A Benchmark for Multi-modal Foundation Models on Low-level Vision: from Single Images to Pairs

Zicheng Zhang*, Haoning Wu*, Erli Zhang,
Guangtao Zhai†, *Senior Member, IEEE,* and Weisi Lin†, *Fellow, IEEE*

***Abstract*—The rapid development of Multi-modality Large Language Models (MLLMs) has navigated a paradigm shift in computer vision, moving towards versatile foundational models. However, evaluating MLLMs in *low-level visual perception and understanding* remains a yet-to-explore domain. To this end, we design benchmark settings to *emulate human language responses* related to low-level vision: the low-level visual *perception* (A1) *via* visual question answering related to low-level attributes (*e.g. clarity, lighting*); and the low-level visual *description* (A2), on evaluating MLLMs for low-level text descriptions. Furthermore, given that pairwise comparison can better avoid ambiguity of responses and has been adopted by many human experiments, we further extend the low-level perception-related question-answering and description evaluations of MLLMs from single images to *image pairs*. Specifically, for *perception* (A1), we carry out the LLVisionQA$^+$ dataset, comprising 2,990 single images and 1,999 image pairs each accompanied by an open-ended question about its low-level features; for *description* (A2), we propose the LLDescribe$^+$ dataset, evaluating MLLMs for low-level descriptions on 499 single images and 450 pairs. Additionally, we evaluate MLLMs on *assessment* (A3) ability, *i.e.* predicting score, by employing a softmax-based approach to enable all MLLMs to generate *quantifiable* quality ratings, tested against human opinions in 7 image quality assessment (IQA) datasets. With 24 MLLMs under evaluation, we demonstrate that several MLLMs have decent low-level visual competencies on single images, but only GPT-4V exhibits higher accuracy on pairwise comparisons than single image evaluations (*like humans*). We hope that our benchmark will motivate further research into uncovering and enhancing these nascent capabilities of MLLMs. Datasets will be available at https://github.com/Q-Future/Q-Bench.**

***Index Terms*—Multi-modality large language models, low-level vision, benchmark, perception, description, assessment**

I. Introduction

The emergent large language models (LLMs) such as ChatGPT and Bard, as well as their excellent open-source counterparts (*e.g.*, LLaMA [1], MPT [2]), have served as powerful general-purpose assistants, which opens a new era for artificial intelligence (AI) from targeting specific tasks towards general intelligence. Following the advancements of LLMs, multi-modality large language models (MLLMs), as represented by LLaVA [3], MiniGPT-4 [4], InstructBLIP [5], and Otter [6], have brought exciting progresses on the vision field as well. They are capable of providing robust general-level abilities on visual perception/understanding and can even seamlessly dialog and interact with humans through natural language. While such abilities of MLLMs have been validated on several vision-language tasks such as image captioning [7], visual question answering [8], cross-modality grounding [9], and traditional vision tasks such as image classification or segmentation [10], most attention is paid to the high-level perception and understanding of visual contents. Meanwhile, the ability of MLLMs remains unclear on **low-level visual perception and understanding**, which play significant roles in image quality assessment (IQA) [11], [12] and its associated tasks on perceiving visual distortions (*noises, blurs*) [13], [14], and other low-level attributes (*color, lighting, composition, style, etc*) [15] that may relate to aesthetics of natural photos [16] as well as human preferences on emerging computer-graphics generated [17] or AI-generated images [18], [19]. These low-level visual abilities are strongly associated with a wide range of applications, such as recommendation [20], guidance on camera systems [21], or visual quality enhancement [22]. Henceforth, it is crucial to evaluate these general-purpose foundation models in low-level visual perception and understanding, to relieve extensive human resources on giving feedback to every specific low-level task.

In this paper, we propose the first systematic benchmark **Q-Bench**$^+$ to measure the low-level visual abilities of MLLMs, which is constructed around a key question:

How do MLLMs emulate human ability related to low-level visual perception and understanding?

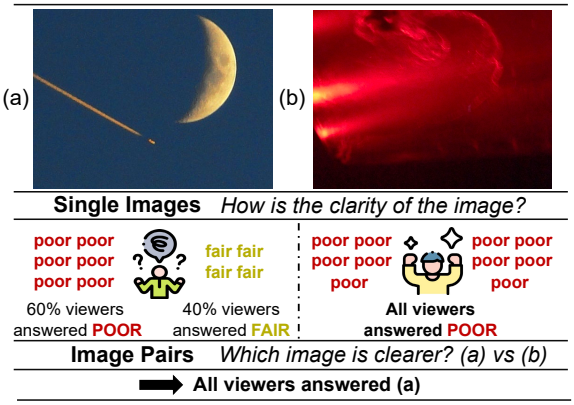

Fig. 1. **Pairwise comparison** is a non-negligible setting for human to perceive and evaluate low-level visual attributes, as it provides additional and non-ambiguous information (***(a)** is clearer than **(b)***). Henceforth, we extend into the **Q-Bench**$^+$ with image pairs to examine whether MLLMs can *extract and compare low-level visual information between a pair of images, like a human*.

Zicheng Zhang and Guangtao Zhai are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China. E-mail:{zzc1998,zhaiguangtao} @sjtu.edu.cn.

Haoning Wu and Erli Zhang are with S-Lab, Nanyang Technological University, Singapore. E-mail: {haoning001,ezhang005}@e.ntu.edu.sg.

Weisi Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. E-mail: wslin@ntu.edu.sg.

*Equal Contributions. †Corresponding Authors.

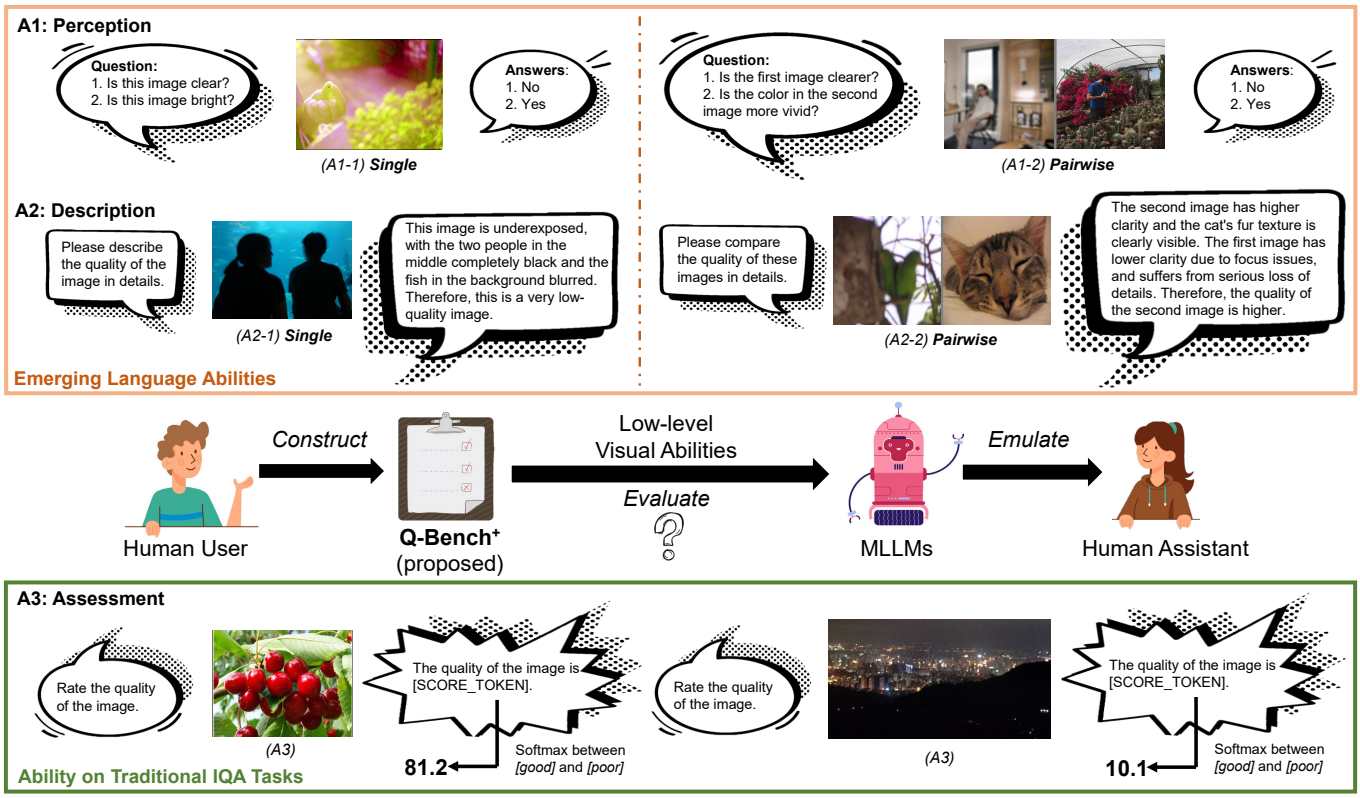

Fig. 2. In the proposed **Q-Bench**$^+$, we build the first benchmark on emerging abilities of MLLMs for low-level vision, including **perception** of single/pairwise low-level attributes (*by correctly answering diverse queries*) and **description** of single/pairwise low-level quality-related information via natural language. Furthermore, the **Q-Bench**$^+$ also evaluates the quantifiable **assessment** ability of MLLMs on traditional IQA tasks.

A basic answer comes from the fundamental capability of MLLMs: **vision-conditioned language generation**. Specifically, for low-level vision, MLLMs should ideally be able to correctly answer low-level visual questions and precisely describe the low-level information of single images. Henceforth, we define the following two emerging abilities of MLLMs that directly arise from their language generation capability:

***Ability 1: Perception** of Low-level Attributes.* As shown in Fig. 2 (A1-1), like a human, an MLLM should be able to respond accurately to simple questions related to low-level attributes, *e.g* answering *'No'* for a blurry image when queried with *'Is this image clear?'*.

***Ability 2: Description** via Natural Language*. As shown in Fig. 2 (A2-1), like a human, an MLLM should be able to describe the quality and other low-level related attributes for single images with natural language. The descriptions should be both complete and accurate.

Although the above two capabilities essentially emulate human perception of low-level vision, they still miss some key capabilities of humans. For example, regarding Fig. 1 (a), some people may consider its clarity to be average, while others may deem it poor, while neither opinion should be considered incorrect; instead, everyone would agree that Fig. 1 (b) *is clearer than* Fig. 1 (a). On the other hand, for those who regard both Fig. 1 (b) and Fig. 1 (a) as blurry, comparing the clarity between the pair can also provide additional valuable information. Noticing these issues, lots of recent subjective studies [22], [23], [24] have adopted the a juxtaposition-based paradigm, that is, collecting human opinions by comparing **a pair of images**. Based on these insights and recent progresses on MLLMs [25], [26], [27], [28] that officially support more than one images as inputs, we further explore whether MLLMs can similarly emulate respective human capabilities:

Can MLLMs adeptly extract and compare low-level visual information between a pair of images?

On answering this question, we further extend the **Perception** and **Description** tasks from single images to image pairs:

***Extended Ability 1: Perception** of Low-level Attributes for **image pairs**.* As shown in Fig. 2 (A1-2), like a human, an MLLM should be able to respond correctly to low-level questions for image pairs, *e.g* answering *'No'* for image pair (first blurrier) when queried with *'Is the first image clearer?'*.

***Extended Ability 2: Description** via Natural Language for **image pairs**.* As shown in Fig. 2 (A2-2), like a human, an MLLM should be able to describe the similarities (*joint information*) and differences (*comparison*) of low-level appearances between a pair of images with natural language.

