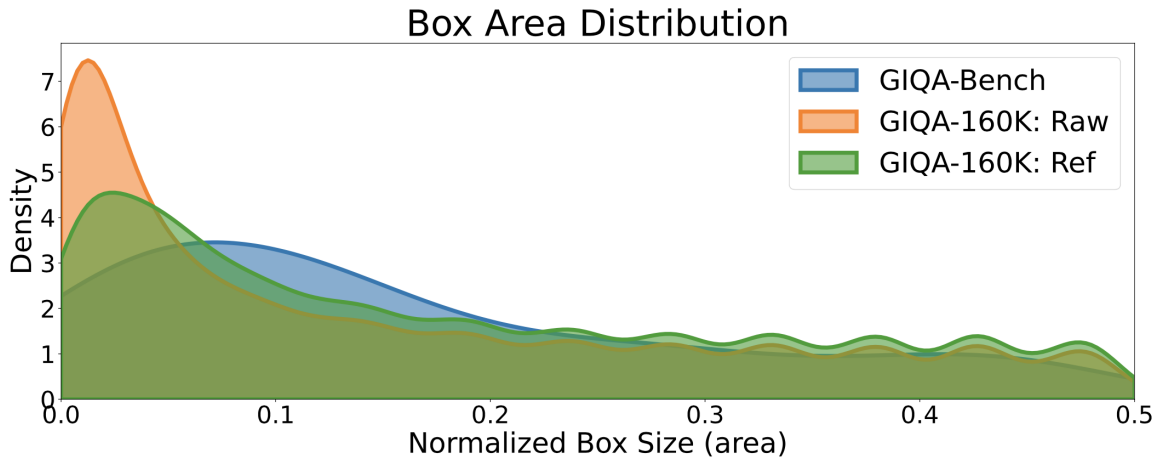

Figure 6. Box area distribution of GIQA-160K (Raw and Ref) and GIQA-Bench, showing data in 0~0.5 to highlight differences.

Box Optimization.

We evaluate box optimization in annotation pipeline, including box refinement (filter and merge) and representation. We compare the models trained on GIQA-DES with (Ref-Box) and without refinement (Raw-Box) in Tab. 2a. The refinement enhances the fine-tuning effect. We also visualize box area distribution in Fig. 6. Refinement reduces the difference between automatically annotated GIQA-160K and human-annotated GIQA-Bench.

Meanwhile, we compare discrete (Disc-Coord) and normalized continuous (Norm-Coord) box representations in Tab. 2b. Results indicate that Disc-Coord enhances description quality (BLEU@4 and LLM-Score) and grounding accuracy (Tag-

Recall), compared with Norm-Coord.

Multi-Task Training.

We conduct an ablation on multi-task (GIQA-DES/VQA) joint training. The results are listed in Tab. 3. We observe that only GIQA-DES (Only-DES) can improve the quality assessment and grounding, while GIQA-VQA improves VQA accuracy, but grounding ability is limited. Moreover, multi-task training (GIQA-160K) enhances performance on both GIQA-DES/VQA. It demonstrates the importance of data diversity.

Data Compatibility.

We fine-tune various baselines using the proposed GIQA-160K. The results are provided in Tab. 4. The results indicate that our dataset is compatible with various MLLMs, effectively enhancing the grounding-IQA ability of the model. Furthermore, we provide more detailed comparisons with more methods in Sec. 4.3.

Table 5. Quantitative results on GIQA-Bench. The best and second-best results are colored red and blue.

