Q-Instruct: Improving Low-level Visual Abilities for Multi-modality Foundation Models

Haoning Wu[1♡], Zicheng Zhang[3♡], Erli Zhang[1♡],
Chaofeng Chen[1], Liang Liao[1], Annan Wang[1], Kaixin Xu[2,5], Chunyi Li[3], Jingwen Hou[2],
Guangtao Zhai[3], Geng Xue[5], Wenxiu Sun[4], Qiong Yan[4], Weisi Lin[1,2♢]
[1]S-Lab for Advanced Intelligence, [2]Nanyang Technological University
[3]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[4]Sensetime Research
[5]Institute for Infocomm Research (I[2]R), Agency for Science, Technology and Research (A*STAR),
1 Fusionopolis Way, 138632, Singapore

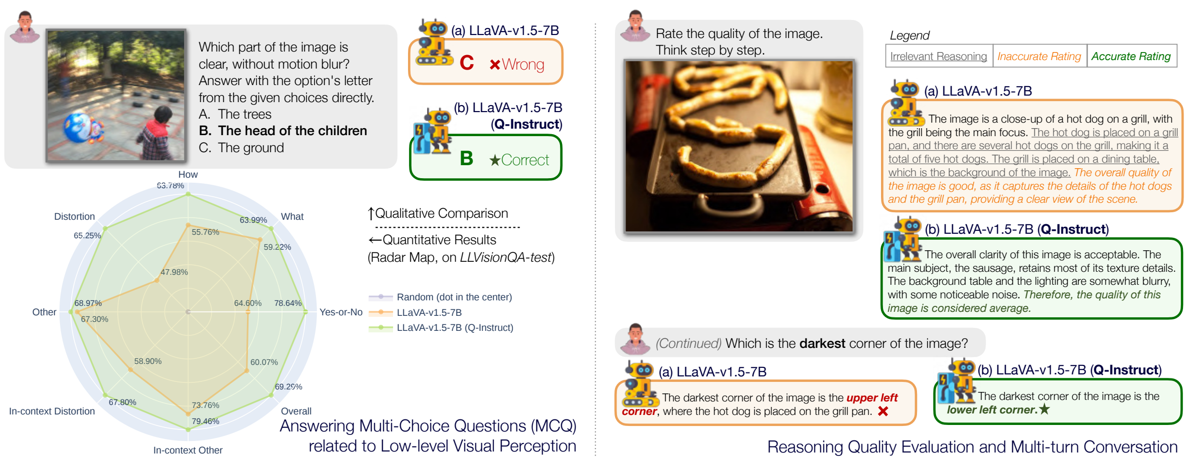


Figure 1. Abilities of **Q-Instruct**-*tuned* LLaVA-v1.5-7B [29] on various low-level visual tasks, in comparison with the baseline version.

Abstract

Multi-modality large language models (MLLMs), as represented by GPT-4V, have introduced a paradigm shift for visual perception and understanding tasks, that a variety of abilities can be achieved within one foundation model. While current MLLMs demonstrate primary ***low-level visual abilities*** *from the identification of low-level visual attributes (e.g., clarity, brightness) to the evaluation on image quality, there's still an imperative to further improve the accuracy of MLLMs to substantially alleviate human burdens. To address this, we collect the first dataset consisting of human natural language feedback on low-level vision. Each feedback offers a comprehensive description of an image's low-level visual attributes, culminating in an overall quality assessment. The constructed* ***Q-Pathway*** *dataset includes 58K detailed human feedbacks on 18,973 multi-sourced images with diverse low-level appearance. To ensure MLLMs can adeptly handle diverse queries, we further propose a GPT-participated transformation to convert these feedbacks into a rich set of 200K instruction-response pairs, termed* ***Q-Instruct****. Experimental results indicate that the* ***Q-Instruct*** *consistently elevates various low-level visual capabilities across multiple base models. We anticipate that our datasets can pave the way for a future that foundation models can assist humans on low-level visual tasks.*

1. Introduction

Computer vision has witnessed a recent paradigm shift attributed to the emergence of multi-modality large language models (MLLMs) [7, 11, 30, 37]. These models aim to transcend traditional task-specific experts, and serve as general-purpose foundation models capable of facilitating humans across a variety of visual tasks [25]. Specifically, these foundation models also bring exciting potentials in the domain of **low-level visual perception and understanding**. This domain includes not only commonly-focused image

♡Equal contribution. ♢Corresponding author.
♣Project Page: *https://q-future.github.io/Q-Instruct*

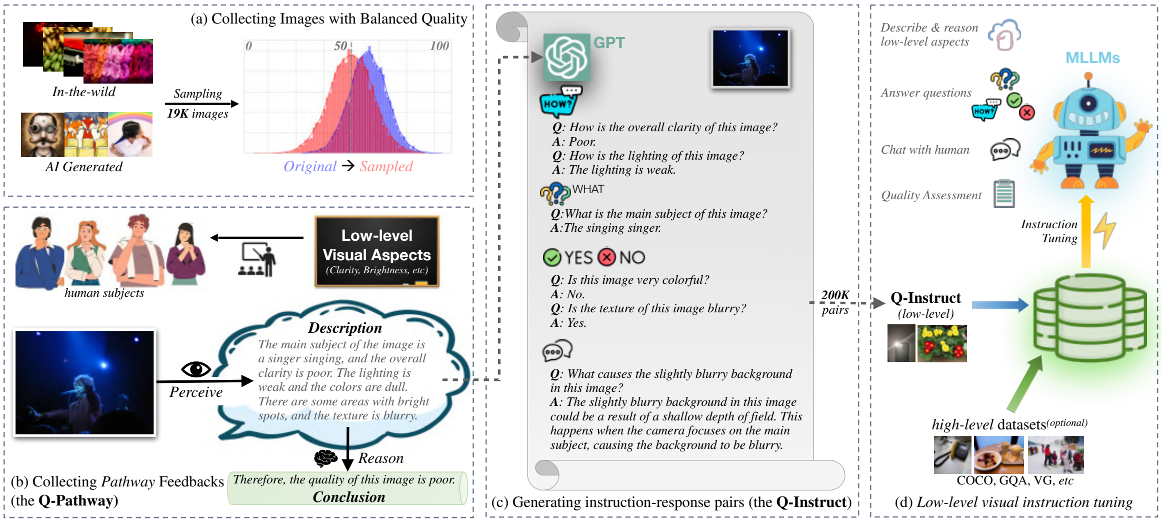


Figure 2. Data construction pipeline. First, we collect **58K** human feedbacks on low-level visual aspects (the **Q-pathway**, a/b); they are then converted into with **200K** instruction-response pairs (the **Q-Instruct**, c), which are used for (d) low-level visual instruction tuning.

quality assessment (IQA) [14, 54, 59] tasks, but also finer-grained abilities to identify the low-level visual attributes (*noise, blur, etc*) [43], or evaluate the low-level visual dimensions (*clarity, brightness, etc*) [9, 55]. As human cognition associated with these tasks is highly interconnected, we aspire for a unified foundation model to establish general abilities across these tasks, which could robustly respond to open-ended human queries on low-level visual aspects.