Despite the direct and concrete abilities above, we also evaluate how MLLMs can perform on the traditional IQA task, a highly abstract task that requires understanding on how the low-level attributes affect human judgements, as follows:

***Ability 3:** Precise **Assessment** Aligned with Human Opinions.* As depicted in Fig. 2 (A3), an MLLM should be able to predict **quantifiable** quality scores for images, which can be aligned with the human-rated mean opinion scores (MOS).

To evaluate the three abovementioned abilities, we formulate their respective benchmark settings, as follows:

*1) **LLVisionQA$^+$ Benchmark Dataset**:* To evaluate the low-level **perception** ability (A1) on various low-level attributes under diverse circumstances, we construct the **LLVisionQA$^+$** dataset, including 2,990 single images and 2,000 image pairs from 10 diverse sources. Aligned with existing practices [29], [30], each single image or image pair in **LLVisionQA$^+$** is equipped with a question, alongside a correct answer and false candidate answers. Specifically, we design three diverse types of questions: *Yes-or-No* questions, *What* questions, and *How* questions. Moreover, we divide low-level concerns for single images into four quadrants, via two axes: **(1)** distortions (*blur, noises, etc*) *vs* other low-level attributes (*color, lighting, composition, etc*) [31]. **(2)** global perception (*e.g., sharpness of the whole picture*) *vs* local content-related in-context perception (*e.g., whether the red flower is in focus*) [32]. On the other hand, we separate the low-level concerns for image pairs into four sub-categories: **(1)** distortions *vs* other low-level attributes (*similar as above*). **(2)** comparison (*e.g., which image is clearer*) *vs* joint analysis (*e.g., are both images underexposure*). With three types of questions and divided concerns, the proposed **LLVisionQA$^+$** dataset provides a holistic benchmark for the low-level perception abilities of MLLMs on both single images and pairs.

*2) **LLDescribe$^+$ Benchmark Dataset**:* For the **description** ability (A2), given that the output description is expected to be open-ended, we propose the **LLDescribe$^+$** dataset by inviting experts with rich experience in the low-level vision field to write long *golden* low-level descriptions (*average **58** words per description*) for 499 single images and 450 image pairs. The long *golden* low-level descriptions then serve as the reference texts for the single-modal GPT to evaluate MLLM output descriptions. To ensure the evaluation is comprehensive, the quality of MLLM descriptions is evaluated through three dimensions: completeness (*punish missing information*), preciseness (*punish outputs controversial with reference*), as well as relevance (*punish outputs irrelevant to low-level attributes*). With *golden* descriptions and the multi-dimensional evaluation process participated by GPT, we comprehensively evaluate the low-level description ability of MLLMs.

*3) **IQA Benchmark**:* For the **assessment** ability, we utilize plenty of existing IQA databases [11], [33], [18], [17], [12], [34] that focus on various low-level appearances of images, to benchmark MLLMs within conventional IQA settings. Specifically, we notice that MLLMs encounter difficulties in providing sufficiently **quantifiable** outputs, whether instructed to directly rate with texts or provide numerical outputs. To solve this challenge, we propose to extract the `softmax` pooling result on the logits of the two most frequent tokens (***good*** and ***poor***) under the response template of MLLMs (Fig. 2 (A3)) as their quality predictions. Our studies prove that the proposed softmax-based strategy is generally better correlated with human perception than direct token outputs of MLLMs (via `argmax`), which bridges between these emergent MLLMs and the traditional IQA task settings. Under this strategy, we evaluate all MLLMs on their precise assessment ability by measuring the correlations between their predictions and human opinion scores in various IQA databases. Furthermore, we propose a **prompt-ensemble** approach to help boost the IQA performance of MLLMs with the softmax-based strategy.

TABLE I
OVERVIEW OF THE 10 DIVERSE IMAGE SOURCE DATASETS IN THE **Q-BENCH$^+$**, AND THE RESPECTIVE BENCHMARK DATASET SIZE FOR EACH LOW-LEVEL ABILITY AMONG **PERCEPTION**, **DESCRPTION** AND **ASSESSMENT**. THE *Corrupted* COCO DENOTES COCO-CAPTIONS IMAGES CORRUPTED BY [36].

Type	**Source Dataset**	**LLVisionQA$^+$** Sampled Size	**LLDescribe$^+$** Sampled Size	Full Dataset Size for A3 Task
In-the-wild	KONiQ-10K [11]	600	200	10,073
	SPAQ [12]	800	200	11,125
	LIVE-FB [37]	300	50	39,810
	LIVE-itw [38]	300	50	1,169
Generated	CGIQA-6K [17]	200	50	6,000
	AGIQA-3K [18]	198	80	2,982
	ImageRewardDB [19]	194	29	*excluded in* (A3)
Manually-distorted	KADID-10K [33]	81	20	10,125
	LIVEMultiDistortion [39]	15	10	*excluded in* (A3)
	Corrupted COCO [7]	302	50	*excluded in* (A3)
Corresponding Task in **Q-Bench$^+$**		(A1) **Perception**	(A2) **Description**	(A3) **Assessment**
Benchmark Size (single+pairwise)		2,990+2,000	499+450	81,284

This work is a substantially extended version of our earlier conference publication [35]. Compared with the conference version, we bring three major changes: **(1)** Most importantly, we extend the **perception** and **description** tasks from single images to image pairs, which provides a more comprehensive benchmark for MLLMs on emulating human low-level visual understanding ability. **(2)** We update the benchmark with the latest popular MLLMs (evaluated MLLMs increased from 15 to **24**), providing a review of the development for MLLMs on low-level vision. **(3)** We further propose a simple yet effective prompt-ensemble approach, which can help boost the zero-shot performance of MLLMs on the **assessment** task.

In summary, we systematically explore the potential of MLLMs on three low-level visual abilities: **perception**, **description**, and **assessment**. The three realms compose into the proposed **Q-Bench$^+$**, a MLLM benchmark on low-level visual tasks. Our contributions can be summarized as three-fold:

- We build a benchmark for MLLMs on low-level **perception** ability. To achieve this, we construct a first-of-its-kind balanced and comprehensive **LLVisionQA$^+$** dataset with 2,990 single images and 2,000 image pairs with one low-level-related question-answer pair for each image. The **LLVisionQA$^+$** dataset includes three question types and multiple low-level concerns to ensure diversity.
- We define a benchmark process to evaluate the low-level **description** ability of MLLMs, including an **LLDescription$^+$** dataset of 499 single images and 450 image pairs with expert-labeled long golden quality descriptions, and a GPT-assisted evaluation to rate MLLM-descriptions in terms of completeness, preciseness, and relevance compared with golden descriptions.
- To evaluate precise quality **assessment** ability, we propose a unified **softmax-based** quality prediction strategy for all MLLMs based on their probability outputs. Furthermore, we propose a prompt-ensemble approach to help boost the IQA performance of MLLMs with the softmax-based strategy. With its effectiveness validated in our experiments, the proposed strategy sets up a bridge between general-purpose MLLMs and traditional IQA tasks that requires **quantifiable** scores as outputs.

II. Constructing the Q-Bench$^+$

A. General Principles

1) Focusing on Low-level Visual Abilities of MLLMs: Unlike existing MLLM benchmarks [40], [29], [30] that aim at all-round abilities, the tasks in **Q-Bench**$^+$ are constrained with two basic principles: a) Requiring perception and/or understanding on low-level attributes of images; b) Not requiring reasoning (*i.e. why*) or outside knowledge [41]. We adhere to the principles in designing the **perception**, **description**, and **assessment** tasks, making the proposed **Q-Bench**$^+$ a focused reflection on the low-level visual abilities of MLLMs.

2) Covering Diverse Low-level Appearances: To cover diverse low-level appearances, we collect multi-sourced images for each task, as depicted in Table I. Among all images in the **perception** and **description** tasks, *two-thirds* are in-the-wild images directly collected from social media posts, smartphones, or professional photography. The rest *one-third* images are collected after various artificial distortions, or via generative processes (CGI, AIGC). Furthermore, we employ k-means clustering for the low-level attribute indicators to certify that the sub-sampled images retain high diversity. In the **assessment** task, full images of 7 IQA datasets within all three source types are evaluated through traditional IQA metrics. The diverse and multiple sources of images morph the **Q-Bench**$^+$ into a holistic and balanced benchmark to fairly evaluate low-level-related abilities.

3) Extending from Single Images to Image Pairs: Evaluating image pairs allows for direct comparison and joint analysis of low-level attributes, which can highlight subtle differences or similarities that might not be evident when images are viewed in isolation. Humans are good at comparing, therefore we believe it is also important to benchmark the *low-level visual perception and understanding ability* of MLLMs on image pairs. Thus we extend the benchmark (only including single images) in our conference version [35] with image pairs to simulate more complex visual tasks that mirror real-world scenarios and challenge the MLLMs to process and compare multiple visual inputs simultaneously.

B. Benchmark on Low-level *Perception* Ability

In the first task of **Q-Bench**$^+$, we evaluate the low-level **perception** ability of MLLMs to examine whether they can answer simple natural queries related to low-level attributes. For this purpose, we first collect 2,990 single images (`I`) from multiple sources (see Table I) with diverse low-level concerns, from which we collect 2,000 image pairs (`I'`) as well. All image pairs are different from each other but may have one repeated image across different pairs. Then, we collect one low-level-related question (`Q`), one correct answer to the question (`C`), and 1-3 candidate false answers (`F`) for each single image or image pair. The 2,990 `(I,Q,C,F)` and 2,000 `(I',Q,C,F)` tuples compose into the **LLVisionQA**$^+$ dataset (as illustrated in Fig. 3), the first visual question answering (VQA) dataset in the low-level computer vision field. Specifically, the questions in **LLVisionQA**$^+$ cover four quadrants of distinct low-level concerns and three question types. After constructing the dataset, the `(I,Q,C,F)` are together fed into MLLMs for evaluation, while their outputs are further examined by GPT to judge correctness. The details are elaborated as follows.

1) Low-level Visual Concerns for Single Images

Axis 1: Distortions *vs* Other Low-level Attributes. The primary axis differentiates two categories of low-level perceptual attributes: **1)** technical **distortions** [13], seen as the low-level characteristics that directly degrade the quality of images [37], and **2)** aesthetic-related **other low-level attributes** [15], [42] which are discernible to human perception and evoke varied emotions. Several studies [43], [37], [31] follow this paradigm and categorize them through a relative golden standard, that whether the attributes *directly improve or degrade picture quality* (*Yes→Distortions; No→Others*).

Axis 2: Global Perception *vs* Local In-context Perception. In recent research on low-level vision, it is observed that human perceptions of low-level visuals often intertwine with higher-level contextual comprehension [32], [44], [45], [46]. For instance, a clear sky might lack complex textures yet display exceptional clarity. Furthermore, localized low-level appearances can deviate from their overall counterparts, as observed by [47], [48]. Acknowledging these differences, we curate **local in-context perception** (Fig. 3 *right top*) questions, that require MLLMs to grasp the content or other context to answer correctly, while other questions are categorized as **global perception** (Fig. 3 *left top*).

2) Low-level Visual Concerns for Image Pairs

Axis 1: Distortions *vs* Other Low-level Attributes. Same as Axis 1 for single images. Please refer to Sec. II-B *1)*.

Axis 2: Compare *vs* Joint. This dual approach mimics human visual perception more closely. Humans often use both comparison (looking at differences and similarities) and joint analysis (perceiving images in a unified context) when viewing images. The **comparison** highlights the differences and similarities between the two images, which is the key component of the full-reference IQA tasks [49] and other low-level enhancement evaluation tasks [50], [51]. The **joint analysis**, on the other hand, looks at the images as a combined entity to understand the overall context or to detect patterns that emerge only when the images are considered together.