Group	Method	GIQA-DES				GIQA-VQA				
		mIoU	Tag-Recall	BLEU@4	LLM-Score	mIoU	Tag-Recall	Acc (Y)	Acc (W)	Acc (Total)
General	LLaVA-v1.5-7B[58]	N/A	N/A	2.82	47.00	N/A	N/A	0.4444	0.5167	0.4733
	LLaVA-v1.5-13B[58]	N/A	N/A	3.00	49.00	N/A	N/A	0.3888	0.5250	0.4433
	LLaVA-v1.6-7B[59]	N/A	N/A	3.04	50.50	N/A	N/A	0.4889	0.5333	0.5067
	mPLUG-Owl2-7B[18]	N/A	N/A	3.62	48.25	N/A	N/A	0.5889	0.5250	0.5633
Ground	Shikra-7B[23]	0.4506	0.4768	0.40	27.00	0.4126	0.4359	0.5333	0.3917	0.4767
	Kosmos-2-1.6B[17]	0.4946	0.3448	2.63	39.25	0.4982	0.4103	0.3889	0.4750	0.4233
	Ferret-7B[26]	**0.6458**	**0.6778**	3.16	43.75	0.5393	0.5769	0.4111	0.4875	0.4417
	GroundingGPT-7B[27]	0.4967	0.5391	1.99	32.50	0.3845	0.5321	0.5444	0.5250	0.5367
IQA	DepictQA-Wild-7B[21]	N/A	N/A	3.34	56.50	N/A	N/A	0.4333	0.5458	0.4783
	Q-Instruct[9] (LLaVA-v1.5-7B)	N/A	N/A	**22.69**	58.25	N/A	N/A	0.6444	0.5375	0.6017
	Q-Instruct[9] (LLaVA-v1.5-13B)	N/A	N/A	19.01	57.25	N/A	N/A	0.6222	0.5417	0.5900
	Q-Instruct[9] (mPLUG-Owl2-7B)	N/A	N/A	21.46	**62.00**	N/A	N/A	0.6111	0.5375	0.5817
Ours	Grounding-IQA (LLaVA-v1.5-7B)	0.5763	0.5283	19.02	60.00	0.5180	0.5961	0.7777	0.5458	0.6850
	Grounding-IQA (LLaVA-v1.5-13B)	0.6302	0.5548	20.24	60.50	**0.6830**	**0.7564**	0.7889	0.5542	0.6950
	Grounding-IQA (LLaVA-v1.6-7B)	**0.6583**	**0.5981**	19.17	60.00	0.5459	0.6538	**0.8333**	**0.5625**	**0.7250**
	Grounding-IQA (mPLUG-Owl2-7B)	0.5955	0.5474	**22.87**	**63.00**	**0.6031**	**0.7372**	**0.8444**	**0.5875**	**0.7417**

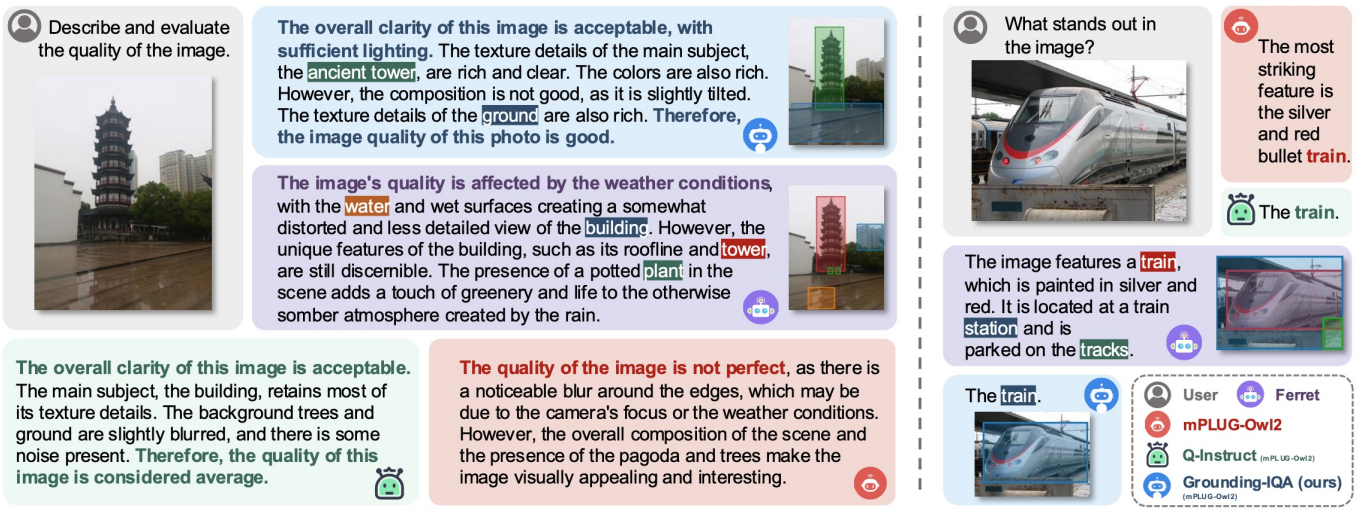

Figure 7. Visual comparisons on GIQA-Bench. Our proposed grounding-IQA (blue module) enables more fine-grained quality descriptions (left instance) and QA (right instance) with precise position (*i.e.*, bounding box).