Nevertheless, though existing MLLMs can basically reply to human queries regarding low-level visual aspects, the accuracy of their responses remains unsatisfactory [31, 56] (Fig. 1(a)). The primary problem is the lack of low-level visual datasets during training MLLMs, where publicly available datasets generally only focus on high-level visual abilities [2, 16, 22, 32]. To solve this problem, we construct the **Q-Instruct**, the first large-scale *low-level visual instruction tuning* dataset, in the following two steps:

Step 1: Collect human feedbacks for low-level vision.

For this step, we invite human subjects to provide direct feedbacks on their low-level perception and understanding over a variety of images (Fig. 2(b)). Specifically, each feedback should include two parts: 1) Primarily, an exhaustive **description** on elemental low-level attributes (*e.g. blurs, noises, clarity, color, brightness*). Such descriptions should also include content [27, 48] or position [51, 59] contexts (*e.g. the duck / the left part of the image is under-exposed*) that are related to low-level attributes. 2) Then, an overall **conclusion** on the image quality based on the description of the attributes. With the two parts, the feedbacks, denoted as ***pathway*** feedbacks, not only record fundamental human low-level perception but also reflect the human reasoning process on evaluating visual quality. The hence-constructed **Q-Pathway** dataset (Fig 2(b)) contains 58K pathway feedbacks on 18,973 multi-sourced images, each image with at least three feedbacks (*avg. 46.4 words per feedback*).

Step 2: Convert these feedbacks for instruction tuning.

While these *pathway* feedbacks themselves make up an important subset for the *low-level visual instruction tuning*, the full instruction tuning dataset should be designed to activate more capabilities. Primarily, it should also include a low-level *visual question answering* (VQA) subset. To generate a reliable VQA subset, we refer to the setting that how COCO-VQA [2] is derived from image captions, and employ GPT [36] to convert the *pathway* feedbacks into question-answer pairs with adjectives (*e.g. good/fair/poor*) or nouns (*e.g. noise/motion blur*) as answers. Similarly, we also collect a balanced *yes-or-no* question-answer set based on the information in the feedbacks (*answered with yes*), or information contrast to the feedbacks (*answered with no*); some context-related question-answer pairs are also created to better ground [61] the low-level attributes. Following existing studies [40], all question-answer pairs in the VQA subset include both multiple-choice (*A/B/C/D*) and direct-answer settings. Furthermore, besides the VQA subset, with the assistance of GPT, we also collect a subset of long conversations related to the low-level concerns (*e.g. why the distortions happen*, *how to improve the picture quality*). The subsets compose into the **Q-Instruct** dataset (Fig. 2(c)) with 200K instruction-response pairs, which are designed to enhance MLLMs on a variety of low-level visual abilities.

The core contributions of our study can be summarized as follows: **1)** We collect the **Q-Pathway**, a multi-modality dataset for low-level visual perception and quality assessment, which includes direct human feedbacks (*with reasoning*) on low-level visual aspects. **2)** Based on **Q-Pathway**, we construct the **Q-Instruct**, the first instruction tuning dataset that focuses on human queries related to low-level vision. **3)** Our rich experiments on *low-level visual instruc-*

tion tuning ((Fig. 2 (d)) validate that the **Q-Instruct** improve various low-level abilities of MLLMs (Fig. 1), and bring insights for future studies to inject various low-level visual abilities into the scope of general foundation models.

2. Related Works

2.1. Low-level Visual Perception

Tasks and Datasets. Image quality assessment (IQA), targeting to predict accurate scores aligned with integrated human opinions on all low-level aspects, has always been the chief task in low-level visual perception. Many datasets are developed to address IQA on artificially-distorted images [17, 28] (*JPEG, AWGN, etc*), in-the-wild photographs [14, 59], or recently-popular AI-generated contents [26, 57], providing important metrics for visual content production and distribution. Despite general IQA, recent studies have started to focus on finer-grained low-level visual aspects, and explored some related tasks such as evaluating on low-level visual dimensions (*e.g. color, brightness*) [9, 55], or distinguishing the existing distortions (*e.g. blur, noise, over-exposure*) in images [43]. Some recent works [52–54] also consider some photography-related dimensions (*e.g. composition, lighting, bokeh*) [21] as a broader sense of low-level aspects. In general, low-level visual perceptual tasks can include all aspects of image appearance (*in contrast to object-level contents*) that can be perceived by human and evoke different human feelings. While these low-level visual tasks used to be tackled separately, the proposed datasets bring the opportunities to include, relate, and learn these tasks together, supporting one foundational model to generally master these tasks.

Approaches. Similarly, the approaches designed for low-level visual perception also basically focus on their general IQA abilities. The traditional IQA metrics, *e.g.* NIQE [34], operate on discipline-based methodologies without training with human opinions, offering robust but less accurate evaluations. In contrast, deep learning-based methods [4, 8, 18, 42, 50, 63] utilize task-specific data, capitalizing on the extensive learning capacities of neural networks to tailor their assessment to particular data distributions, while they also suffer from compromised generalization abilities. Notably, recent methods [15, 19, 47, 64, 66] explore CLIP [38] for IQA, which stand out for their pioneer efforts on ***multi-modality integration*** for low-level vision, and exciting zero-shot performance. Their zero-shot IQA abilities are also inherited by most recent MLLMs [3, 29, 62]. Similar to NIQE, these multi-modality IQA methods are robust on various scenarios, yet not enough accurate on each single case. While these methods present improving performance on general IQA, the other finer-grained low-level visual perception abilities are still yet to be deeply investigated; moreover, tackling all these tasks separately may overlook the underlying relationships between them, refraining from reasoning among these sections. After instruction tuning with the proposed **Q-Instruct**, MLLMs can significantly improve their abilities on various low-level visual abilities, forecasting a future to unify these tasks through one model.

Table 1. The **Q-Pathway** compared to its sources. We sub-sample the source images to reduce the ***skews*** in their MOS distributions, resulting in the sampled distribution to be further <u>balanced</u>.