3) Question Types

In the **LLVisionQA**$^+$ dataset, we curate three question types, *Yes-or-No*, *What*, and *How* to simulate multiple query forms from humans. The details of the three question types are defined as follows.

Type 1: *Yes-or-No* Questions. The fundamental type of questions is *Yes-or-No*, *i.e.*, judgments. Specifically, we notice that some MLLMs especially prefer to respond with *yes* rather than *no*. To reduce such biases in our benchmark, though designing questions with answers as *yes* is easier, we ensure that around 40% of all judgments are with correct answers as *no*, via querying on **contrastive** low-level attributes or **non-existing** low-level attributes.

Type 2: *What* Questions. Despite *Yes-or-No* judgments, the *what* questions are also a common type of queries in recent MLLM benchmarks such as [30]. In **Q-bench**$^+$, they classify low-level attributes in pictures (*e.g., What distortion occurs in the image?*), or associated context given specific low-level

Fig. 3. A dataset card of **LLVisionQA**$^+$ that evaluates the low-level **perception** ability of MLLMs. 2,990 (I,Q,C,F) and 1,999 (I',Q,C,F) tuples are collected to cover three question types and various low-level visual concerns, providing an all-around evaluation of low-level visual perception for MLLMs.

appearances (for in-context perception questions, *e.g., Which object in the image is under-exposed?*). Unlike *Yes-or-No* questions, the *What* questions examine more comprehensive low-level attribute understanding of MLLMs, by requiring correct perception on **multiple** attributes.

Type 3: *How* Questions. Despite the two common types, we also include a special type, the *How* questions, to cover non-extreme appearances [14] of low-level attribute dimensions into our benchmark, as an extension to *Yes-or-No* questions. As shown in Fig. 3, we can query *How is the clarity of the image?* for the image with both clear and blurry areas, and answer with **Medium**. With this special question type, we broaden the **Q-bench**$^+$ into **finer-grained** low-level perception.

4) GPT-assisted Evaluation Process

The input query format for MLLMs is as follows:

- **Single Images:**
  #User: How is the clarity of the image? (Question)
  [IMAGE_TOKEN] (Image)
  Choose between one of the following options:
  A. High (Correct) *B. Medium* (Wrong) *C. Low* (Wrong)
  #Assistant:
- **Image Pairs:**
  #User: Which image is brighter? (Question)
  The first image: [IMAGE_TOKEN] (Image 1)
  The second image: [IMAGE_TOKEN] (Image 2)
  Choose between one of the following options:
  A. The first image (Wrong) *B. The second image* (Correct)
  #Assistant:

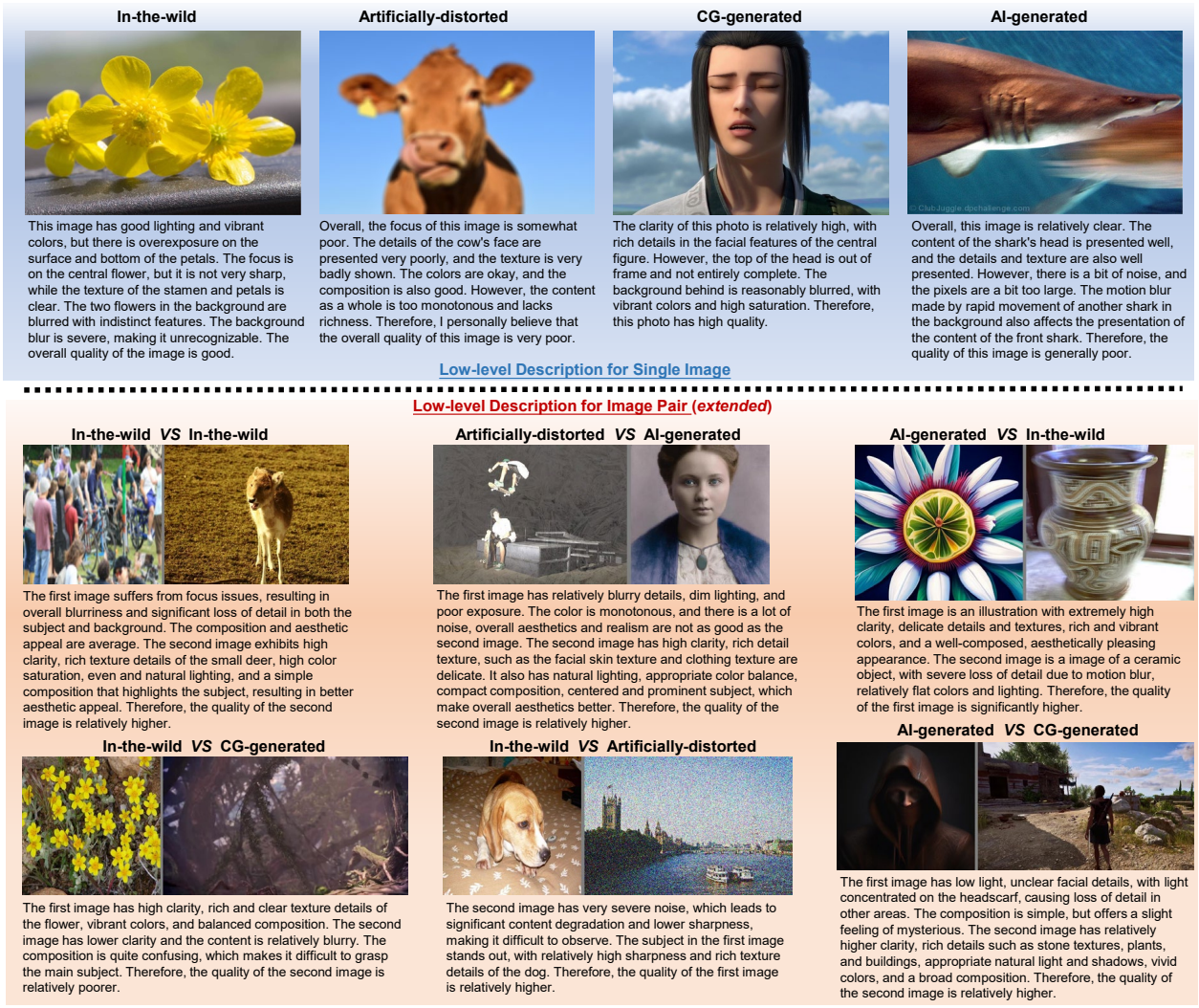

Fig. 4. A dataset card of **LLDescribe**$^+$ that evaluates the low-level **description** ability of MLLMs. 499 single images and 450 image pairs from 10 diverse sources are labeled with *golden* descriptions, to serve as **text** references to evaluate the completeness, preciseness, and relevance of MLLM outputs.

The correct answer has been shuffled and finally uniformly distributed among all choices (A/B/C/D). Moreover, while traditional visual question answering [8], [41] tasks typically employ traditional language metrics (BLEU-4, CIDEr) to compare performance, as observed by recent studies [52] and validated by us, most MLLMs cannot consistently provide outputs on **instructed formats**. Given the question above, different MLLMs may reply *"A."*, *"High"*, *"The clarity of the image is high."*, *"The image is of high clarity."* (all correct), which are difficult to be exhaustively-included under traditional metrics. To solve this problem, we design, validate, and employ a **5-round** GPT-assisted evaluation process inspired by [29]. Under this process, the question, correct answers, and MLLM replies are fed into GPT for evaluation.

C. Benchmark on Low-level *Description* Ability

In the second task of **Q-Bench**$^+$, we evaluate the language **description** ability of MLLMs on low-level information. This task is a sibling task of image captioning [7], [53], [54] that describes image content with natural language, with a specific concern on the low-level appearance of images. To evaluate this ability automatically, we first derive a *golden* low-level description dataset, denoted as **LLDescribe**$^+$ (Sec. II-C1), including one long (*average 58 words*) *golden* description provided by experts for each of 499 images. With these *golden* text descriptions, we are able to measure the quality of output low-level descriptions from MLLMs with a single-modal GPT, under the three dimensions: **completeness**, **preciseness**, as well as **relevance** (Sec II-C2). The discussions of the *golden* descriptions and the evaluation process are as follows.

1) Defining Golden Low-level Descriptions for Images: For the description ability, MLLMs should accurately and completely describe low-level visual information of images. Thus, the *ground truths* for these MLLMs are also built within a basic principle to cover as many low-level concerns as possible, so long as they are enumerated in Sec. II-B and occur in images. The resulting *golden* descriptions in **LLDescribe**$^+$ have an average duration of **58** words, notably longer than common high-level image caption datasets (**11** for [54], **10** for [7]). Similar to the **LLVisionQA**$^+$ dataset for the perception task, the 499 single images and 450 image pairs in **LLDescribe**$^+$ dataset also include all 10 sources (as in

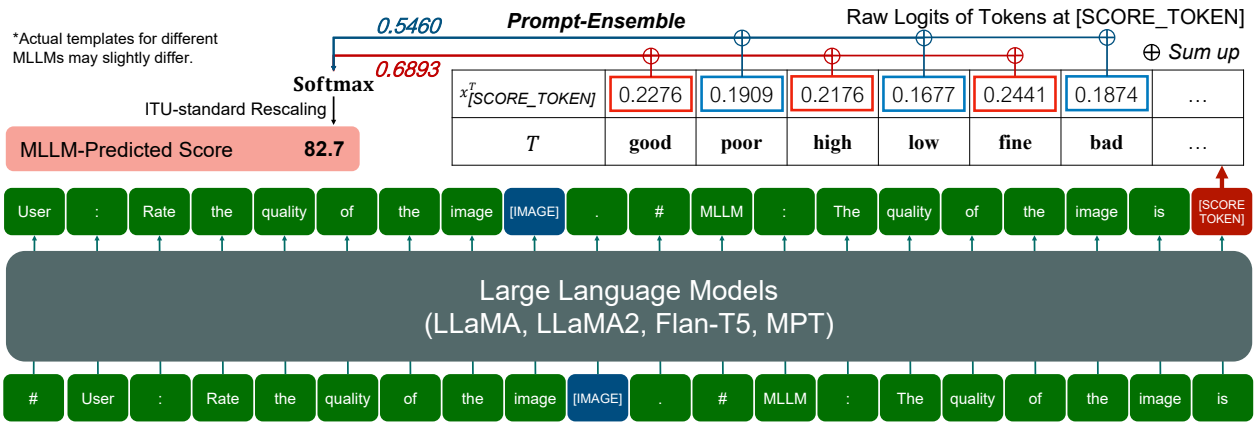

Fig. 5. The proposed softmax-based quality **assessment** strategy for MLLMs. Instead of directly decoding tokens from the *[SCORE_TOKEN] position*, the strategy extracts log probabilities (logits) of ***good*** and ***poor***, and predicts *quantifiable* score via a `softmax` pooling between the two logits.

Table I) to cover images with diverse low-level appearances. The *golden* descriptions on different sources of images are depicted in Fig. 4.

2) Evaluation with Single-modal GPT: After collecting the *golden* descriptions, we design an input prompt to acquire the output descriptions from MLLMs:

- **Single Images:**
  #User: Describe the quality, aesthetics and other low-level appearance of the image in details. `(Prompt)`
  [IMAGE_TOKEN] `(Image)`
  #Assistant:
- **Image Pairs:**
  #User: Compare and jointly analyze the quality, aesthetics and other low-level appearance of the images in details. `(Prompt)`
  The first image: [IMAGE_TOKEN] `(Image 1)`
  The second image: [IMAGE_TOKEN] `(Image 2)`
  #Assistant:

Recent studies [55] have proved single-modal GPT [28] to be a reliable evaluation tool for pure language tasks. Via the **LLDescribe**$^+$ dataset, we convert the multi-modality problem into a text-only setting, by matching the MLLM outputs with the *golden* descriptions with single-modal GPT under three dimensions: **(1) Completeness.** More matched information with the *golden* description is encouraged. **(2) Preciseness.** The controversial information with the *golden* description is punished. **(3) Relevance.** More proportions of MLLM outputs should be related to low-level information, instead of others. Each dimension is scored among [0,1,2]. Similar as Sec. II-B, we repeat **5 rounds** for each single evaluation and collect the weighted average as the final score.