4.3. Results on GIQA-Bench

In GIQA-Bench, we compare four groups of MLLMs with different functionalities,*i.e.*, (1) General models (General): LLaVA-v1.5-7B[58], LLaVA-v1.5-13B[58], LLaVA-v1.6-7B[59], and mPLUG-Owl2-7B[18]; (2)Multimodal referring and grounding models (Ground): Shikra-7B[23], Kosmos-2-1.6B[17], Ferret-7B[26], and GroundingGPT-7B[27]; (3) IQA models (IQA): DepictQA-Wild-7B[21] and Q-Instruct[9] (fine-tuned three base models); and (4) Our methods (Ours): Four general models fine-tuned on GIQA-160K. Detailed testing prompts for all models are provided in the supplementary material.

Quantitative Results.

We evaluate all models on GIQA-DES and GIQA-VQA tasks from two aspects: quality assessment and referring/grounding ability, as in Tab. 5. General models perform poorly on both tasks, while task-specific models are more effective in their respective domains. Specifically, grounding MLLMs (*e.g.*, Shikra[23]) perform well on general grounding tasks, but show decreased performance when dealing with quality-related objects/areas (GIQA-VQA, Tag-Recall). Conversely, IQA models (*e.g.*, Q-Instruct[9]) excel in description quality (GIQA-DES, LLM-Score), but exhibit low accuracy in GIQA-VQA. In contrast, our models outperform MLLMs specialized in grounding or IQA tasks in both aspects.

Qualitative Results.

We provide some visual comparisons in Fig. 7. For GIQA-DES (left instance), the quality descriptions generated by general (mPLUG-Owl2-7B[18]) and grounding (Ferret[26]) MLLMs are unsatisfactory. In addition, the ground results of Ferret include many quality irrelevant objects, which is impractical for real applications. In contrast, our method describes image quality more properly with coordinates of key objects affecting the quality. Furthermore, in the GIQA-VQA task (right instance), our method produces more accurate responses to image quality VQA involving spatial perception. More qualitative results are provided in the supplementary material.

5. Conclusion

In this paper, we introduce a new IQA task paradigm called Grounding-IQA for fine-grained quality assessments. The grounding-IQA combines multimodal referring and grounding with IQA, and comprises two subtasks: GIQA-DES and GIQA-VQA. Under the task paradigm, we construct a corresponding dataset, GIQA-160K, by an automated annotation pipeline. Meanwhile, we develop a benchmark, GIQA-Bench, to evaluate the grounding-IQA. Experiments indicate that our proposed task, dataset, and benchmark facilitate more fine-grained IQA applications.

References

1. [a, b, c] *Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018). "The unreasonable effectiveness of deep features as a perceptual metric". CVPR.*
2. [^] *Lin H, Hosu V, Saupe D (2019). "Kadid-10k: A large-scale artificially distorted iqa database".QoMEX.*
3. [a, b] *Ying Z, Niu H, Gupta P, Mahajan D, Ghadiyaram D, Bovik A (2020). "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality". CVPR. 2020.*
4. [a, b] *Li C, Zhang Z, Wu H, Sun W, Min X, Liu X, Zhai G, Lin W (2023). "Agiqa-3k: An open database for ai-generated image quality assessment". TCSVT. 2023.*
5. [a, b, c] *Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004). "Image quality assessment: from error visibility to structural similarity". TIP.*
6. [a, b] *Mittal A, Moorthy AK, Bovik AC (2012). "No-reference image quality assessment in the spatial domain".TIP.*
7. [^] *Ding K, Ma K, Wang S, Simoncelli EP (2020). "Image quality assessment: Unifying structure and texture similarity". TPAMI.*
8. [a, b] *Wang J, Chan KC, Loy CC (2023). "Exploring clip for assessing the look and feel of images".AAAI.*
9. [a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q] *Wu H, Zhang Z, Zhang E, Chen C, Liao L, Wang A, Xu K, Li C, Hou J, Zhai G, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In CVPR, 2024.*
10. [a, b] *Mittal A, Soundararajan R, Bovik AC (2012). "Making a 'completely blind' image quality analyzer".SPL.*
11. [a, b] *Kang L, Ye P, Li Y, Doermann D. "Convolutional neural networks for no-reference image quality assessment". CVPR. 2014.*
12. [a, b] *Bosse S, Maniry D, Müller KR, Wiegand T, Samek W (2017). "Deep neural networks for no-reference and full-reference image quality assessment". TIP.*
13. [a, b] *Gu J, Cai H, Chen H, Ye X, Ren JS, Dong C (2020). "Pipal: a large-scale image quality assessment dataset for perceptual image restoration". ECCV. 2020.*
14. [a, b] *Ke J, Wang Q, Wang Y, Milanfar P, Yang F (2021). "Musiq: Multi-scale image quality transformer".ICCV.*
15. [a, b] *You Z, Li Z, Gu J, Yin Z, Xue T, Dong C. "Depicting beyond scores: Advancing image quality assessment through multi-modal language models." In: ECCV; 2024.*
16. [a, b, c] *Liu H, Li C, Wu Q, Lee YJ. "Visual instruction tuning." In:NeurIPS, 2023.*