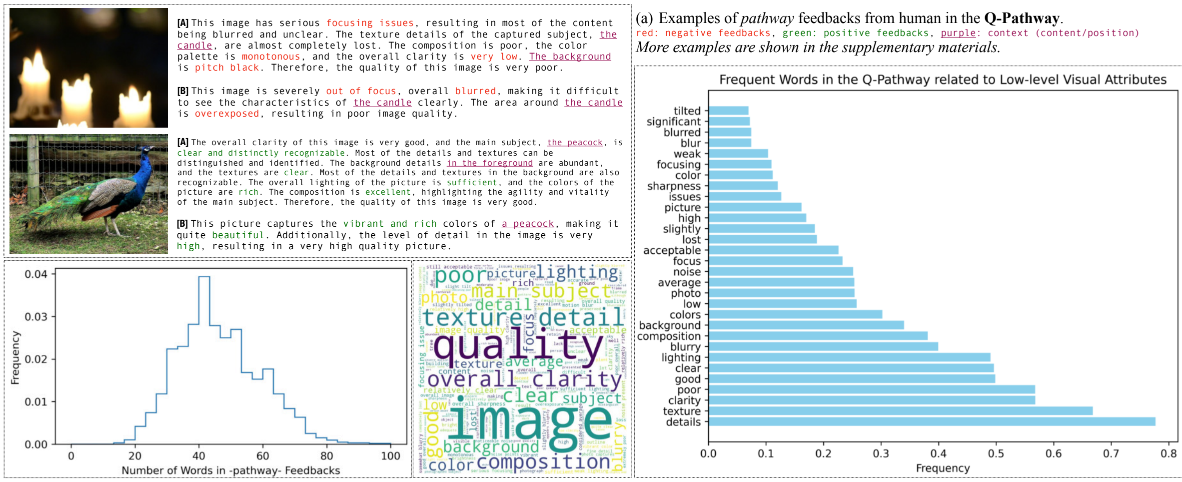
Image Sources	Original Distribution			Sampled Distribution		
MOS $\in [0, 100)$	Size	μ_{MOS}	σ_{MOS}	Size	μ_{MOS}	σ_{MOS}
KonIQ-10k [14]	10,073	58.73	15.43	5,182	49.53	15.72
SPAQ [9]	11,125	50.32	20.90	10,797	49.46	20.63
LIVE-FB [59]	39,810	72.13	6.16	800	60.68	17.38
LIVE-itw [12]	1,169	55.38	20.27	200	55.70	19.83
AGIQA-3K [26]	2,982	50.00	19.80	400	40.80	21.80
ImageRewardDB [57]	50,000	- *w/o* MOS -		584	- *w/o* MOS -	
15-*distortion* COCO [5]	330,000	- *w/o* MOS -		1,012	- *w/o* MOS -	
Overall	445,159	65.02	16.51	18,973	<u>49.87</u>	<u>19.08</u>

2.2. Multi-modality Large Language Models

Large language models (LLMs), *e.g.* GPT-4 [37], T5 [6], LLaMA [45], has shown great language abilities regarding general human knowledge. With CLIP [38] and additional adapting modules to involve visual inputs into LLMs, the multi-modality large language models (MLLMs) [7, 11, 24, 30, 62] can tackle a variety of multi-modality tasks for high-level vision, such as *image captioning* [1, 5, 60], *visual question answering* (VQA) [2, 32, 40], and more language-related capabilities [10, 23, 31]. Nevertheless, the evaluation results in the recent benchmark [56] reveal that MLLMs' low-level visual abilities are still unsatisfactory, especially when it comes to the *finer-grained* low-level perception questions. While we notice that this is mainly due to the lack of respective data, we collect the first *low-level visual instruction tuning* dataset, the **Q-Instruct**, to improve low-level visual abilities for different MLLMs, and bring them into the realm of low-level visual perception.

3. the *Q-Pathway*

As the fundamental part of the dataset construction, we introduce the **Q-Pathway**, the first large-scale dataset that collects **text** feedbacks from human on low-level visual aspects. To diversify and balance different low-level appearances, we sub-sample images from **seven** sources (Sec. 3.1) and reduce the ***skews*** in the source distributions (Tab. 1). After the preparation of images, we discuss the rationality and the detailed task definition for the *pathway* feedbacks (Sec. 3.2), a kind of natural language feedback, as collected in the **Q-Pathway**. The subjective study is conducted **in-lab** (Sec. 3.3), where all subjects are trained before providing feedback. The analysis of the **Q-Pathway** is in Sec. 3.4.



(a) Examples of *pathway* feedbacks from human in the **Q-Pathway**.
More examples are shown in the supplementary materials.

(b) The distribution of feedback lengths in the **Q-Pathway**. (c) Wordcloud of the **Q-Pathway**. (d) Top-frequency words related to low-level vision in the **Q-Pathway**.

Figure 3. (a) Example *pathway* feedbacks, each containing a detailed description followed by an overall evaluation, with context included. (b) The distribution of *pathway* feedback lengths. (c) *Wordcloud* of the **Q-Pathway**. (d) Top-frequency words related to low-level vision.

3.1. Preparation of Images

The images in the **Q-Pathway** are sampled from various sources, including four *in-the-wild* IQA datasets [9, 12, 14, 59], and two datasets with *AI-generated* images [26, 57]. Specifically, as compared in Tab. 1, the sub-sampled population of images is carefully constructed to introduce more diverse low-level appearances in the **Q-Pathway**, with a balance between high-quality and low-quality images. Moreover, to further diversify the low-level appearances of the collected images, we design a custom variant of *imagecorruptions* [33] to randomly corrupt 1,012 originally-pristine images from COCO [5] dataset with one in *15* artificial distortions. The assembled sub-sampled dataset consists of **18,973** images, which are further fed to human subjects to provide *pathway* feedbacks.

3.2. Task Definition: the *pathway* Feedbacks

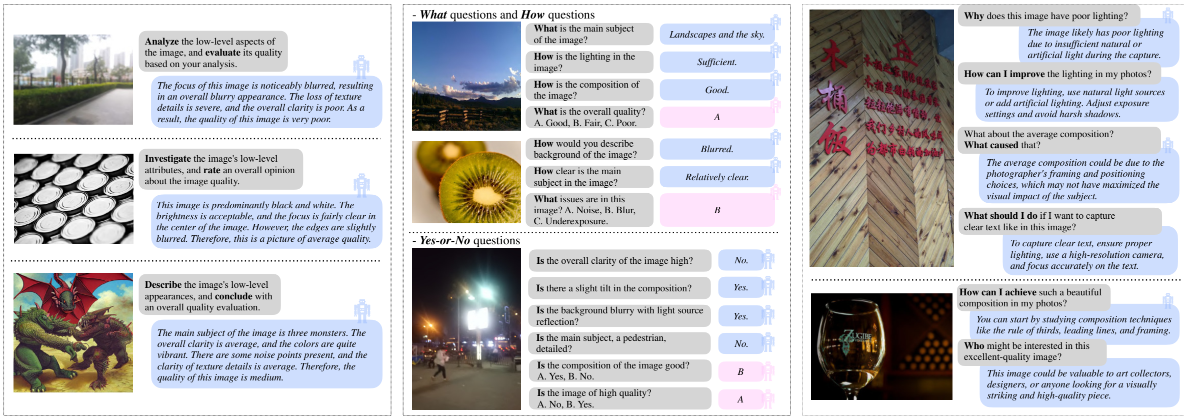
For the **Q-Pathway**, to collect a richer and more nuanced understanding of human perception on low-level visual aspects, instead of collecting multi-dimensional scores as in existing studies [9, 55], we opt to collect a new format of annotation, termed *pathway* feedbacks, with an exhaustive natural language description on low-level visual attributes *e.g. noise, brightness, clarity*) followed by a general conclusion. The rationales for this format are as follows: **(1)** Primarily, the descriptions can preserve what humans perceive more *completely* and *precisely*. For instance, if an image has both dark and bright areas such as Fig 3(a) *upper*, the brightness score might not properly record [51, 59] this situation: the positional context cannot be preserved, and the reliability of the score could also be compromised, as neither labeling it as 'dark' nor as 'bright' is accurate. **(2)** Moreover, unlike free-form text feedbacks, the order of the two parts in *pathway* feedbacks generally aligns with the human reasoning process. For instance, while human subjects are shown with an *underexposed* yet **clear** image, they can provide intuitive reasoning leading to eclectic conclusions like "*Thus, the quality of the image is acceptable*". This reasoning will help MLLMs to better emulate human perception and understanding related to low-level vision. While this *pathway*-style format faces challenges to be transformed into machine learning objectives in the past, the emergence of MLLMs has provided the opportunity to learn from these direct human feedbacks, in order to allow machines to more precisely and robustly align with human perception.