*D. Benchmark on Precise Quality **Assessment** Ability*

In the third task, we benchmark the ability of MLLMs to provide **quantifiable assessment** on the overall low-level appearance of images. Unlike the two tasks above, we utilize existing IQA datasets that are collected across a variety of low-level appearances to evaluate how MLLMs can predict **quantifiable** quality scores aligned with human opinions. All the three types of IQA datasets (*in-the-wild*, *generated*, *artificially-distorted*) as mentioned in Sec. II-A are evaluated, to provide a broad range measurement of the assessment ability of MLLMs. Nevertheless, how to collect **quantifiable** quality scores from MLLMs remains challenging as their outputs only have weak measurability (Sec. II-D1). Noticing that MLLMs can provide probabilities of tokens, we employ softmax pooling on the logits of ***good*** and ***poor*** under a simple and direct prompt template, deriving into **quantifiable** quality scores (Sec. II-D2), as illustrated in Fig. 5. Details as follows.

1) Weak Measurability of MLLM Outputs: In **Q-Bench**$^+$, we aim to fairly compare the **assessment** ability between different MLLMs on diverse low-level appearances. Henceforth, our principle is to define a unified, simplest instruction that is applicable for all MLLMs on all IQA datasets. Under this principle, we conduct toy experiments on Shikra [64] and LLaVA-v1 [3], with two simple instruction strategies: **(A) Direct Instruction,** in which the prompt is designed as simple as *"Rate the quality of the image"*. The top-frequency answers are ***good*** (78%), and ***poor*** (20%), with other outputs almost negligible. **(B) Numerical Instruction,** in which we specifically instruct numerical ratings, with the prompt: *"Score the quality of the image from 1 to 5, with 1 as lowest and 5 as highest."*. Under the numerical strategy, the top-frequency answers are **5** (84%), **1** (9%), and **3** (5%); though within the score range, the frequencies of scores **2** and **4** are both less than 1%. The toy experiments imply the weak measurability of MLLM outputs, given that the answers are statistically **1)** biased towards *positive*, **2)** biased towards *extreme*, and **3)** with *only two* effective scales. Therefore, it is necessary to explore extended strategies for MLLMs to provide truly **quantifiable** outputs for low-level **assessment**.

2) A Softmax-based Evaluation Strategy: Given the above observations, we design the softmax-based evaluation strategy (Fig. 5) to reduce the negative impacts of the biases and lack of scales. To start with, we design our strategy within the **Direct Instruction**, which is more general and less biased than the **Numerical Instruction**. The strategy is based on the observation that two top-frequency outputs, ***good*** and ***poor***, can be considered as anchors for better and worse human perception, and the **Direct Strategy** can be approximated into a binary classification problem on the *[SCORE_TOKEN]* position, or technically, an `argmax` between the logits of ***good*** ($x^{\mathbf{good}}_{SCORE\_TOKEN}$) and ***poor*** ($x^{\mathbf{poor}}_{SCORE\_TOKEN}$) on this position. In our revised strategy, we modify the `argmax` into `softmax`

TABLE II
RESULTS ON THE dev AND test SUBSETS OF **LLVISIONQA**$^+$ FOR THE LOW-LEVEL **PERCEPTION** ABILITY OF MLLMS. OPEN-SOURCE MLLMS WITH *top-3* PERFORMANCE IN EACH SUB-CATEGORY ARE MARKED WITH BEST IN **BOLD** AND SECOND/THIRD UNDERLINED.

Sub-categories	**Question Types**			**Quadrants of Low-level Concerns**				
Model *(variant)*	*Yes-or-No*↑	*What*↑	*How*↑	*Distortion*↑	*Other*↑	*In-context Distortion*↑	*In-context Other*↑	*Overall*↑
***Dev Set** / random guess*	50.00%	27.86%	33.31%	37.89%	38.48%	38.28%	35.82%	37.80%
InfiMM (*Zephyr-7B*) [56]	57.45%	57.96%	44.62%	47.27%	57.17%	49.67%	64.08%	53.37%
Emu2-Chat (*LLaMA-33B*) [25]	71.81%	67.25%	56.18%	**64.78%**	63.19%	63.48%	72.24%	65.28%
Fuyu-8B (*Persimmon-8B*) [57]	53.33%	43.70%	38.00%	40.81%	47.40%	45.45%	49.23%	45.05%
BakLLava (*Mistral-7B*) [26]	66.00%	56.16%	51.12%	51.15%	61.57%	53.72%	72.00%	57.48%
SPHINX [58]	**74.18%**	**68.81%**	**62.07%**	63.62%	**71.76%**	**66.12%**	**76.33%**	**68.56%**
mPLUG-Owl2 *(LLaMA-7B)* [59]	72.18%	57.96%	56.19%	56.68%	69.21%	53.29%	72.65%	61.61%
LLaVA-v1.5 (*Vicuna-v1.5-7B*) [60]	66.36%	58.19%	50.51%	49.42%	65.74%	54.61%	70.61%	58.66%
LLaVA-v1.5 (*Vicuna-v1.5-13B*) [60]	65.27%	64.38%	56.59%	56.03%	67.13%	61.18%	67.35%	62.14%
InternLM-XComposer-VL *(InternLM)* [61]	69.45%	65.27%	60.85%	61.67%	70.14%	56.91%	75.10%	65.35%
IDEFICS-Instruct *(LLaMA-7B)* [62]	56.18%	44.69%	44.02%	42.80%	54.17%	44.74%	56.33%	48.70%
Qwen-VL *(QwenLM)* [63]	63.09%	58.19%	56.39%	50.58%	62.73%	57.89%	73.88%	59.40%
Shikra *(Vicuna-7B)* [64]	65.64%	47.35%	49.09%	48.83%	59.49%	50.00%	64.08%	54.65%
Otter-v1 *(MPT-7B)* [6]	57.09%	40.71%	39.55%	42.22%	49.31%	44.08%	52.65%	46.35%
InstructBLIP *(Flan-T5-XL)* [5]	67.64%	59.96%	55.98%	56.23%	65.51%	58.22%	69.39%	61.47%
InstructBLIP *(Vicuna-7B)* [5]	71.64%	52.65%	43.81%	48.64%	62.50%	55.59%	64.90%	56.72%
VisualGLM-6B *(GLM-6B)* [65]	60.18%	54.20%	46.25%	51.75%	54.40%	53.62%	57.14%	53.78%
mPLUG-Owl *(LLaMA-7B)* [52]	66.00%	54.87%	44.02%	51.36%	55.09%	54.28%	65.71%	55.38%
LLaMA-Adapter-V2 [66]	66.18%	59.29%	52.13%	57.39%	56.25%	63.16%	64.90%	59.46%
LLaVA-v1 (*Vicuna-13B*) [3]	54.00%	53.10%	55.38%	48.64%	54.63%	55.59%	63.27%	54.18%
MiniGPT-4 *(Vicuna-13B)* [4]	55.82%	50.22%	40.37%	42.02%	48.38%	51.97%	61.22%	49.03%
Qwen-VL-Plus (*Close-Source*) [63]	73.77%	69.47%	53.88%	66.21%	65.72%	63.81%	68.75%	66.04%
Qwen-VL-Max (*Close-Source*) [63]	75.60%	79.43%	66.09%	73.39%	74.08%	71.0%	76.92%	73.63%
Gemini-Pro (*Close-Source*) [27]	68.80%	73.74%	62.34%	66.30%	71.34%	63.91%	73.09%	68.16%
GPT-4V (*Close-Source*) [28]	76.85%	79.17%	67.52%	73.53%	76.18%	72.83%	76.47%	74.51%
***Test Set** / random guess*	50.00%	28.48%	33.30%	37.24%	38.50%	39.13%	37.10%	37.94%
InfiMM (*Zephyr-7B*) [56]	61.31%	56.61%	49.58%	47.79%	62.05%	51.71%	67.68%	56.05%
Emu2-Chat (*LLaMA-33B*) [25]	70.09%	65.12%	54.11%	**66.22%**	62.96%	63.47%	73.21%	64.32%
Fuyu-8B (*Persimmon-8B*) [57]	62.22%	35.79%	36.62%	41.07%	49.40%	45.89%	49.04%	45.75%
BakLLava (*Mistral-7B*) [26]	66.46%	61.48%	54.83%	51.33%	63.76%	56.52%	**78.16%**	61.02%
SPHINX [58]	**74.45%**	**65.50%**	**62.13%**	59.11%	**73.26%**	**66.09%**	77.56%	**67.69%**
mPLUG-Owl2 *(LLaMA-7B)* [59]	72.26%	55.53%	58.64%	52.59%	71.36%	58.90%	73.00%	62.68%
LLaVA-v1.5 (*Vicuna-v1.5-7B*) [60]	64.60%	59.22%	55.76%	47.98%	67.30%	58.90%	73.76%	60.07%
LLaVA-v1.5 (*Vicuna-v1.5-13B*) [60]	64.96%	64.86%	54.12%	53.55%	66.59%	58.90%	71.48%	61.40%
InternLM-XComposer-VL *(InternLM)* [61]	68.43%	62.04%	61.93%	56.81%	70.41%	57.53%	77.19%	64.35%
IDEFICS-Instruct *(LLaMA-7B)* [62]	60.04%	46.42%	46.71%	40.38%	59.90%	47.26%	64.77%	51.51%
Qwen-VL *(QwenLM)* [63]	65.33%	60.74%	58.44%	54.13%	66.35%	58.22%	73.00%	61.67%
Shikra (*Vicuna-7B*) [64]	69.09%	47.93%	46.71%	47.31%	60.86%	53.08%	64.77%	55.32%
Otter-v1 *(MPT-7B)* [6]	57.66%	39.70%	42.59%	42.12%	48.93%	47.60%	54.17%	47.22%
InstructBLIP *(Flan-T5-XL)* [5]	69.53%	59.00%	56.17%	57.31%	65.63%	56.51%	71.21%	61.94%
InstructBLIP *(Vicuna-7B)* [5]	70.99%	51.41%	43.00%	45.00%	63.01%	57.19%	64.39%	55.85%
VisualGLM-6B *(GLM-6B)* [65]	61.31%	53.58%	44.03%	48.56%	54.89%	55.48%	57.79%	53.31%
mPLUG-Owl *(LLaMA-7B)* [52]	72.45%	54.88%	47.53%	49.62%	63.01%	62.67%	66.67%	58.93%
LLaMA-Adapter-V2 [66]	66.61%	54.66%	51.65%	56.15%	61.81%	59.25%	54.55%	58.06%
LLaVA-v1 (*Vicuna-13B*) [3]	57.12%	54.88%	51.85%	45.58%	58.00%	57.19%	64.77%	54.72%
MiniGPT-4 *(Vicuna-13B)* [4]	60.77%	50.33%	43.00%	45.58%	52.51%	53.42%	60.98%	51.77%
Qwen-VL-Plus (*Close-Source*) [63]	75.74%	73.25%	57.33%	64.88%	73.24%	68.67%	70.56%	68.93%
Qwen-VL-Max (*Close-Source*) [63]	73.20%	81.02%	68.39%	70.84%	74.57%	73.11%	80.44%	73.90%
Gemini-Pro (*Close-Source*) [27]	71.26%	71.39%	65.59%	67.30%	73.04%	65.88%	73.60%	69.46%
GPT-4V (*Close-Source*) [28]	77.72%	78.39%	66.45%	71.01%	71.07%	79.36%	78.91%	74.10%
Junior-level Human	82.48%	79.39%	60.29%	75.62%	72.08%	76.37%	73.00%	74.31%
Senior-level Human	84.31%	88.94%	72.02%	79.65%	79.47%	83.90%	87.07%	81.74%

to collect better **quantifiable** scores:

$$q_{\text{pred}} = \frac{e^{x^{\textbf{good}}_{SCORE\_TOKEN}}}{e^{x^{\textbf{good}}_{SCORE\_TOKEN}} + e^{x^{\textbf{poor}}_{SCORE\_TOKEN}}} \tag{1}$$

This simple and generally-applicable strategy enables us to collect **quantifiable** outputs (q_{pred}) from MLLMs with higher correlation to human ratings, as verified in our experimental analysis (Table VII).

