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

Supplementary Material

A. Details for Data Collection

A.1. Interface for Subjective Experiments

The interface for the subjective experiments is built upon Gradio 3.34.0, set up locally on Ubuntu 20.04 workstations. All participants need to record their ID and write down their *pathway* feedbacks for a given image. The MOS for the image and possible low-level attributes are listed as reference. A screenshot of the interface is shown in Fig. 1.

A.2. Prompts for Building Q-Instruct with GPT

***What/How* questions.** *Generate multiple question and answer pairs based on the following description of an image quality. The questions can start with "What/Why/How". The answer should be concise and only contain the core information with minimum words. You should also generate several false answers for each question under the key of "false candidates", which are also reasonable given the question by contradicts with the description. Organize the output a list in JSON format and when you respond, please only output the json, no other words are needed:*
Description: $DESC

***Yes/No* questions.** *Generate multiple yes-or-no question and answer pairs based on the following description of an image quality. The answer should be concise and only contain "Yes" or "No". The number of questions with the answer "Yes" should be close to the number of questions with the answer "No". You can also ask questions about quality issues that are not mentioned in the analysis. The answer for those unsure questions should be "No". Organize the output a list in JSON format and when you respond, please only output the json, no other words are needed:*
Description: $DESC

***Extended* conversations.** *Generate conversations based on the following description of quality and other low-level visual attributes of an image. These conversations can include one of the aspects in the folow 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. Recommending the image to respective consumers; 5. Other conversations that may happen given the descriptions. Remember to be relevant to the image. Organize the output a list in JSON format (interleaved with "query" and "response" keys for each conversation) and when you respond, please only output the json, no other words are needed:*
Description: $DESC

B. Hyper-parameters during Training

Hyper-parameters for LLaVA-v1.5. The *low-level visual instruction tuning* for LLaVA-v1.5 (7B/13B) is conducted with 8 NVIDIA A100-SMX4-80GB GPU (*requiring 16 hours for 7B, 22 hours for 13B*, for the ***mix*** version). We record all hyper-parameters in Tab. 1.

Hyper-parameter	***mix*** with high-level	***after*** high-level
ViT init.	CLIP-L/14-336 [3]	
LLM init.	Vicuna-v1.5 [8]	LLaVA-v1.5
image resolution	336 × 336	336 × 336
group modality length	True	False
batch size	128	
lr max	2e-5	
lr schedule	cosine decay	
warmup epochs	0.03	
weight decay	0	
gradient acc.	1	
numerical precision	`bfloat16`	
epoch	1	
optimizer	AdamW	
optimizer sharding	✓	
activation checkpointing	✓	
deepspeed stage	3	

Table 1. **Hyper-parameters** of *low-level visual instruction tuning* on LLaVA-v1.5 (7B/13B), the same as original LLaVA-v1.5.

Hyper-parameters for mPLUG-Owl-2. The *low-level visual instruction tuning* for mPLUG-Owl-2 is conducted with 32 NVIDIA A100-SMX4-80GB GPU (requiring *8 hours* for the ***mix*** version). Hyper-parameters in Tab. 2.