3.3. The subjective study process.

The subjective study is carried out in a well-controlled laboratory environment, during which a total of 39 **trained** human subjects are invited. Based on task definition, training material includes not only calibration on *overall quality*, but also on the *respective text descriptions* of different low-level appearances shown in visuals. Furthermore, as the majority of images come from IQA datasets, the mean opinion scores (MOSs) of them are also displayed to subjects to better calibrate them with a common understanding of *quality*. To facilitate their feedback process, we also show a reference attribute set that can be used in the descriptions. To avoid test fatigue of subjects, consecutive feedbacks on more than 30 images will be warned and discouraged; it will be further forcefully paused after 50 images. 58K *pathway* feedbacks are collected during the study, as exemplified in Fig. 3(a).

3.4. Analysis

After the subjective study, we briefly analyze the collected feedbacks. Qualitatively (Fig. 3(a)), the *pathway* feedbacks can generally preserve the respective contexts related to low-level attributes. Moreover, feedbacks from different human subjects for the same image (as exemplified in [A]



(a) Pathway Reasoning for *explainable* IQA (b) Low-level *Visual Question Answering* (VQA) (c) Extended Conversations

Figure 4. The composition of the **Q-Instruct** dataset, in which the **200K** instruction-response pairs include (a) **58K** pathway reasoning, (b) *visual question answering*, with **76K** *what/how* questions and **57K** balanced *yes-or-no* questions, and (c) **12K** extended conversations.

and [B] for each image) shows decent consistency (*no controversial information*), and slightly complements one another. Statistically, the length of feedbacks generally ranges from 20 to 100 words, with an average of **46.4** words, 4 times as long as common high-level image captions [5, 60] (Fig 3(b)). We also visualize the wordcloud [35] and the bar chart for the top frequency words related to low-level vision, demonstrating that the collected **Q-Pathway** covers a wide range of low-level attributes, and includes positive and negative feedbacks within similar proportions.

4. the *Q-Instruct*

The long and diverse feedbacks in the **Q-Pathway** provides sufficient reference for the automatic generation process of instruction-response pairs to be used for low-level visual instruction tuning. While the *pathway* feedbacks themselves can teach MLLMs to reason low-level aspects and predict quality (Sec. 4.1), we design more instruction types to allow MLLMs to respond to a variety of human queries, including a *visual question answering* subset (Sec. 4.2) for more accurate low-level perception ability [56], and an extended conversation subset (Sec. 4.3) to allow MLLMs to seamlessly *chat* with human about topics related to low-level visual aspects. Overall, the **Q-Instruct** dataset includes 200K instruction-response pairs, with its details as follows.

4.1. Low-level Reasoning with *pathway* Feedbacks

Similar as image captioning [1, 5, 60], a general low-level visual description ability is also vital for MLLMs. As analyzed in Fig. 3, the pathway *feedbacks* are direct and holistic human responses that generally describe low-level visual appearances. Furthermore, these feedbacks provide ***reasoning*** from low-level attributes (*brightness, clarity*) to overall quality ratings (*good/poor*), which could activate the potential reasoning abilities [20, 49] of MLLMs on IQA. Henceforth, with each *pathway* feedback as response and a general prompt as instruction, we include **58K** pathway reasoning (Fig. 4(a)) as the primary part of the **Q-Instruct** dataset.

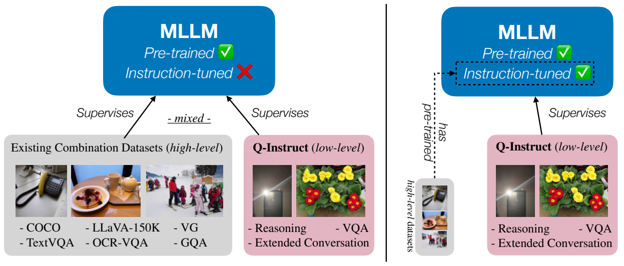
4.2. Visual Question Answering (VQA)

Besides directly applying the **Q-Pathway** into low-level visual instruction tuning, we also design a GPT [36]-participated pipeline to convert them into a *visual question answering* (VQA) subset. In general, we ask GPT to generate diverse-style questions related to low-level-vision from the *pathway* feedbacks, and provide answers with *as few words as possible*. Via this process, we convert the feedbacks into **76K** questions, including *how* questions answered with opinion-related adjectives (*e.g. good/poor, high/low*), or *i.e.* ***what*** questions answered with attribute-related (*blur/noise/focus*) or context-related (*left/the peacock/the background*) nouns, as shown in the *upper* part of Fig. 4(b). We further instruct GPT to generate binary judgments (*yes/no*, Fig. 4(b) *lower*) from the feedbacks, and balance *yes* and *no* into 1:1 ratio, with **57K** *yes-or-no* questions collected at last. As for the answering format, following A-OKVQA [40], despite the direct answers, we also create several distracting answers for the questions, and convert them into an additional multi-choice question (MCQ) format (*the pink boxes* in Fig. 4(b)).

4.3. Extended Conversations

While the first two subsets are designed to enhance the fundamental language-related abilities for low-level vision, the third subset of the **Q-Instruct**, the *extended conversations* (Fig. 4(c)), focuses on improving the ability to discuss with human grounded on the low-level visual aspects of an input image. These discussions include five major scopes: **1)** Examining the causes of low-level visual patterns; **2)** Providing improvement suggestions on photography; **3)** Providing tools to restore, enhance, or edit the image; **4)** Recommend-

For better visualization, the two words that appear in every feedback, ***image*** and ***quality***, are removed from the bar chart in Fig. 3(d).



(a) ***mix*** **Q-Instruct** with *high-level* instruction tuning datasets

(b) **Q-Instruct** ***after*** *high-level* instruction tuning

Figure 5. Training strategies for *low-level visual instruction tuning* evaluated in our study, including (a) ***mix*** the **Q-Instruct** with high-level visual instruction tuning datasets, (b) conduct a further low-level tuning stage with only **Q-Instruct** ***after*** high-level tuning.

ing the image to respective consumers; **5)** Other conversations that may happen given the low-level visual descriptions provided in the *pathway* feedbacks. Similarly, the extended conversation subset is also generated by GPT, with in total **12K** conversations collected for the **Q-Instruct**.