17. a, b, c, d, e, f, g *Peng Z, Wang W, Dong L, Hao Y, Huang S, Ma S, Wei F. Kosmos-2: Grounding multimodal large language models to the world. ICLR. 2024.*
18. a, b, c, d, e, f, g, h, i *Ye Q, Xu H, Ye J, Yan M, Hu A, Liu H, Qian Q, Zhang J, Huang F. "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration." In CVPR, 2024.*
19. ^ *Wu H, Zhang Z, Zhang E, Chen C, Liao L, Wang A, Li C, Sun W, Yan Q, Zhai G, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In ICLR, 2024.*
20. ^ *Wu T, Ma K, Liang J, Yang Y, Zhang L. "A comprehensive study of multimodal large language models for image quality assessment." In: ECCV, 2024.*
21. a, b, c, d, e, f, g, h, i *You Z, Gu J, Li Z, Cai X, Zhu K, Xue T, Dong C (2024). "Descriptive image quality assessment in the wild". arXiv preprint arXiv:2405.18842. arXiv:2405.18842.*
22. ^ *Mao J, Huang J, Toshev A, Camburu O, Yuille AL, Murphy K. Generation and comprehension of unambiguous object descriptions. In: CVPR; 2016.*
23. a, b, c, d, e, f *Chen K, Zhang Z, Zeng W, Zhang R, Zhu F, Zhao R (2023). "Shikra: Unleashing multimodal llm's referential dialogue magic". arXiv preprint arXiv:2306.15195. Available from: https://arxiv.org/abs/2306.15195.*
24. a, b *Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Jiang Q, Li C, Yang J, Su H, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV; 2024.*
25. a, b, c, d, e *Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024.*
26. a, b, c, d, e, f, g *You H, Zhang H, Gan Z, Du X, Zhang B, Wang Z, Cao L, Chang SF, Yang Y. "Ferret: Refer and ground anything anywhere at any granularity." In: ICLR, 2024.*
27. a, b, c, d, e, f *Li Z, Xu Q, Zhang D, Song H, Cai Y, Qi Q, Zhou R, Pan J, Li Z, Tu V, et al. Groundinggpt: Language enhanced multi-modal grounding model. In ACL, 2024.*
28. ^ *Zhang L, Zhang L, Mou X, Zhang D (2011). "Fsim: a feature similarity index for image quality assessment".TIP.*
29. ^ *Moorthy AK, Bovik AC (2011). "Blind image quality assessment: From natural scene statistics to perceptual quality". TIP.*
30. ^ *Saad MA, Bovik AC, Charrier C (2012). "Blind image quality assessment: A natural scene statistics approach in the DCT domain". TIP.*
31. ^ *Liu X, Van De Weijer J, Bagdanov AD. Rankiqa: Learning from rankings for no-reference image quality assessment. In: ICCV, 2017.*
32. ^ *Yang S, Wu T, Shi S, Lao S, Gong Y, Cao M, Wang J, Yang Y (2022). "Maniqa: Multi-dimension attention network for no-reference image quality assessment". CVPRW.*
33. ^ *Chen C, Mo J, Hou J, Wu H, Liao L, Sun W, Yan Q, Lin W. Topiq: A top-down approach from semantics to distortions for image quality assessment. TIP. 2024.*
34. ^ *Zhu H, Li L, Wu J, Dong W, Shi G (2020). "Metaiqa: Deep meta-learning for no-reference image quality assessment". CVPR.*