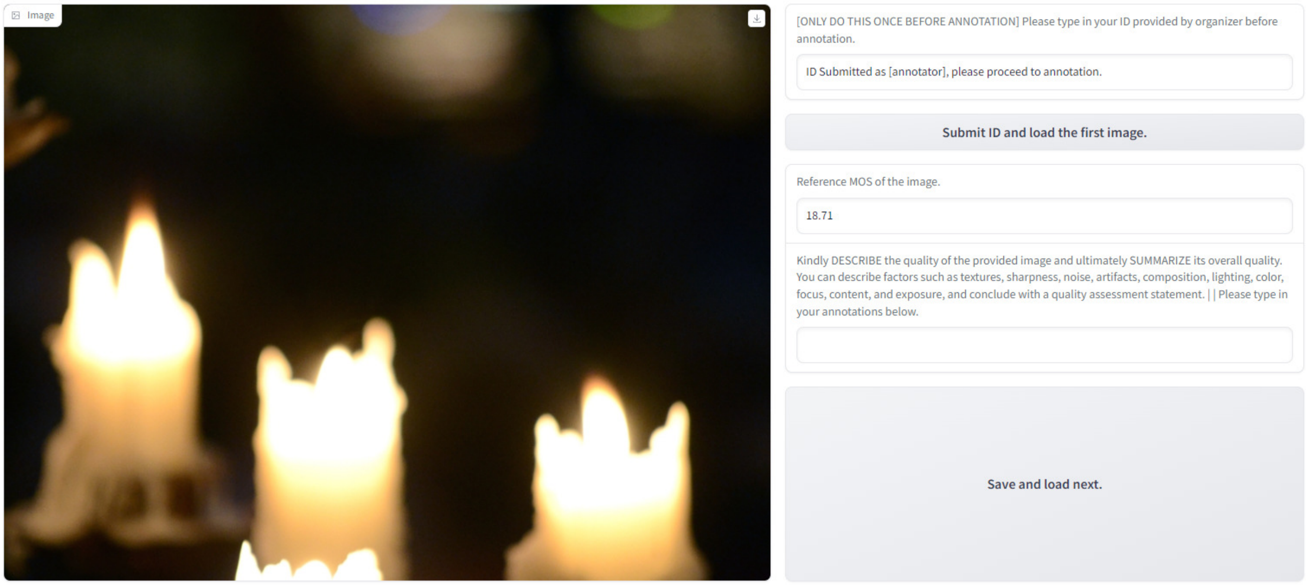
Hyper-parameters for InternLM-XComposer-VL. Similar as mPLUG-Owl-2, the *low-level visual instruction tuning* for InternLM-XComposer-VL is conducted with 32 NVIDIA A100-SMX4-80GB GPU (requiring *13 hours* for the ***mix*** version). Hyper-parameters are listed in Tab. 3.

C. Evaluation Details

C.1. Prompt Settings on (A1) Perception (*via* MCQ)

Denote the image tokens as `<image>`, the question as `<QUESTION>`, choices as <CHOICE$_i$>, the prompt settings for different models on answering Multi-Choice Questions (MCQ) are slightly different, listed as follows. To ensure optimal results, during training, we also transform the VQA subset under the same settings, respectively.

Prompt Settings for LLaVA-v1.5 (7B/13B). *A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite an-*



[ONLY DO THIS ONCE BEFORE ANNOTATION] Please type in your ID provided by organizer before annotation.

ID Submitted as [annotator], please proceed to annotation.

Reference MOS of the image.

18.71

Kindly DESCRIBE the quality of the provided image and ultimately SUMMARIZE its overall quality. You can describe factors such as textures, sharpness, noise, artifacts, composition, lighting, color, focus, content, and exposure, and conclude with a quality assessment statement. | | Please type in your annotations below.

Figure 1. The gradio interface for subjects to provide ***pathway*** feedbacks. While the quality scores (MOS) of images are available, these scores will be provided to the subjects as a reference, allowing the feedbacks to truly become explanations of these quality scores.

Hyper-parameter	***mix*** with high-level	***after*** high-level
ViT init.	Pre-train stage (updated CLIP-L/14 [3])	
LLM init.	LLaMA-2 [5]	
visual abstractor init.	Pre-train stage	mPLUG-Owl-2
image resolution	448 × 448	448 × 448
batch size	256	
lr max	2e-5	
lr schedule	cosine decay	
lr warmup ratio	0.03	
weight decay	0	
gradient acc.	16	
numerical precision	`bfloat16`	
epoch	1	
warm-up steps	250	
optimizer	AdamW	
optimizer sharding	✓	
activation checkpointing	✓	
model parallelism	2	
pipeline parallelism	1	

Table 2. **Hyper-parameters** of *low-level visual instruction tuning* on mPLUG-Owl-2, the same as the original model.