5. Low-level Visual Instruction Tuning

In this section, we discuss the standard training strategies for *low-level visual instruction tuning*, *i.e.* when to involve the **Q-Instruct** dataset during the training of MLLMs. In general, the training of open-source MLLMs [7, 24, 62] includes two stages: **First,** aligning the representation space of the visual backbone and the LLM with million-scale web data [39, 41]. **Second,** visual instruction tuning with a combination of human-labeled datasets [2, 5, 32, 61]. Considering the scale of the **Q-Instruct**, a general strategy is to ***mix*** its instruction-response pairs with the high-level datasets in the **second** stage, so as to ideally built their low-level visual abilities within general high-level awareness, as shown in Fig. 5(a). Another faster and more convenient strategy (without requiring to train high-level datasets again) is a further **third** stage only with the **Q-Instruct** (Fig. 5(b)) ***after*** original high-level tuning. In our experiments, we validate that they both bring notable improvements on various low-level visual tasks, and involving *high-level* awareness contributes to the effectiveness of both strategies.

6. Experiments

6.1. Experimental Setups

Baseline models. We pick four variants of three state-of-the-art MLLMs within diverse meta structures (Tab. 2) as baseline models to evaluate their low-level visual abilities *before* and *after* training with the **Q-Instruct**. Each model is evaluated under both strategies as in Fig. 5, with the original combination of *high-level* datasets unchanged.

Table 2. Baseline MLLMs for *low-level visual instruction tuning*.

Month/Year **Model Name**	**Visual Backbone**	**V→L Module**	**Language Model**
Oct/23 LLaVA-v1.5 (*7B*) [29]	CLIP-ViT-L14$^{\uparrow 336}$	MLP	Vicuna-v1.5-7B [67]
Oct/23 LLaVA-v1.5 (*13B*) [29]	CLIP-ViT-L14$^{\uparrow 336}$	MLP	Vicuna-v1.5-13B [67]
Oct/23 mPLUG-Owl-2 [58]	CLIP-ViT-L14$^{\uparrow 448}$	Abstractor	LLaMA2-7B [46]
Sep/23 InternLM-XComposer-VL [62]	EVA-CLIP-G	Perceive Sampler	InternLM-7B [44]

6.2. Tasks for Evaluation

The low-level visual abilities of MLLMs after *low-level visual instruction tuning* are quantitatively evaluated in three tasks defined by [56]: **(A1) Perception.** The perception task examines whether the MLLM correctly answers questions for low-level attributes (*e.g. clarity, brightness*). We use **LLVisionQA** dataset as evaluation set, which contains 2990 multi-choice questions (MCQs) about low-level vision, and is equally split into two subsets (*dev/test*) with each 1495 MCQs. **(A2) Description.** The description task evaluates whether the MLLM can precisely describe the low-level appearance of the given image. We adopt the **LLDescribe** dataset as evaluation set with 499 images. This task employs GPT to compare between MLLM outputs and GT descriptions in terms of *precision, completenss,* and *relevance*, each scored in range [0,2] as the metric. **(A3) Quality Assessment.** The quality assessment task examines whether MLLMs (***w/o directly trained to score***) can precisely predict visual quality through output logits (*good* against *poor*), on 8 major quality assessment datasets.

6.3. Main Results

(A1) Perception (MCQ). From Tab. 3 and Tab. 4, we observe that either strategy of including **Q-Instruct** into the training of MLLMs can significantly improve their low-level perception ability. The results demonstrate the effectiveness of the proposed pipeline to automatically generate the VQA subset (*including MCQ*) from the pathway feedbacks via GPT, which could be expected to extend to further query types. Specifically, among all dimensions, we notice that the accuracy on *Yes-or-No* question type is most significantly enhanced (*avg. more than 10%*). Moreover, improvements on **distortions** are more significant than on **other** low-level attributes (*aesthetics, photography techniques*), suggesting that the major concerns as raised by human in the **Q-Pathway** are still related to distortions. We hope that our pipeline can be extended to cover more types of questions and a broader range of concerns in the future.
(A2) Description. As shown in Tab. 5, the Q-Instruct tuning also notably improves the low-level description ability of MLLMs, especially on the *relevance* (*+0.31*), with all *tuned* variants obtaining more than 1.5/2 average score. In contrast, the improvements on *completeness* (*+0.17*) and *precision* (*+0.04*) are less significant and can still be improved in the future. **(A3) Image Quality Assessment**

Quality Score $= e^{x^{good}}/(e^{x^{good}}+e^{x^{poor}})$. x^T denotes *logprob* for T.

Table 3. Comparison of the low-level **Perception** ability between baseline MLLMs and **Q-Instruct**-*tuned* versions, on **LLVisionQA**-*dev*. *Yes-or-No* denotes questions with answers only *yes* or *no*, while *What* and *How* denote questions starting with *what* and *how*.

Model *(variant)*	*Q-Instruct* Strategy	*Yes-or-No*↑	*What*↑	*How*↑	*Distortion*↑	*Other*↑	*I-C Distortion*↑	*I-C Other*↑	*Overall*↑
random guess	–	50.00%	27.86%	33.31%	37.89%	38.48%	38.28%	35.82%	37.80%
LLaVA-v1.5 (*7B*)	*no* (Baseline)	66.36%	58.19%	50.51%	49.42%	65.74%	54.61%	70.61%	58.66%
	(a) ***mix*** with high-level	76.18%$_{+9.82\%}$	66.37%$_{+8.18\%}$	57.61%$_{+7.10\%}$	65.18%$_{+15.76\%}$	67.59%$_{+1.85\%}$	64.80%$_{+10.19\%}$	73.06%$_{+2.55\%}$	67.09%$_{+8.43\%}$
	(b) ***after*** high-level	76.91%$_{+10.45\%}$	65.04%$_{+6.85\%}$	55.78%$_{+5.27\%}$	64.01%$_{+14.59\%}$	67.13%$_{+1.39\%}$	64.80%$_{+10.19\%}$	71.84%$_{+1.23\%}$	66.35%$_{+7.69\%}$
LLaVA-v1.5 (*13B*)	*no* (Baseline)	65.27%	64.38%	56.59%	56.03%	67.13%	61.18%	67.35%	62.14%
	(a) ***mix*** with high-level	76.18%$_{+10.91\%}$	65.71%$_{+1.33\%}$	59.23%$_{+2.64\%}$	64.39%$_{+8.36\%}$	69.91%$_{+2.78\%}$	62.50%$_{+1.32\%}$	75.51%$_{+8.16\%}$	67.42%$_{+5.28\%}$
	(b) ***after*** high-level	76.36%$_{+11.09\%}$	65.04%$_{+0.66\%}$	58.42%$_{+1.83\%}$	65.56%$_{+9.53\%}$	66.44%$_{-0.69\%}$	64.47%$_{+3.29\%}$	74.29%$_{+6.94\%}$	67.02%$_{+4.88\%}$
mPLUG-Owl-2	*no* (Baseline)	72.18%	57.96%	56.19%	56.68%	69.21%	53.29%	72.65%	61.61%
	(a) ***mix*** with high-level	75.64%$_{+3.46\%}$	67.04%$_{+9.08\%}$	59.03%$_{+2.84\%}$	71.01%$_{+14.33\%}$	65.28%$_{-3.93\%}$	63.16%$_{+9.87\%}$	69.80%$_{-2.85\%}$	67.56%$_{+5.95\%}$
	(b) ***after*** high-level	76.00%$_{+3.82\%}$	65.04%$_{+7.08\%}$	61.66%$_{+5.47\%}$	65.95%$_{+9.27\%}$	68.75%$_{-0.46\%}$	65.46%$_{+12.17\%}$	73.88%$_{+1.23\%}$	67.96%$_{+6.35\%}$
InternLM-XComposer-VL	*no* (Baseline)	69.45%	65.27%	60.85%	61.67%	70.14%	56.91%	75.10%	65.35%
	(a) ***mix*** with high-level	76.73%$_{+7.28\%}$	69.91%$_{+4.64\%}$	63.89%$_{+3.04\%}$	70.23%$_{+8.56\%}$	71.53%$_{+1.39\%}$	67.43%$_{+10.52\%}$	72.65%$_{-2.45\%}$	70.43%$_{+5.08\%}$
	(b) ***after*** high-level	78.36%$_{+8.91\%}$	68.58%$_{+3.31\%}$	63.08%$_{+2.23\%}$	65.37%$_{+3.70\%}$	73.15%$_{+3.01\%}$	68.42%$_{+11.51\%}$	78.37%$_{+3.27\%}$	70.37%$_{+5.02\%}$