3) Prompt Ensemble for Boosting Quantitative Abilities of MLLMs: Multiple synonym prompts can broaden the semantic range, allowing for a more nuanced understanding that might be missed by a single term. Additionally, multiple synonym prompts diminish uncertainty since diverse terms have subtly different meanings, resulting in a more dependable assessment. Specifically, we further choose the combination prompts of [***good***, ***high***, ***fine***] and [***poor***, ***low***, ***bad***] to replace ***good*** and ***poor*** respectively. The *quantifiable* outputs (q_{pred}) can then be altered as:

$$q_{\text{pred}} = \frac{e^{\sum_t^{t\in\mathcal{P}} x^t_{SCORE\_TOKEN}}}{e^{\sum_t^{t\in\mathcal{P}} x^t_{SCORE\_TOKEN}} + e^{\sum_t^{t\in\mathcal{N}} x^t_{SCORE\_TOKEN}}} \tag{2}$$

where $\mathcal{P}$ indicates the positive token set (from *good*, *fine*, *high*, etc.), while $\mathcal{N}$ represents the negative token set (from *poor*, *bad*, *low*, etc.). **The implementation of the prompt ensemble approach does not add extra computational complexity.** The core computation occurs once the input prompt is entered and the language model generates the *[SCORE_TOKEN]*. After this, we only require tokenization of the words used, followed by the calculation of logits for the *[SCORE_TOKEN]*. The boosted performance is listed in Table VIII.

TABLE III
RESULTS ON THE dev AND test SUBSETS OF **LLVISIONQA$^+$** FOR THE LOW-LEVEL **PERCEPTION-PAIR** ABILITY OF MLLMS. TOPEN-SOURCE MLLMS WITH *top-3* PERFORMANCE IN EACH SUB-CATEGORY ARE MARKED WITH BEST IN **BOLD** AND SECOND/THIRD UNDERLINED.

Sub-categories	**Question Types**			**Low-level Concerns**		**Pairwise Concerns**		
Model *(variant)*	*Yes-or-No*↑	*What*↑	*How*↑	*Distortion*↑	*Other*↑	*Compare*↑	*Joint*↑	*Overall*↑
***Dev Set** / random guess*	50.00%	32.16%	33.30%	38.59%	41.74%	38.66%	43.89%	39.60%
InfiMM (*Zephyr-7B*) [56]	48.11%	39.04%	40.06%	42.56%	43.78%	41.77%	48.33%	42.95%
Emu2-Chat (*LLaMA-33B*) [25]	56.64%	41.15%	49.62%	49.12%	51.91%	47.86%	60.00%	50.05%
Fuyu-8B (*Persimmon-8B*) [57]	**68.76%**	33.56%	38.78%	46.83%	54.03%	47.86%	55.00%	49.15%
BakLLava (*Mistral-7B*) [26]	56.92%	**43.83%**	**50.00%**	49.33%	**54.34%**	**50.66%**	52.22%	**50.94%**
mPLUG-Owl2 (*Q-Instruct*) [67]	59.19%	42.12%	47.43%	**49.63%**	52.48%	49.81%	53.88%	50.54%
mPLUG-Owl2 (*LLaMA-7B*) [59]	58.43%	39.72%	48.39%	49.04%	51.55%	47.50%	**60.55%**	49.85%
LLaVA-v1.5 (*Vicuna-v1.5-7B*) [60]	60.46%	42.85%	41.53%	47.88%	51.89%	46.55%	59.57%	49.32%
LLaVA-v1.5 (*Vicuna-v1.5-13B*) [60]	56.42%	42.46%	48.38%	48.15%	53.41%	48.84%	54.44%	49.85%
Qwen-VL-Plus (*Close-Source*) [63]	63.63%	55.55%	55.71%	61.61%	56.52%	65.81%	58.45%	60.70%
Qwen-VL-Max (*Close-Source*) [63]	71.96%	62.87%	65.53%	69.21%	62.69%	67.54%	66.01%	67.27%
Gemini-Pro (*Close-Source*) [27]	64.98%	51.36%	54.16%	58.17%	56.52%	57.73%	57.22%	57.64%
GPT-4V (*Close-Source*) [28]	79.34%	70.54%	78.52%	75.84%	77.95%	78.80%	66.11%	76.52%
***Test Set** / random guess*	50.00%	32.03%	33.16%	38.95%	41.95%	38.69%	43.70%	39.82%
InfiMM (*Zephyr-7B*) [56]	54.21%	43.38%	45.32%	49.57%	45.67%	48.32%	48.88%	48.44%
Emu2-Chat (*LLaMA-33B*) [25]	51.94%	29.78%	**53.84%**	42.01%	55.71%	46.26%	49.09%	47.08%
Fuyu-8B (*Persimmon-8B*) [57]	**70.36%**	28.13%	35.98%	44.08%	57.43%	47.02%	51.11%	47.94%
BakLLava (*Mistral-7B*) [26]	60.09%	45.42%	50.86%	**53.09%**	58.82%	**54.52%**	55.55%	52.75%
mPLUG-Owl2 (*Q-Instruct*) [67]	60.24%	**47.46%**	48.78%	52.81%	53.97%	51.42%	59.11%	**53.15%**
mPLUG-Owl2 (*LLaMA-7B*) [59]	58.07%	36.61%	48.44%	47.74%	51.90%	45.73%	**60.00%**	48.94%
LLaVA-v1.5 (*Vicuna-v1.5-7B*) [60]	60.72%	42.37%	50.17%	49.15%	**59.86%**	52.97%	49.77%	52.25%
LLaVA-v1.5 (*Vicuna-v1.5-13B*) [60]	57.34%	47.45%	49.13%	49.01%	59.51%	52.06%	52.00%	52.05%
Qwen-VL-Plus (*Close-Source*) [63]	66.85%	55.79%	59.91%	62.46%	58.77%	62.17%	59.20%	61.48%
Qwen-VL-Max (*Close-Source*) [63]	67.65%	67.56%	65.35%	69.09%	61.18%	68.65%	61.29%	66.99%
Gemini-Pro (*Close-Source*) [27]	65.78%	56.61%	56.74%	60.42%	60.55%	60.46%	60.44%	60.46%
GPT-4V (*Close-Source*) [28]	79.75%	69.49%	84.42%	77.32%	79.93%	81.00%	68.00%	78.07%
Junior-level Human	78.11%	77.04%	82.33%	78.17%	77.22%	80.26%	76.39%	80.12%
Senior-level Human	83.00%	84.81%	89.85%	83.13%	90.78%	86.55%	82.25%	85.48%

III. EXPERIMENT

In **Q-Bench$^+$**, we evaluate the performance of up to **20** up-to-date popular and competitive open-source as well as **4** close-source commercial MLLMs under **zero-shot** settings.

A. Findings on *Perception*

For a holistic examination of the **perception** ability of MLLMs, we evaluate the multi-choice correctness of MLLMs on different sub-categories of the **LLVision$^+$** dataset, which is equally divided as `dev` (*will be released*) and `test` (*will keep private*) subsets as shown in Table II and Table III respectively. **Only the MLLMs that support multiple images input** are included for the **perception-pair** ability benchmark.

1) Perception for Single Images: a) We are glad that the majority of MLLMs can significantly outperform *random guess* on all sub-categories as shown in Table II. Considering that all participating MLLMs are without any explicit training on low-level visual attributes, these results show strong potentials for these general-purpose models when further fine-tuned with respective low-level datasets. b) Among all open-source MLLMs, the recently-released SPHINX reaches the best accuracy on this question-answering task, followed by Emu2-Chat and InternLM-XComposer-VL, which show rather close results. By achieving **more than 64%** accuracy on both subsets, these models show exciting potential as robust low-level visual assistants in the future. c) Another key observation is that almost all methods **perceive worse on distortions** than other low-level attributes, which indicates that distortion questions are relatively more challenging. d) **Close-source MLLMs and Humans.** It is widely acknowledged that commercial close-source MLLMs are the leading models in various tasks. To evaluate the low-level **perception** abilities of these MLLMs, we gauge the accuracy of Qwen-VL-Plus (Alibaba), Qwen-VL-Max (Alibaba), Gemini-Pro (Google), and GPT-4V (OpenAI) on the subsets of **LLVision$^+$** dataset. All close-source MLLMs achieve superior performance than all open-source MLLMs, which indicates that open-source MLLMs still fall behind on low-level visual ability. GPT-4V exhibits the most competitive performance and outperforms the best open-source MLLM (SPHINX) by a large margin (**+6%**), and on par accuracy with the *Junior-level Human*. Despite its prowess, there is still a way to go for GPT-4V before it can match the overall proficiency of the *Senior-level Human (with experiences on low-level visual tasks*, **7%** better than GPT-4V). Furthermore, across all categories, the results show that GPT-4V, much like its open-source counterparts, faces challenges in recognizing **distortions**.

2) Perception for Image Pairs: Perception for image pairs is far more difficult for MLLMs since this task not only requires MLLMs to have stable low-level visual capabilities, but also requires MLLMs to be able to analyze two images simultaneously and conduct discerning comparisons. To enrich the MLLM diversity, we further include the mPLUG-Owl2 fine-tuned with the single image low-level visual dataset **Q-Instruct** [67] for comparison. The performance is exhibited in Table III. With closer inspections, we can obtain several interesting findings. a) **Open-source MLLMs are poor low-level comparators.** It seems that although they might show strong performance for single image perception, they are quite confused by the image pairs. Most of them get worse performance on the **Compare** subset than the **Joint** subset, which further confirms this point. For mPLUG-Owl2 (*Q-Instruct*), despite being fine-tuned with the single image low-

TABLE IV
RESULTS ON THE LOW-LEVEL **DESCRIPTION** ABILITY OF MLLMS. P_i DENOTES FREQUENCY FOR SCORE i.