35. ^Zhang W, Zhai G, Wei Y, Yang X, Ma K (2023). "Blind image quality assessment via vision-language correspondence: A multitask learning perspective". CVPR.
36. ^Sun S, Yu T, Xu J, Zhou W, Chen Z (2022). "Graphiqa: Learning distortion graph representations for blind image quality assessment". TMM.
37. ^Zhu D, Chen J, Shen X, Li X, Elhoseiny M (2023). "Minigpt-4: Enhancing vision-language understanding with advanced large language models". arXiv preprint arXiv:2304.10592. Available from: https://arxiv.org/abs/2304.10592.
38. ^Zhang P, Dong X, Wang B, Cao Y, Xu C, Ouyang L, Zhao Z, Duan H, Zhang S, Ding S, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112. 2023.
39. ^Li J, Li D, Savarese S, Hoi S. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." In ICML, 2023.
40. ^Wu H, Zhu H, Zhang Z, Zhang E, Chen C, Liao L, Li C, Wang A, Sun W, Yan Q, et al. Towards open-ended visual quality comparison. In: ECCV, 2024.
41. a, b Chen C, Yang S, Wu H, Liao L, Zhang Z, Wang A, Sun W, Yan Q, Lin W. "Q-ground: Image quality grounding with large multi-modality models." ACM MM. 2024.
42. ^Wu H, Zhang Z, Zhang W, Chen C, Liao L, Li C, Gao Y, Wang A, Zhang E, Sun W, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In ICML, 2024.
43. ^Krahmer E, Van Deemter K (2012). "Computational generation of referring expressions: A survey".Computational Linguistics.
44. ^Zellers R, Bisk Y, Farhadi A, Choi Y (2019). "From recognition to cognition: Visual commonsense reasoning".CVPR.
45. ^Kazemzadeh S, Ordonez V, Matten M, Berg T. Referitgame: Referring to objects in photographs of natural scenes. EMNLP. 2014.
46. ^Luo R, Shakhnarovich G (2017). "Comprehension-guided referring expressions".CVPR.
47. ^Lai X, Tian Z, Chen Y, Li Y, Yuan Y, Liu S, Jia J. "Lisa: Reasoning segmentation via large language model." In: CVPR, 2024.
48. ^Ren Z, Huang Z, Wei Y, Zhao Y, Fu D, Feng J, Jin X (2024). "Pixellm: Pixel reasoning with large multimodal model". CVPR. 2024.
49. ^Zhang S, Sun P, Chen S, Xiao M, Shao W, Zhang W, Liu Y, Chen K, Luo P (2023). "Gpt4roi: Instruction tuning large language model on region-of-interest". arXiv preprint arXiv:2307.03601. Available from: https://arxiv.org/abs/2307.03601.
50. ^Rasheed H, Maaz M, Shaji S, Shaker A, Khan S, Cholakkal H, Anwer RM, Xing E, Yang MH, Khan FS. "Glamm: Pixel grounding large multimodal model". In: CVPR, 2024.
51. a, b, c Zhou Y, Wang M, Liu D, Hu Z, Zhang H (2020). "More grounded image captioning by distilling image-text matching model". CVPR.
52. ^Liu H, Li C, Wu Q, Lee YJ. "Visual instruction tuning." In:NeurIPS, 2023.

53. ^Hosu V, Lin H, Sziranyi T, Saupe D (2020). "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment". TIP.
54. ^Fang Y, Zhu H, Zeng Y, Ma K, Wang Z (2020). "Perceptual quality assessment of smartphone photography"CVPR.
55. ^Ghadiyaram D, Bovik AC (2015). "Massive online crowdsourced study of subjective and objective picture quality". TIP.
56. ^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: NeurIPS, 2024.
57. ^Lin H, Hosu V, Saupe D (2020). "DeepFL-IQA: Weak supervision for deep IQA feature learning".arXiv preprint arXiv:2001.08113. Available from: https://arxiv.org/abs/2001.08113.
58. a, b, c, d, e, fLiu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. In:CVPR; 2024.
59. a, b, cLiu H, Li C, Li Y, Li B, Zhang Y, Shen S, Lee YJ (2024). "Llava-next: Improved reasoning, OCR, and world knowledge".
60. ^Loshchilov I, Hutter F, et al. Fixing weight decay regularization in adam. In:ICLR, 2018.
61. ^Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS; 2019.