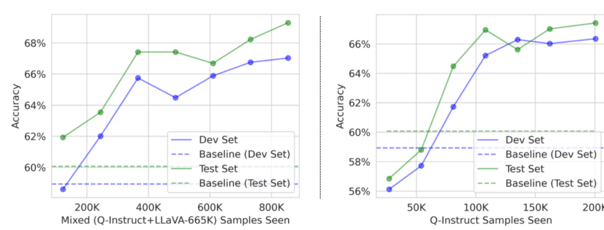
Table 4. Comparison of the low-level **Perception** ability between baseline MLLMs and **Q-Instruct**-*tuned* versions, on **LLVisionQA**-*test*.

Model *(variant)*	*Q-Instruct* Strategy	*Yes-or-No*↑	*What*↑	*How*↑	*Distortion*↑	*Other*↑	*I-C Distortion*↑	*I-C Other*↑	*Overall*↑
random guess	–	50.00%	28.48%	33.30%	37.24%	38.50%	39.13%	37.10%	37.94%
LLaVA-v1.5 (*7B*)	*no* (Baseline)	64.60%	59.22%	55.76%	47.98%	67.30%	58.90%	73.76%	60.07%
	(a) ***mix*** with high-level	78.65%$_{+14.05\%}$	63.99%$_{+4.77\%}$	63.79%$_{+8.03\%}$	65.26%$_{+17.28\%}$	68.97%$_{+1.67\%}$	67.81%$_{+8.91\%}$	79.47%$_{+5.71\%}$	69.30%$_{+9.23\%}$
	(b) ***after*** high-level	78.46%$_{+13.86\%}$	63.34%$_{+4.12\%}$	58.85%$_{+3.09\%}$	60.46%$_{+12.48\%}$	68.74%$_{+1.44\%}$	69.52%$_{+10.62\%}$	76.81%$_{+3.05\%}$	67.42%$_{+7.35\%}$
LLaVA-v1.5 (*13B*)	*no* (baseline)	64.96%	64.86%	54.12%	53.55%	66.59%	58.90%	71.48%	61.40%
	(a) ***mix*** with high-level	77.19%$_{+13.23\%}$	68.55%$_{+3.69\%}$	65.43%$_{+11.31\%}$	64.68%$_{+11.13\%}$	71.12%$_{+4.43\%}$	67.47%$_{+8.57\%}$	85.55%$_{+14.07\%}$	70.70%$_{+9.30\%}$
	(b) ***after*** high-level	80.66%$_{+15.70\%}$	67.25%$_{+2.39\%}$	61.93%$_{+7.81\%}$	66.03%$_{+12.48\%}$	70.41%$_{+3.82\%}$	69.86%$_{+10.96\%}$	79.85%$_{+8.37\%}$	70.43%$_{+9.03\%}$
mPLUG-Owl-2	*no* (Baseline)	72.26%	55.53%	58.64%	52.59%	71.36%	58.90%	73.00%	62.68%
	(a) ***mix*** with high-level	78.47%$_{+6.21\%}$	67.90%$_{+12.37\%}$	63.37%$_{+4.73\%}$	68.52%$_{+15.93\%}$	68.02%$_{-3.34\%}$	70.21%$_{+11.31\%}$	77.57%$_{+4.57\%}$	70.30%$_{+7.62\%}$
	(b) ***after*** high-level	78.47%$_{+6.21\%}$	60.74%$_{+5.21\%}$	66.46%$_{+7.82\%}$	63.34%$_{+10.75\%}$	71.36%$_{\pm 0}$	68.15%$_{+9.25\%}$	77.95%$_{+4.95\%}$	69.10%$_{+6.42\%}$
InternLM-XComposer-VL	*no* (Baseline)	68.43%	62.04%	61.93%	56.81%	70.41%	57.53%	77.19%	64.35%
	(a) ***mix*** with high-level	78.65%$_{+10.22\%}$	68.33%$_{+6.29\%}$	66.26%$_{+4.33\%}$	70.24%$_{+13.43\%}$	71.12%$_{+0.81\%}$	68.15%$_{+10.62\%}$	77.95%$_{+0.76\%}$	71.44%$_{+7.09\%}$
	(b) ***after*** high-level	79.56%$_{+11.13\%}$	64.64%$_{+2.60\%}$	65.43%$_{+3.50\%}$	64.30%$_{+7.49\%}$	71.60%$_{+1.19\%}$	66.44%$_{+8.91\%}$	84.79%$_{+7.60\%}$	70.37%$_{+6.02\%}$

Table 5. Comparison of the low-level **Description** ability between baseline MLLMs and **Q-Instruct**-*tuned* versions, under the same prompt: *"Describe and evaluate the quality of the image."*

Model *(variant)*	*Q-Instruct* Strategy	*completeness*	*precision*	*relevance*	*sum*
LLaVA-v1.5 (*7B*)	*no* (Baseline)	0.90	1.13	1.18	3.21
	(a) ***mix*** w/ high-level	**1.12**	**1.17**	**1.57**	**3.86**
	(b) ***after*** high-level	1.11	1.16	1.54	3.82
LLaVA-v1.5 (*13B*)	*no* (Baseline)	0.91	1.28	1.29	3.47
	(a) ***mix*** w/ high-level	**1.14**	**1.29**	1.58	**4.01**
	(b) ***after*** high-level	1.13	1.26	**1.61**	4.00
mPLUG-Owl-2	*no* (Baseline)	1.06	1.24	1.36	3.67
	(a) ***mix*** w/ high-level	**1.18**	**1.29**	**1.57**	**4.04**
	(b) ***after*** high-level	1.16	1.27	**1.57**	3.99
InternLM-XComposer-VL	*no* (Baseline)	1.03	1.26	1.27	3.56
	(a) ***mix*** w/ high-level	1.16	**1.35**	**1.63**	**4.14**
	(b) ***after*** high-level	**1.18**	1.34	1.62	**4.14**
Average Improvement		*+0.17*	*+0.04*	*+0.31*	*+0.52*