Dimensions	**Completeness**				**Precision**				**Relevance**				*Sum.*↑
Model (*variant*)	P_0	P_1	P_2	*score*↑	P_0	P_1	P_2	*score*↑	P_0	P_1	P_2	*score*↑	
InfiMM (*Zephyr-7B*) [56]	29.61%	62.32%	7.77%	0.77	29.25%	31.90%	38.51%	1.08	2.16%	22.72%	74.58%	1.71	3.58
Emu2-Chat (*LLaMA-33B*) [25]	20.01%	52.77%	27.22%	1.07	24.66%	27.12%	48.22%	1.24	1.21%	9.91%	88.88%	**1.88**	4.19
Fuyu-8B (*Persimmon-8B*) [57]	25.54%	61.00%	13.46%	0.88	41.96%	32.76%	25.28%	0.83	2.99%	11.34%	85.67%	1.82	3.53
BakLLava (*Mistral-7B*) [26]	24.31%	51.22%	24.47%	1.00	49.23%	24.11%	26.66%	0.77	1.25%	36.22%	62.53%	1.61	3.38
SPHINX [58]	27.96%	64.36%	7.33%	0.79	26.16%	32.42%	41.01%	1.14	1.69%	23.00%	74.61%	1.72	3.65
mPLUG-Owl2 (*LLaMA-7B*) [59]	27.71%	38.58%	33.71%	1.06	28.11%	19.78%	52.11%	1.24	7.91%	48.18%	43.91%	1.36	3.67
LLaVA-v1.5 (*Vicuna-v1.5-7B*) [60]	27.48%	54.74%	17.78%	0.90	30.51%	26.04%	43.45%	1.13	10.85%	60.34%	28.81%	1.18	3.21
LLaVA-v1.5 (*Vicuna-v1.5-13B*) [60]	27.68%	53.78%	18.55%	0.91	25.45%	21.47%	53.08%	**1.28**	6.31%	58.75%	34.94%	1.29	3.47
InternLM-XComposer-VL *(InternLM)* [61]	19.94%	51.82%	28.24%	1.08	22.59%	28.99%	48.42%	1.26	1.05%	10.62%	88.32%	**1.87**	**4.21**
IDEFICS-Instruct *(LLaMA-7B)* [62]	28.91%	59.16%	11.93%	0.83	34.68%	27.86%	37.46%	1.03	3.90%	59.66%	36.44%	1.33	3.18
Qwen-VL *(QwenLM)* [63]	26.34%	49.13%	24.53%	0.98	50.62%	23.44%	25.94%	0.75	0.73%	35.56%	63.72%	1.63	3.36
Shikra (*Vicuna-7B*) [64]	21.14%	68.33%	10.52%	0.89	30.33%	28.30%	41.37%	1.11	1.14%	64.36%	34.50%	1.33	3.34
Otter-v1 *(MPT-7B)* [6]	22.38%	59.36%	18.25%	0.96	40.68%	35.99%	23.33%	0.83	1.95%	13.20%	84.85%	1.83	3.61
Kosmos-2 [9]	8.76%	70.91%	20.33%	**1.12**	29.45%	34.75%	35.81%	1.06	0.16%	14.77%	85.06%	1.85	4.03
InstructBLIP *(Flan-T5-XL)* [5]	23.16%	66.44%	10.40%	0.87	34.85%	26.03%	39.12%	1.04	14.71%	59.87%	25.42%	1.11	3.02
InstructBLIP *(Vicuna-7B)* [5]	29.73%	61.47%	8.80%	0.79	27.84%	23.52%	48.65%	1.21	27.40%	61.29%	11.31%	0.84	2.84
VisualGLM-6B *(GLM-6B)* [65]	30.75%	56.64%	12.61%	0.82	38.64%	26.18%	35.18%	0.97	6.14%	67.15%	26.71%	1.21	2.99
mPLUG-Owl *(LLaMA-7B)* [52]	28.28%	37.69%	34.03%	1.06	26.75%	18.18%	55.07%	**1.28**	3.03%	33.82%	63.15%	1.60	3.94
LLaMA-Adapter-V2 [66]	30.44%	53.99%	15.57%	0.85	29.41%	25.79%	44.80%	1.15	1.50%	52.75%	45.75%	1.44	3.45
LLaVA-v1 (*Vicuna-13B*) [3]	34.10%	40.52%	25.39%	0.91	30.02%	15.15%	54.83%	1.25	1.06%	38.03%	60.91%	1.60	3.76
MiniGPT-4 (*Vicuna-13B*) [4]	34.01%	32.15%	33.85%	1.00	29.20%	15.27%	55.53%	1.26	6.88%	45.65%	47.48%	1.41	3.67

TABLE V
RESULTS ON THE LOW-LEVEL **DESCRIPTION-PAIR** ABILITY OF MLLMS. P_i DENOTES FREQUENCY FOR SCORE i.

Dimensions	**Completeness**				**Precision**				**Relevance**				*Sum.*↑
Model (*variant*)	P_0	P_1	P_2	*score*↑	P_0	P_1	P_2	*score*↑	P_0	P_1	P_2	*score*↑	
InfiMM (*Zephyr-7B*) [56]	30.75%	62.66%	6.22%	0.75	34.17%	38.84%	26.35%	0.91	2.57%	30.84%	65.28%	1.61	3.28
Emu2-Chat (*LLaMA-33B*) [25]	41.25%	54.33%	4.42%	0.63	38.11%	36.41%	25.48%	0.87	4.12%	38.61%	57.27%	1.53	3.03
Fuyu-8B (*Persimmon-8B*) [57]	37.95%	52.17%	9.11%	0.70	37.68%	37.33%	23.73%	0.84	3.95%	31.15%	62.84%	1.56	3.12
BakLLava (*Mistral-7B*) [26]	29.46%	59.77%	10.57%	0.80	40.0%	38.08%	21.33%	0.80	2.26%	15.06%	82.04%	1.79	3.40
mPLUG-Owl2 (*Q-Instruct*) [59]	15.25%	65.76%	18.32%	**1.02**	39.44%	40.18%	19.62%	0.79	0.09%	9.86%	89.02%	**1.87**	**3.69**
mPLUG-Owl2 *(LLaMA-7B)* [59]	19.43%	65.54%	14.45%	0.94	30.94%	43.71%	24.63%	**0.92**	3.79%	26.94%	68.28%	1.63	3.50
LLaVA-v1.5 (*Vicuna-v1.5-7B*) [60]	19.68%	72.57%	7.19%	0.86	38.00%	40.04%	20.97%	0.82	2.13%	39.77%	56.66%	1.53	3.22
LLaVA-v1.5 (*Vicuna-v1.5-13B*) [60]	18.77%	73.44%	7.79%	0.89	34.66%	38.72%	26.62%	**0.92**	1.02%	34.59%	64.39%	1.63	3.44

level visual dataset **Q-Instruct** [67], the overall performance improvement from the low-level knowledge infusion of single images is relatively weak. This also suggests that there is a necessity to build open-source low-level datasets for multiple images to cultivate the corresponding capabilities of open-source MLLMs. b) **Close-source MLLMs are more robust in this task.** This may be because these close-source MLLMs are supported by training on multiple-image data, allowing them to make better comparative judgments. Particularly with GPT-4V, its performance in the **compare** subset is significantly higher than in the **joint** subset, and it far exceeds all other models, even reaching the level of a junior human. c) **Perception for image pairs is easier for humans.** Comparing image pairs is simpler for humans, as the answers to related questions tend to be more objective. Especially for junior-level humans with no professional experience, they may have stronger subjectivity in grasping absolute sensations, but it is easier to remain objective when they are faced with comparison concepts. For example, it's difficult for a junior-level human to judge whether the lighting in a dimly lit single image is appropriate. However, if presented with another image with even weaker lighting, they can easily determine which image is worse. This may explain the notable **6%** improvements for junior-level human from single images to image pairs.

In conclusion, the performance of open-source MLLMs on low-level **perception** for image pairs is still far from satisfactory, which needs to be enhanced and optimized.

*B. Findings on **Description***

1) Description for single images: For the **description** ability exhibited in Table IV, InternLM-XComposer-VL reaches the best proficiency, especially in terms of the relevance dimension. Nevertheless, in the perspective of the completeness and precision of the descriptions, even the best of all MLLMs cannot obtain an excellent score; on the contrary, almost all MLLMs reach an acceptable standard (0.8/2.0). In general, all MLLMs at present are only with relatively limited and primary ability to provide low-level visual descriptions.

2) Description for image pairs: We also include the mPLUG-Owl2 (*Q-Instruct*) for **description-pair** ability benchmark. As shown in Table V, similarly, all MLLMs perform better in the aspect of relevance than completeness and precision. Furthermore, fine-tuned with the single image low-level visual dataset **Q-Instruct** [67], mPLUG-Owl2 (*Q-Instruct*) achieves the best performance on completeness and relevance but gets the lowest score on precision. This indicates the knowledge infusion from single images can effectively enhance an MLLM to focus on corresponding low-level dimensions for targeted responses, but it does not improve the accuracy of the content, meaning it cannot enhance the core analytical ability of image pairs.

*C. Findings on **Assessment***

1) MLLM Performance: To measure the **assessment** ability, we evaluate the performance of 20 open-source MLLMs on

TABLE VI
MAIN EVALUATION RESULTS ON THE ZERO-SHOT **ASSESSMENT** ABILITY OF MLLMS, IN COMPARISON WITH NIQE AND CLIP-VIT-LARGE-14, THE VISUAL BACKBONE OF MOST MLLMS. METRICS ARE *SRCC/PLCC*.

Dataset Type	In-the-wild				Generated		Artificial	*Average*
Model / Dataset	*KONiQ-10k*	*SPAQ*	*LIVE-FB*	*LIVE-itw*	*CGIQA-6K*	*AGIQA-3K*	*KADID-10K*	
NIQE [68]	0.316/0.377	0.693/0.669	0.211/0.288	0.480/0.451	0.075/0.056	0.562/0.517	0.374/0.428	0.387/0.398
CLIP-ViT-Large-14 [69]	0.468/0.505	0.385/0.389	0.218/0.237	0.307/0.308	0.285/0.290	0.436/0.458	0.376/0.388	0.354/0.368
InfiMM (*Zephyr-7B*) [56]	0.507/0.547	0.616/0.633	0.269/0.299	0.548/0.580	0.229/0.245	0.706/0.767	0.466/0.452	0.477/0.503
Emu2-Chat (*LLaMA-33B*) [25]	**0.664/0.714**	0.712/0.698	0.355/0.341	0.597/0.611	0.224/0.269	**0.759**/0.751	**0.841/0.790**	**0.593/0.596**
Fuyu-8B (*Persimmon-8B*) [57]	0.124/0.123	0.125/0.179	0.164/0.133	0.225/0.176	0.118/0.116	0.368/0.317	0.099/0.088	0.174/0.161
BakLLava (*Mistral-7B*) [26]	0.389/0.390	0.406/0.398	0.227/0.216	0.335/0.337	0.179/0.209	0.542/0.561	0.344/0.361	0.346/0.353
mPLUG-Owl2 *(LLaMA-7B)* [59]	0.196/0.252	0.589/0.614	0.217/0.286	0.293/0.342	-0.024/-0.032	0.473/0.492	0.541/0.546	0.326/0.357
LLaVA-v1.5 (*Vicuna-v1.5-7B)* [60]	0.463/0.459	0.443/0.467	0.305/0.321	0.344/0.358	**0.321/0.333**	0.672/0.738	0.417/0.440	0.424/0.445
LLaVA-v1.5 (*Vicuna-v1.5-13B)* [60]	0.448/0.460	0.563/0.584	0.310/0.339	0.445/0.481	0.285/0.297	0.664/0.754	0.390/0.400	0.444/0.474
InternLM-XComposer-VL *(InternLM)* [61]	0.564/0.615	**0.730/0.750**	**0.360/0.416**	**0.612/0.676**	0.243/0.265	0.732/**0.775**	0.546/0.572	0.541/0.581
IDEFICS-Instruct *(LLaMA-7B)* [62]	0.375/0.400	0.474/0.484	0.235/0.240	0.409/0.428	0.244/0.227	0.562/0.622	0.370/0.373	0.381/0.396
Qwen-VL *(QwenLM)* [63]	0.470/0.546	0.676/0.669	0.298/0.338	0.504/0.532	0.273/0.284	0.617/0.686	0.486/0.486	0.475/0.506
Shikra (*Vicuna-7B)* [64]	0.314/0.307	0.320/0.337	0.237/0.241	0.322/0.336	0.198/0.201	0.640/0.661	0.324/0.332	0.336/0.345
Otter-v1 *(MPT-7B)* [6]	0.406/0.406	0.436/0.441	0.143/0.142	-0.008/0.018	0.254/0.264	0.475/0.481	0.557/0.577	0.323/0.333
Kosmos-2 [9]	0.255/0.281	0.644/0.641	0.196/0.195	0.358/0.368	0.210/0.225	0.489/0.491	0.359/0.365	0.359/0.367
InstructBLIP *(Flan-T5-XL)* [5]	0.334/0.362	0.582/0.599	0.248/0.267	0.113/0.113	0.167/0.188	0.378/0.400	0.211/0.179	0.290/0.301
InstructBLIP *(Vicuna-7B)* [5]	0.359/0.437	0.683/0.689	0.200/0.283	0.253/0.367	0.263/0.304	0.629/0.663	0.337/0.382	0.389/0.446
VisualGLM-6B *(GLM-6B)* [65]	0.247/0.234	0.498/0.507	0.146/0.154	0.110/0.116	0.209/0.183	0.342/0.349	0.127/0.131	0.240/0.239
mPLUG-Owl *(LLaMA-7B)* [52]	0.409/0.427	0.634/0.644	0.241/0.271	0.437/0.487	0.148/0.180	0.687/0.711	0.466/0.486	0.432/0.458
LLaMA-Adapter-V2 [66]	0.354/0.363	0.464/0.506	0.275/0.329	0.298/0.360	0.257/0.271	0.604/0.666	0.412/0.425	0.381/0.417
LLaVA-v1 *(Vicuna-13B)* [3]	0.462/0.457	0.442/0.462	0.264/0.280	0.404/0.417	0.208/0.237	0.626/0.684	0.349/0.372	0.394/0.416
MiniGPT-4 (*Vicuna-13B)* [4]	0.239/0.257	0.238/0.253	0.170/0.183	0.339/0.340	0.252/0.246	0.572/0.591	0.239/0.233	0.293/0.300