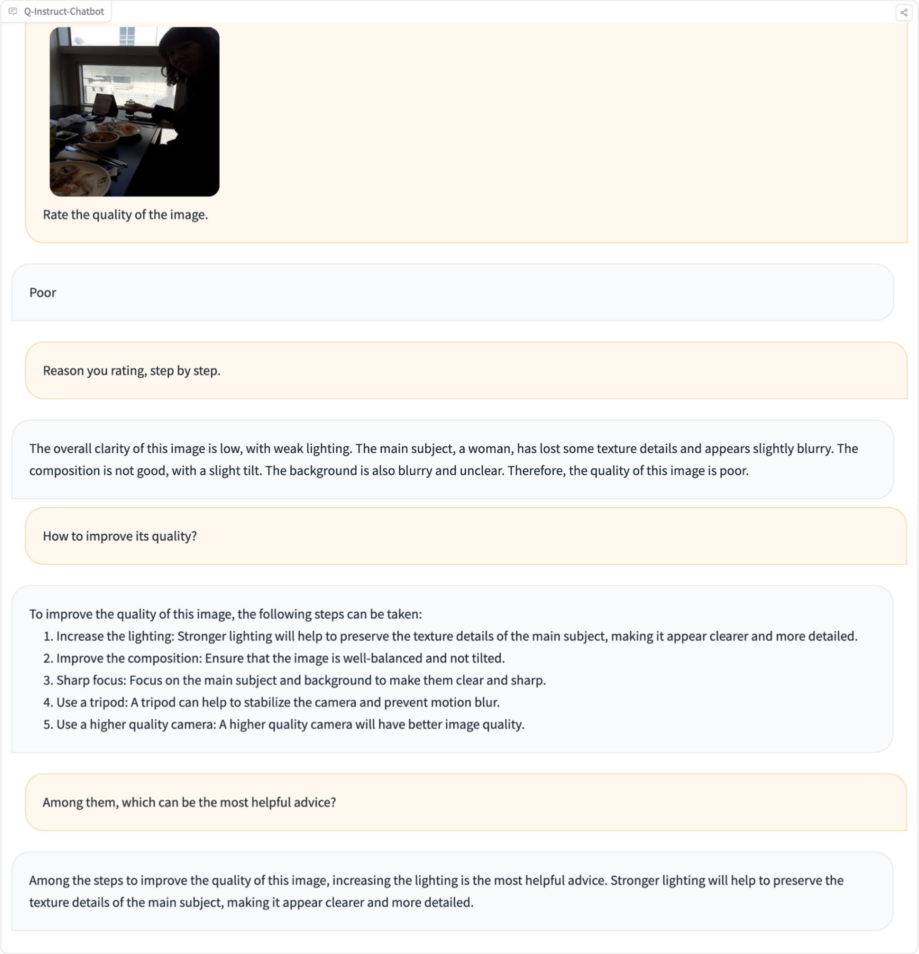
Hyper-parameter	***mix*** with high-level	***after*** high-level
ViT init.	EVA-CLIP-G [4]	
LLM init.	Pre-train stage	InternLM-XComposer-VL
perceive sampler init.	Pre-train stage	InternLM-XComposer-VL
image resolution	224 × 224	224 × 224
batch size	256	
lr max	5e-5	
lr schedule	cosine decay	
lr warmup ratio	0.05	
weight decay	0	
gradient acc.	1	
numerical precision	`float16`	
epoch	1	
warm-up steps	250	
optimizer	AdamW	
special setting	low-rank adaptation (*LORA*)	
activation checkpointing	✓	

Table 3. **Hyper-parameters** of *low-level visual instruction tuning* on InternLM-XComposer-VL, the same as the original model.

swers to the human's questions. USER:`<image>`
`<QUESTION>`
Answer with the option's letter from the given choices directly.
A. `<CHOICE`$_A$`>`
B. `<CHOICE`$_B$`>`
C. `<CHOICE`$_C$`>`
ASSISTANT:

Prompt Settings for mPLUG-Owl-2. *USER:* `<image>`
`<QUESTION>`
Answer with the option's letter from the given choices directly.
A. `<CHOICE`$_A$`>`
B. `<CHOICE`$_B$`>`
C. `<CHOICE`$_C$`>`
ASSISTANT:

Prompt Settings for InternLM-XComposer-VL.
`<|User|>`: `<image>`*Please answer this question by choosing the correct choice.Context: N/A*
Question: `<QUESTION>`
Options: A. `<CHOICE`$_A$`>`
B. `<CHOICE`$_B$`>`
C. `<CHOICE`$_C$`>`
`<TOKENS_UNUSED_0>` `<|Bot|>`: *Answer: The answer is*



Rate the quality of the image.

Poor

Reason you rating, step by step.

The overall clarity of this image is low, with weak lighting. The main subject, a woman, has lost some texture details and appears slightly blurry. The composition is not good, with a slight tilt. The background is also blurry and unclear. Therefore, the quality of this image is poor.

How to improve its quality?

To improve the quality of this image, the following steps can be taken:

1. Increase the lighting: Stronger lighting will help to preserve the texture details of the main subject, making it appear clearer and more detailed.
2. Improve the composition: Ensure that the image is well-balanced and not tilted.
3. Sharp focus: Focus on the main subject and background to make them clear and sharp.
4. Use a tripod: A tripod can help to stabilize the camera and prevent motion blur.
5. Use a higher quality camera: A higher quality camera will have better image quality.

Among them, which can be the most helpful advice?