(IQA). Despite the two directly tuned tasks, we follow the *softmax* pooling strategy [56] to extract quality scores from MLLMs and evaluate their IQA ability, as listed in Tab. 6. Primarily, we notice the excellent performance on two "*mostly seen*" datasets. As we do not directly use any MOS values during training, it suggests that we may teach MLLMs to effectively learn to score **without any numerical values** as supervision. This result by-side suggests the high reliability of the proposed datasets. The more exciting results are the huge improvements on "*barely seen*" (with a small proportion of images sampled into the **Q-Instruct**) and even "*never seen*" (cross-set) datasets. Considering the



(a) under ***mix*** strategy (b) under ***after*** strategy

Figure 6. Accuracy on MCQ questions with respect to data samples seen during training (*in comparison with baseline*), demonstrating the effectiveness of scaling up the **Q-Instruct** dataset.

three "*never seen*" datasets [13, 28, 65] (*with computer-generated images, artificially-degraded image, and even videos respectively*) have notable domain gap with the major part of the **Q-Instruct** dataset (*mostly in-the-wild photographs*), the *+0.243* average SRCC gain on them demonstrates that the *low-level instruction tuning* can broadly improve low-level perception abilities of MLLMs.

6.4. Ablation Studies

Despite the main results for *low-level visual instruction tuning*, we also compare among several data variations during tuning on LLaVA-v1.5 (*7B*), analyzed as follows.

#1: Effects of scaling up the Q-Instruct. The first group of variations discusses the effects of data amount during *low-level visual instruction tuning*. As illustrated in Fig. 6, under either ***mix*** or ***after*** strategy, scaling up the **Q-Instruct**

Table 6. Comparison of the **Quality Assessment (A3)** ability between baseline MLLMs and **Q-Instruct**-*tuned* versions, where *"Mostly Seen"* datasets denote those with the majority of their images sampled in the Q-Instruct, and *"Barely Seen"* represent those with only a small proportion (<20%) sampled. The *"Never Seen"* datasets have **zero** overlap with the **Q-Instruct**. Metrics are SRCC / PLCC.

Dataset Group		*Mostly Seen*		*Barely Seen*			*Never Seen*		
% of dataset seen during training		48.92%	95.26%	2.00%	17.11%	13.41%	**0%**	**0%**	**0%**
Model (*variant*)	*Q-Instruct* Strategy	*KonIQ-10k*	*SPAQ*	*LIVE-FB*	*LIVE-itw*	*AGIQA-3K*	*CGIQA-6K*	*KADID-10K*	*KonViD-1k*
NIQE	–	0.316 / 0.377	0.693 / 0.669	0.211 / 0.288	0.480 / 0.451	0.562 / 0.517	0.075 / 0.056	0.374 / 0.428	0.541 / 0.553
LLaVA-v1.5 (*7B*)	*no* (Baseline)	0.463 / 0.459	0.443 / 0.467	0.310 / 0.339	0.445 / 0.481	0.664 / 0.754	0.285 / 0.297	0.390 / 0.400	0.461 / 0.495
	(a) ***mix*** w/ high-level	0.809 / 0.852	0.880 / 0.883	0.377 / 0.436	0.800 / 0.806	0.724 / 0.828	0.521 / 0.535	0.688 / 0.695	0.766 / 0.717
	(b) ***after*** high-level	0.793 / 0.850	0.887 / 0.888	0.385 / 0.447	0.805 / 0.810	0.729 / 0.830	0.501 / 0.524	0.695 / 0.702	**0.780 / 0.731**
LLaVA-v1.5 (*13B*)	*no* (Baseline)	0.471 / 0.541	0.563 / 0.584	0.305 / 0.321	0.344 / 0.358	0.672 / 0.738	0.321 / 0.333	0.417 / 0.440	0.518 / 0.577
	(a) ***mix*** w/ high-level	0.732 / 0.787	0.858 / 0.848	0.371 / 0.463	0.629 / 0.701	0.709 / 0.814	0.471 / 0.488	0.627 / 0.626	0.720 / 0.733
	(b) ***after*** high-level	0.748 / 0.798	0.867 / 0.869	0.359 / 0.417	0.695 / 0.719	0.696 / 0.766	0.494 / 0.516	0.633 / 0.641	0.706 / 0.692
mPLUG-Owl-2	*no* (Baseline)	0.196 / 0.252	0.589 / 0.614	0.217 / 0.286	0.293 / 0.342	0.473 / 0.492	-0.024 / -0.032	0.541 / 0.546	0.409 / 0.442
	(a) ***mix*** w/ high-level	0.899 / 0.916	0.899 / **0.903**	0.432 / 0.545	0.829 / 0.822	0.743 / 0.806	**0.624 / 0.636**	0.698 / 0.676	0.693 / 0.663
	(b) ***after*** high-level	**0.911 / 0.921**	0.901 / 0.898	0.442 / **0.535**	**0.842 / 0.840**	0.700 / 0.763	0.572 / 0.578	0.682 / 0.683	0.769 / 0.721
InternLM-XComposer-VL	*no* (Baseline)	0.568 / 0.616	0.731 / 0.751	0.358 / 0.413	0.619 / 0.678	0.734 / 0.777	0.246 / 0.268	0.540 / 0.563	0.620 / 0.649
	(a) ***mix*** w/ high-level	0.874 / 0.892	**0.909** / 0.897	0.442 / 0.518	0.820 / 0.811	**0.785** / 0.830	0.391 / 0.411	**0.706 / 0.710**	0.739 / 0.702
	(b) ***after*** high-level	0.816 / 0.858	0.879 / 0.884	**0.443** / 0.510	0.771 / 0.801	0.772 / **0.847**	0.394 / 0.420	0.677 / 0.645	0.743 / 0.730
Average Improvement		*+0.398/+0.392*	*+0.304/+0.280*	*+0.108/+0.144*	*+0.349/+0.324*	*+0.097/+0.120*	*+0.289/+0.297*	*+0.204/+0.185*	*+0.238/+0.170*

Table 7. Comparison on low-level **Description** ability between *full* **Q-Instruct** and *only* **Q-Pathway** as low-level training dataset.

Q-Instruct Strategy	low-level dataset	*completeness*	*precision*	*relevance*	***sum***
no (Baseline)	None	0.90	1.13	1.18	3.21
(a) ***mix*** w/ high-level	*only* **Q-Pathway**	1.07	1.13	1.54	3.74
	full **Q-Instruct**	**1.12**	**1.17**	**1.57**	**3.86**
(b) ***after*** high-level	*only* **Q-Pathway**	1.02	1.12	**1.55**	3.69
	full **Q-Instruct**	**1.11**	**1.16**	1.54	**3.82**

Table 8. Comparison on low-level **Perception** ability (*test set*) between training with *full* **Q-Instruct** dataset and *only* VQA subset.