TABLE VII
EFFECTIVENESS OF THE PROPOSED `softmax` PROBABILITY-BASED STRATEGY AGAINST THE BASELINE `argmax` STRATEGY, ON MULTIPLE MLLMS AND DIFFERENT IQA DATASETS. METRICS ARE *SRCC/PLCC*.

Dataset Type		In-the-wild				Generated		Artificial
Model / Dataset	Strategy	*KONiQ-10k*	*SPAQ*	*LIVE-FB*	*LIVE-itw*	*CGIQA-6K*	*AGIQA-3K*	*KADID-10K*
Shikra (*Vicuna-7B*) [64]	`argmax`	0.178/0.201	0.277/0.281	0.152/0.169	0.248/0.267	0.071/0.065	0.513/0.562	0.245/0.246
Shikra (*Vicuna-7B*) [64]	`softmax`	**0.314/0.307**	**0.327/0.337**	**0.237/0.241**	**0.322/0.336**	**0.198/0.201**	**0.640/0.661**	**0.324/0.332**
InstructBLIP *(Vicuna-7B)* [5]	`argmax`	0.284/0.352	0.662/0.664	0.156/0.249	0.195/0.264	0.141/0.142	0.505/0.567	0.305/0.307
InstructBLIP *(Vicuna-7B)* [5]	`softmax`	**0.359/0.437**	**0.683/0.689**	**0.200/0.283**	**0.253/0.367**	**0.263/0.304**	**0.629/0.663**	**0.337/0.382**
mPLUG-Owl *(LLaMA-7B)* [52]	`argmax`	0.111/0.154	0.463/0.469	0.081/0.123	0.169/0.237	0.082/0.067	0.410/0.466	0.203/0.204
mPLUG-Owl *(LLaMA-7B)* [52]	`softmax`	**0.409/0.427**	**0.634/0.644**	**0.241/0.271**	**0.437/0.487**	**0.148/0.180**	**0.687/0.711**	**0.466/0.486**
LLaMA-Adapter-V2 [66]	`argmax`	0.218/0.237	0.417/0.423	0.222/0.257	0.205/0.239	0.152/0.116	0.545/0.579	0.228/0.229
LLaMA-Adapter-V2 [66]	`softmax`	**0.354/0.363**	**0.464/0.506**	**0.275/0.329**	**0.298/0.360**	**0.251/0.257**	**0.604/0.666**	**0.412/0.425**
LLaVA-v1 *(Vicuna-13B)* [3]	`argmax`	0.038/0.045	0.101/0.108	0.036/0.035	0.059/0.075	0.112/0.109	0.240/0.297	0.005/0.005
LLaVA-v1 *(Vicuna-13B)* [3]	`softmax`	**0.462/0.457**	**0.442/0.462**	**0.264/0.280**	**0.404/0.417**	**0.285/0.297**	**0.626/0.684**	**0.349/0.372**

7 IQA datasets that are with at least **1,000** images and **15** human ratings per image [70]. The experimental results are illustrated in Table VI. a) Primarily, we notice that the majority of MLLMs are notably better than NIQE on **non-natural** circumstances (CGI, AIGC, artificial distortions), showing their potential towards general-purpose evaluators on a broader range of low-level appearances. b) We also notice that without explicit alignment with human opinions during training, the most excellent MLLM, Emu2-Chat (*which is based on the heaviest LLM, LLaMA-33B*), can already outperform CLIP-ViT-Large-14 by a large margin (**25%**). These results have demonstrated that, though most MLLMs are still based on CLIP as visual encoders, their high capacity in the strong language decoder can do help them perform much better on visual quality assessment even without any explicit training.

2) Superiority of `softmax`*:* In this section, we quantitatively evaluate the correlation with human perception on a simple `argmax` strategy between *good↔bad* and our proposed softmax strategy. In Table VII, we select 5 MLLMs of different architectures and confirm that for all IQA datasets, the more measurable `softmax` strategy predicts better than the `argmax` strategy, which degenerates into only two scores, 0 and 1. Though the result is generally expected, the experiments validate that MLLMs have quantitative **assessment** ability hidden behind their word outputs, and prove the effectiveness of our softmax-based IQA strategy.

3) Prompt Ensemble Effectiveness: As shown in Table VIII, the *prompt ensemble* strategy (as proposed in Eq. 2) on top-7 MLLMs (*i.e.* Emu2-Chat, InternLM-XComposer-VL, QWen-VL, InfiMM, LLaVA-v1.5 (*13B*), mPLUG-Owl, and LLaVA-v1.5 (*7B*)) can lead to up to 5% accuracy improvement (*in average 1.7%*). We believe it is a useful boost technique to improve the performance of MLLMs on the IQA task. Nevertheless, we also notice that different MLLMs perform best with different specific prompt combos. For example, the *good+fine↔poor+bad* performs best on InternLM-XComposer-VL, but comes with reduced accuracy on QWen-VL compared with only *good↔poor*. While *good↔poor* is proved *overall best single word pair* (except Emu2-Chat and LLaVA-v1.5 (*13B*)) for the evaluation and shows stable results across MLLMs, we decide to keep the current strategy (using the *good↔poor* combo) in **Q-Bench**$^+$.

TABLE VIII
EVALUATION RESULTS ON THE *synonym ensemble* STRATEGY FOR THE (A3) **ASSESSMENT** ABILITY ON MLLMS WITH TOP-7 RESULTS IN THE DEFAULT A3 LEADERBOARD OF THE **Q-BENCH**$^+$. AFTER *ensemble*, THE RANKINGS AMONG THEM ARE NOT CHANGED. METRICS ARE *SRCC/PLCC*.