Q-Instruct Strategy	low-level dataset	*Yes-or-No*	*What*	*How*	*Overall*
no (Baseline)	None	64.6%	59.2%	55.8%	60.1%
(a) ***mix*** w/ high-level	*only* VQA subset	78.1%	61.5%	61.5%	67.6%
	full **Q-Instruct**	**78.7%**	**64.0%**	**63.8%**	**69.3%**
(b) ***after*** high-level	*only* VQA subset	77.9%	61.8%	56.8%	66.1%
	full **Q-Instruct**	**78.5%**	**63.3%**	**58.9%**	**67.4%**

during training can continuously improve the low-level perceptual accuracy. Moreover, the results suggest that the performance of MLLMs is still not saturated even with the current 200K data scale, encouraging us to further unleash their vast underlying power on tackling low-level visual tasks.

#2: Effects of joint training. In the *low-level visual instruction tuning*, we combine different subsets together and train them jointly under one unified model. To validate its effectiveness, we compare this approach with traditional task-separate tuning, on both low-level description (Tab. 7) and question-answering (Tab. 8) capabilities. Both experiments indicate that a joint learning scheme can improve the accuracy on these abilities, especially when low-level data is independently used during tuning. While the different subsets in the **Q-Instruct** come from the same original human feedbacks, the improvement is cost-efficient, and inspires further explorations for *low-level visual instruction tuning* to expand to even more tasks, so as to further improve the low-level capabilities of these MLLMs.

#3: Effects of high-level awareness. While we notice generally on par abilities between the ***mix*** strategy and the ***after*** strategy, we further investigate the performance if we ***replace*** the second stage datasets into the **Q-Instruct**, while no high-level instruction tuning datasets are involved during training. As compared in Tab. 9, the "*replace*" strategy is notably worse than the two adopted strategies in Sec. 5, suggesting that fundamental high-level awareness is important on general low-level visual recognition for MLLMs.

Table 9. Comparison between the proposed two strategies (as in Sec. 5, and another variant that ***replaces*** high-level tuning into the low-level tuning, on their low-level **Perception** ability (*test set*).

Q-Instruct Strategy	*Yes-or-No*	*What*	*How*	*Overall*
no (Baseline)	64.6%	59.2%	55.8%	60.1%
replace high-level (*not adopted*)	75.0%	59.4%	56.4%	64.1%
mix with high-level (*ours*, strategy (a))	**78.7%**	**64.0%**	**63.8%**	**69.3%**
after high-level (*ours*, strategy (b))	**78.5%**	**63.3%**	**58.9%**	**67.4%**

7. Conclusion

Our work proposes the first-of-a-kind multi-modal datasets on low-level visual aspects, including the **Q-Pathway** with **58K** human *text* feedbacks, and the derived **Q-Instruct** with **200K** instruction-response pairs, to facilitate *low-level visual instruction tuning* for MLLMs. They allow MLLMs to significantly improve their question-answering accuracy related to low-level visual perception, and showcase the potential for providing more reliable low-level descriptions for images and eventually relieving human burdens on this task. Further, their IQA performance reveals an intriguing phenomenon, that pure *text-driven* instruction tuning can sufficiently align MLLMs with numerical quality scores, with impressive generalization on unseen types of visual inputs. In summary, our work has advanced a solid step forward on improving the low-level visual abilities of MLLMs, and we hope that our progress and insights can encourage future explorations towards an eventual goal that foundation models understand the low-level visual world like a human.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 3, 5

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2, 3, 6

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3

[4] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment, 2023. 3

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 3, 4, 5, 6

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. 3

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 3, 6

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2022. 3

[9] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, 2020. 2, 3, 4

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2023. 3

[11] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1, 3

[12] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE*, 25(1):372–387, 2016. 3, 4

[13] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *QoMEX*, pages 1–6, 2017. 7

[14] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP*, 29:4041–4056, 2020. 2, 3, 4

[15] Jingwen Hou, Weisi Lin, Yuming Fang, Haoning Wu, Chaofeng Chen, Liang Liao, and Weide Liu. Towards transparent deep image aesthetics assessment with tag-based content descriptors. *IEEE TIP*, 2023. 3

[16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[17] Dinesh Jayaraman, Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. Objective quality assessment of multiply distorted images. In *ASILOMAR*, pages 1693–1697, 2012. 3

[18] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 3

[19] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining, 2023. 3

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. 5

[21] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 3

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, 2017. 2

[23] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 3

[24] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 3, 6

[25] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants, 2023. 1

[26] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment, 2023. 3, 4

[27] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE TMM*, 21(5):1221–1234, 2019. 2

[28] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *QoMEX*, pages 1–3, 2019. 3, 7

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1, 3, 6

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 3

[31] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2023. 2, 3

[32] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 6

[33] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 4

[34] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 3

[35] Layla Oesper, Daniele Merico, Ruth Isserlin, and Gary D Bader. Wordcloud: a cytoscape plugin to create a visual semantic summary of networks. *Source code for biology and medicine*, 6(1):7, 2011. 5

[36] OpenAI. Chatgpt (june 13 version), 2023. Large language model. 2, 5

[37] OpenAI. Gpt-4 technical report, 2023. 1, 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3

[39] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 6

[40] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*, 2022. 2, 3, 5

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 6

[42] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, 2020. 3

[43] Shaolin Su, Vlad Hosu, Hanhe Lin, Yanning Zhang, and Dietmar 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)*, pages 1–12, 2021. 2, 3

[44] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM, 2023. 6

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 3

[46] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 6

[47] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images, 2022. 3

[48] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *CVPR*, pages 13435–13444, 2021. 2

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022. 5

[50] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *ECCV*, 2022. 3

[51] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE TPAMI*, 45(12):15185–15202, 2023. 2, 4

[52] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring opinion-unaware video quality assessment with semantic affinity criterion. In *International Conference on Multimedia and Expo (ICME)*, 2023. 3

[53] Haoning Wu, Liang Liao, Annan Wang, Chaofeng Chen, Jingwen Hou Hou, Erli Zhang, Wenxiu Sun Sun, Qiong Yan,

and Weisi Lin. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment, 2023.

[54] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 2, 3

[55] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable video quality assessment: A database and a language-prompted approach. In *ACM MM*, 2023. 2, 3, 4

[56] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision, 2023. 2, 3, 5, 6, 7

[57] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. 3, 4

[58] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023. 6

[59] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *CVPR*, 2020. 2, 3, 4

[60] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3, 5

[61] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. 2, 6

[62] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023. 3, 6

[63] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE TCSVT*, 30(1):36–47, 2020. 3

[64] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[65] Zicheng Zhang, Wei Sun, Tao Wang, Wei Lu, Quan Zhou, Qiyuan Wang, Xiongkuo Min, Guangtao Zhai, et al. Subjective and objective quality assessment for in-the-wild computer graphics images. *arXiv preprint arXiv:2303.08050*, 2023. 7

[66] Zicheng Zhang, Wei Sun, Yingjie Zhou, Haoning Wu, Chunyi Li, Xiongkuo Min, and Xiaohong Liu. Advancing zero-shot digital human quality assessment through text-prompted evaluation, 2023. 3

[67] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 6