Dataset Type	In-the-wild				Generated		Artificial	*Average*
Prompt / Dataset	*KONiQ-10k*	*SPAQ*	*LIVE-FB*	*LIVE-itw*	*CGIQA-6K*	*AGIQA-3K*	*KADID-10K*	
InfiMM (*Zephyr-7B*) [56]								
good↔poor	0.507/0.546	0.616/0.633	0.268/0.299	0.548/0.580	0.229/0.245	0.706/0.767	0.466/0.452	0.477/0.503
fine↔bad	0.331/0.368	0.500/0.527	0.190/0.251	0.305/0.366	0.309/0.324	0.555/0.651	0.411/0.430	0.372/0.417
high↔low	0.412/0.382	0.539/0.492	0.216/0.194	0.586/0.524	0.173/0.171	0.674/0.698	0.429/0.429	0.433/0.413
good+high↔poor+low	0.475/0.492	0.589/0.583	0.249/0.253	0.582/0.578	0.198/0.201	0.697/0.750	0.454/0.456	0.463/0.473
*good+fine↔*poor+bad	0.463/0.502	0.591/0.613	0.255/0.299	0.488/0.530	0.272/0.287	0.675/0.749	0.479/0.467	0.460/0.493
good+high+fine↔poor+low+bad	0.496/0.533	0.605/0.618	0.266/0.291	0.569/0.593	0.239/0.248	0.708/0.768	0.492/0.490	**0.482/0.506**
Emu2-Chat (*LLaMA-33B*) [25]								
good↔poor	0.664/0.714	0.712/0.698	0.355/0.341	0.597/0.611	0.224/0.269	0.759/0.751	0.841/0.790	0.593/0.596
fine↔bad	0.663/0.540	0.711/0.702	0.359/0.362	0.601/0.631	0.285/0.334	0.770/0.599	0.846/0.830	0.605/0.571
high↔low	0.685/0.644	0.721/0.703	0.333/0.334	0.633/0.647	0.255/0.237	0.779/0.793	0.830/0.795	0.605/0.593
good+high↔poor+low	0.696/0.732	0.744/0.721	0.341/0.320	0.656/0.671	0.307/0.347	0.775/0.796	0.841/0.794	0.622/**0.625**
*good+fine↔*poor+bad	0.674/0.678	0.731/0.735	0.360/0.356	0.632/0.654	0.298/0.343	0.771/0.743	0.847/0.830	0.616/0.619
good+high+fine↔poor+low+bad	0.694/0.712	0.732/0.738	0.363/0.366	0.644/0.613	0.321/0.342	0.779/0.772	0.844/0.820	**0.625**/0.623
InternLM-XComposer-VL (*InternLM*) [61]								
good↔poor	0.564/0.615	0.730/0.750	0.360/0.416	0.612/0.676	0.243/0.265	0.732/0.775	0.546/0.572	0.541/0.581
fine↔bad	0.546/0.597	0.720/0.736	0.341/0.389	0.626/0.671	0.213/0.227	0.681/0.708	0.494/0.479	0.517/0.544
high↔low	0.543/0.590	0.704/0.720	0.331/0.372	0.612/0.656	0.223/0.251	0.716/0.755	0.490/0.500	0.517/0.549
good+high↔poor+low	0.564/0.613	0.723/0.743	0.354/0.405	0.621/0.676	0.238/0.264	0.734/0.775	0.522/0.546	0.537/0.575
good+fine↔poor+bad	0.573/0.626	0.735/0.755	0.366/0.420	0.629/0.687	0.236/0.260	0.732/0.771	0.531/0.551	**0.543/0.581**
good+high+fine↔poor+low+bad	0.571/0.621	0.728/0.748	0.360/0.410	0.629/0.683	0.236/0.261	0.734/0.773	0.521/0.538	0.540/0.576
LLaVA-v1.5 (*Vicuna-v1.5-7B*) [60]								
good↔poor	0.463/0.459	0.443/0.467	0.305/0.321	0.344/0.358	0.321/0.333	0.672/0.738	0.417/0.440	0.424/0.445
fine↔bad	0.453/0.469	0.457/0.482	0.258/0.288	0.303/0.333	0.294/0.302	0.558/0.617	0.389/0.420	0.388/0.416
high↔low	0.474/0.476	0.370/0.386	0.261/0.262	0.432/0.429	0.266/0.269	0.669/0.716	0.304/0.331	0.397/0.410
good+high↔poor+low	0.491/0.491	0.416/0.436	0.293/0.300	0.696/0.751	0.413/0.416	0.298/0.304	0.359/0.389	0.424/0.441
good+fine↔poor+bad	0.482/0.482	0.461/0.485	0.300/0.320	0.644/0.708	0.339/0.357	0.327/0.336	0.425/0.451	0.425/0.449
good+high+fine↔poor+low+bad	0.512/0.513	0.443/0.465	0.303/0.315	0.408/0.415	0.318/0.324	0.697/0.752	0.392/0.421	**0.439/0.458**
LLaVA-v1.5 (*Vicuna-v1.5-13B*) [60]								
good↔poor	0.448/0.460	0.563/0.584	0.310/0.339	0.445/0.481	0.285/0.297	0.664/0.754	0.390/0.400	0.444/0.473
fine↔bad	0.449/0.487	0.583/0.597	0.316/0.360	0.466/0.513	0.349/0.365	0.650/0.749	0.425/0.437	0.463/**0.501**
high↔low	0.456/0.482	0.529/0.553	0.286/0.306	0.489/0.513	0.276/0.284	0.683/0.752	0.316/0.331	0.434/0.460
good+high↔poor+low	0.462/0.484	0.548/0.573	0.303/0.327	0.480/0.509	0.283/0.294	0.687/0.763	0.350/0.363	0.445/0.473
good+fine↔poor+bad	0.463/0.483	0.579/0.596	0.321/0.356	0.467/0.505	0.326/0.339	0.670/0.762	0.420/0.426	**0.464**/0.495
good+high+fine↔poor+low+bad	0.474/0.498	0.565/0.588	0.314/0.345	0.488/0.521	0.311/0.322	0.692/0.771	0.382/0.392	0.461/0.491
Qwen-VL (*QwenLM*) [63]								
good↔poor	0.470/0.546	0.676/0.669	0.298/0.339	0.504/0.532	0.273/0.284	0.617/0.686	0.486/0.486	0.475/0.506
fine↔bad	0.467/0.507	0.352/0.365	0.205/0.238	0.451/0.472	0.188/0.185	0.599/0.627	0.354/0.378	0.374/0.396
high↔low	0.531/0.578	0.626/0.616	0.281/0.290	0.574/0.560	0.286/0.314	0.637/0.692	0.332/0.344	0.467/0.485
good+high↔poor+low	0.539/0.600	0.684/0.673	0.299/0.324	0.565/0.568	0.306/0.330	0.660/0.721	0.414/0.422	**0.495/0.520**
good+fine↔poor+bad	0.495/0.558	0.596/0.581	0.264/0.307	0.521/0.548	0.270/0.270	0.640/0.691	0.435/0.449	0.460/0.486
good+high+fine↔poor+low+bad	0.541/0.600	0.632/0.617	0.286/0.316	0.570/0.577	0.301/0.318	0.664/0.719	0.416/0.429	0.487/0.511
mPLUG-Owl (*LLaMA-7B*) [52]								
good↔poor	0.409/0.427	0.634/0.644	0.241/0.271	0.437/0.487	0.148/0.180	0.687/0.711	0.466/0.486	0.432/0.458
fine↔bad	0.357/0.398	0.622/0.636	0.260/0.290	0.422/0.475	0.178/0.224	0.606/0.646	0.536/0.534	0.426/0.457
high↔low	0.353/0.369	0.610/0.624	0.176/0.187	0.436/0.464	0.110/0.124	0.662/0.663	0.361/0.378	0.387/0.401
good+high↔poor+low	0.382/0.402	0.626/0.642	0.208/0.228	0.446/0.483	0.125/0.144	0.684/0.697	0.409/0.432	0.411/0.432
good+fine↔poor+bad	0.403/0.430	0.635/0.645	0.260/0.292	0.444/0.493	0.172/0.213	0.664/0.694	0.525/0.527	**0.443/0.471**
good+high+fine↔poor+low+bad	0.395/0.421	0.633/0.647	0.233/0.258	0.455/0.496	0.147/0.173	0.685/0.704	0.463/0.483	0.430/0.455

IV. CONCLUSION

In this research, we introduce **Q-Bench**$^+$, a benchmark designed to evaluate the advancements of MLLMs in low-level visual skills. We evaluate the MLLMs from three aspects: **perception** of low-level visual attributes, **description** of low-level visual content, and **assessment** of image quality. Additionally, acknowledging the importance of discerning differences and similarities in image pairs, our benchmark encompasses both single images and image pairs in the **perception** and **description** tasks. For assessing these skills, we have compiled two multi-modal benchmark datasets focused on low-level vision, and introduced a unified softmax-based method for quantitative image quality assessment (IQA) in MLLMs. Our findings demonstrate that some advanced MLLMs exhibit commendable low-level visual skills even without specialized low-level training. However, there's still a significant journey ahead before MLLMs can become fully reliable assistants in general low-level visual tasks. We hope that the insights gained from **Q-Bench**$^+$ will spur further development in MLLMs, particularly in improving their perception and understanding of low-level visual elements.

REFERENCES

[1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[2] M. N. Team. (2023) Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05. [Online]. Available: www.mosaicml.com/blog/mpt-7b

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[4] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[5] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

[6] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[7] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," 2015.

[8] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *ICCV*, 2015.

[9] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding multimodal large language models to the world," *ArXiv*, vol. abs/2306, 2023.

[10] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," *arXiv preprint arXiv:2308.00692*, 2023.

[11] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE TIP*, vol. 29, pp. 4041–4056, 2020.

[12] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *CVPR*, 2020.

[13] S. Su, V. Hosu, H. Lin, Y. Zhang, and D. Saupe, "Koniq++ : Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects," in *The British Machine Vision Conference (BMVC)*, 2021, pp. 1–12.

[14] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Towards explainable video quality assessment: A database and a language-prompted approach," in *ACM MM*, 2023.

[15] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*, 2016.

[16] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *CVPR*, 2012, pp. 2408–2415.

[17] Z. Zhang, W. Sun, T. Wang, W. Lu, Q. Zhou, Q. Wang, X. Min, G. Zhai *et al.*, "Subjective and objective quality assessment for in-the-wild computer graphics images," *ACM TOMM*, 2023.

[18] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, and W. Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," 2023.

[19] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," 2023.

[20] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *ICCV*, 2023.

[21] C. Zhang, S. Su, Y. Zhu, Q. Yan, J. Sun, and Y. Zhang, "Exploring and evaluating image restoration potential in dynamic scenes," in *CVPR*, 2022, pp. 2057–2066.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[23] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in *CVPR*, June 2018.

[24] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," 2020.

[25] Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Z. Luo, Y. Wang, Y. Rao, J. Liu, T. Huang *et al.*, "Generative multimodal models are in-context learners," *arXiv preprint arXiv:2312.13286*, 2023.

[26] SkunkworksAI, "Bakllava," 2024. [Online]. Available: https://github.com/SkunkworksAI/BakLLaVA

[27] Google, "Gemini pro," 2023. [Online]. Available: https://deepmind.google/technologies/gemini

[28] OpenAI, "Gpt-4 technical report," 2023.

[29] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, "Mmbench: Is your multi-modal model an all-around player?" 2023.

[30] J. Lu, J. Rao, K. Chen, X. Guo, Y. Zhang, B. Sun, C. Yang, and J. Yang, "Evaluation and mitigation of agnosia in multimodal large language models," 2023.

[31] T. Guha, V. Hosu, D. Saupe, B. Goldlücke, N. Kumar, W. Lin, V. Martinez, K. Somandepalli, S. Narayanan, W.-H. Cheng, K. McLaughlin, H. Adam, J. See, and L.-K. Wong, "Atqam/mast'20: Joint workshop on aesthetic and technical quality assessment of multimedia and media analytics for societal trends," in *ACM MM*, 2020, p. 4758–4760.

[32] D. Li, T. Jiang, W. Lin, and M. Jiang, "Which has better visual quality: The clear blue sky or a blurry animal?" *IEEE TMM*, vol. 21, no. 5, pp. 1221–1234, 2019.

[33] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *QoMEX*, 2019, pp. 1–3.

[34] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE TIP*, vol. 25, no. 1, pp. 372–387, 2015.

[35] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai, and W. Lin, "Q-bench: A benchmark for general-purpose foundation models on low-level vision," in *ICLR*, 2024.

[36] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv preprint arXiv:1907.07484*, 2019.

[37] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *CVPR*, 2020.

[38] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE*, vol. 25, no. 1, pp. 372–387, 2016.

[39] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *ASILOMAR*, 2012, pp. 1693–1697.

[40] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," 2023.

[41] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] J. Hou, W. Lin, Y. Fang, H. Wu, C. Chen, L. Liao, and W. Liu, "Towards transparent deep image aesthetics assessment with tag-based content descriptors," *IEEE TIP*, 2023.

[43] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE TIP*, 2018.

[44] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, "Rich features for perceptual quality assessment of ugc videos," in *CVPR*, June 2021, pp. 13 435–13 444.

[45] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, J. Gu, and W. Lin, "Neighbourhood representative sampling for efficient end-to-end video quality assessment," 2023.

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

[47] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *ECCV*, 2022.

[48] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-vq: 'patching up' the video quality problem," in *CVPR*, 2021.

[49] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *International Conference on Pattern Recognition*. IEEE, 2010, pp. 2366–2369.

[50] Z. Zhang, W. Sun, X. Min, W. Zhu, T. Wang, W. Lu, and G. Zhai, "A no-reference evaluation metric for low-light image enhancement," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2021, pp. 1–6.

[51] Z. Zhang, W. Sun, X. Min, W. Zhu, T. Wang, and G. Zhai, "A no-reference deep learning quality assessment method for super-resolution images based on frequency maps," in *IEEE International Symposium on Circuits and Systems*, 2022, pp. 3170–3174.

[52] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Jiang, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, and F. Huang, "mplug-owl: Modularization empowers large language models with multimodality," 2023.

[53] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[54] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "nocaps: novel object captioning at scale," in *ICCV*, 2019.

[55] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023.

[56] I. Team, "Infimm: Advancing multimodal understanding from flamingo's legacy through diverse llm integration," 2024. [Online]. Available: https://huggingface.co/Infi-MM/

[57] R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşırlar, "Introducing our multimodal models," 2023. [Online]. Available: https://www.adept.ai/blog/fuyu-8b

[58] Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen *et al.*, "Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models," *arXiv preprint arXiv:2311.07575*, 2023.

[59] Q. Ye, H. Xu, J. Ye, M. Yan, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," *arXiv preprint arXiv:2311.04257*, 2023.

[60] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[61] P. Zhang, X. Dong, B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, S. Ding, S. Zhang, H. Duan, W. Zhang, H. Yan, X. Zhang, W. Li, J. Li, K. Chen, C. He, X. Zhang, Y. Qiao, D. Lin, and J. Wang, "Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition," 2023.

[62] Huggingface, "Introducing idefics: An open reproduction of state-of-the-art visual language model," 2023. [Online]. Available: https://huggingface.co/blog/idefics

[63] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[64] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.

[65] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 320–335.

[66] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, H. Li, and Y. Qiao, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[67] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, K. Xu, C. Li, J. Hou, G. Zhai *et al.*, "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," *arXiv preprint arXiv:2311.06783*, 2023.

[68] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[69] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[70] "Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures," ITU-R Rec. BT.500, 2000.