Among the steps to improve the quality of this image, increasing the lighting is the most helpful advice. Stronger lighting will help to preserve the texture details of the main subject, making it appear clearer and more detailed.

Figure 2. **Qualitative Analysis (I)**: A multi-turn conversation that the user subsequently queries the **Q-Instruct**-*tuned* MLLM on (1) *rating image quality*, (2) *reasoning the rating*, (3) *providing improvement suggestions*, and (4) *discerning the most important suggestion*.

C.2. Prompt Setting on (A2) Description

For the **(A2) Description** task, we unify all models under the same prompt: *"Describe and evaluate the quality of the image."*, as this is the only prompt that can effectively allow every base model to describe low-level visual attributes and then evaluate image quality. For the alternate prompt as shown in Fig. 1 (main paper), *"Rate the quality of the image. Think step by step."*, the base InternLM-XComposer-VL only provides numbers (*1/2/3/4/5*) without explanations or reasonings. Therefore, we choose the current prompt to evaluate the description ability among all variants.

C.3. Prompt Setting on (A3) Assessment

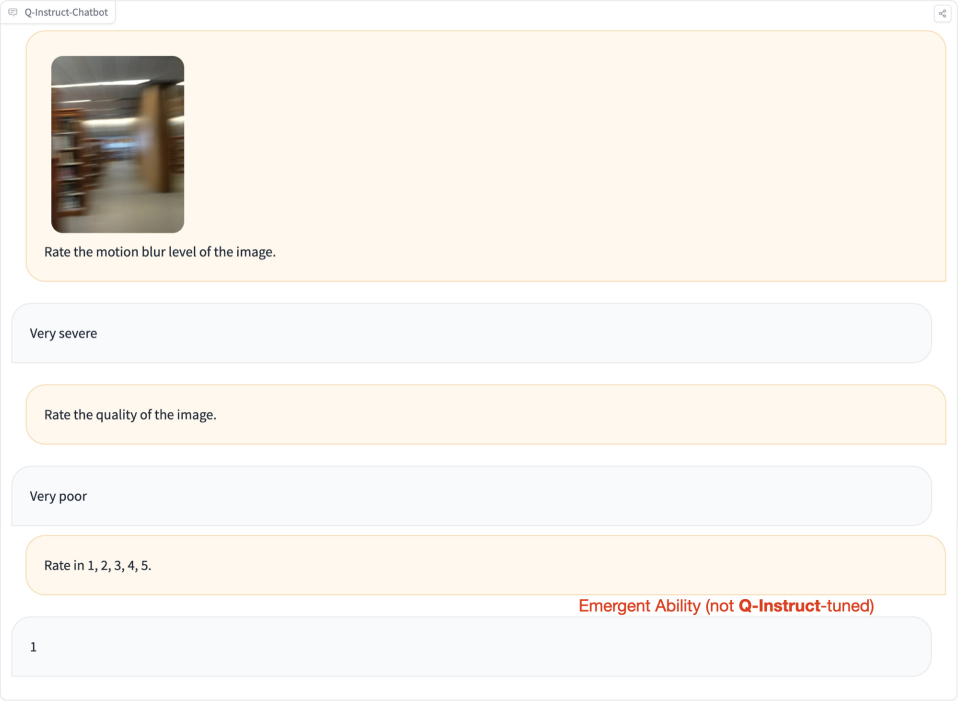
For the **(A3) Quality Assessment** task, we follow the strategy as proposed by Q-Bench [6], with the `softmax` output between *good* and *poor* to collect better *quantifiable* scores for images, under the first output token of MLLMs:

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

For KoNViD-1k, the video quality assessment dataset as evaluated, we sample *one frame per second* (the same as NIQE [2]), and average the image quality scores as calculated by Eq. 1 among frames. While we acknowledge that this is not the optimal strategy for evaluating video quality, the excellent results in main paper already significantly outperform all methods that are not trained for video quality assessment, suggesting that if better strategies are adopted, MLLMs can do even better on video quality assessment.

D. Qualitative Analyses

Considering that different base MLLMs reach similar performance after *low-level visual instruction tuning*, we build a publicly-accessible demo for **Q-Instruct**-mPLUG-Owl-2 under the ***mix*** strategy, and invite our team members to really *chat* with it. These conversations exhibit their excellent abilities on various scenarios, including multi-turn queries (Fig. 2), multi-purpose assessment (Fig. 3), correction (Fig. 4), and local in-context perception (Fig. 5).



Rate the motion blur level of the image.

Very severe

Rate the quality of the image.

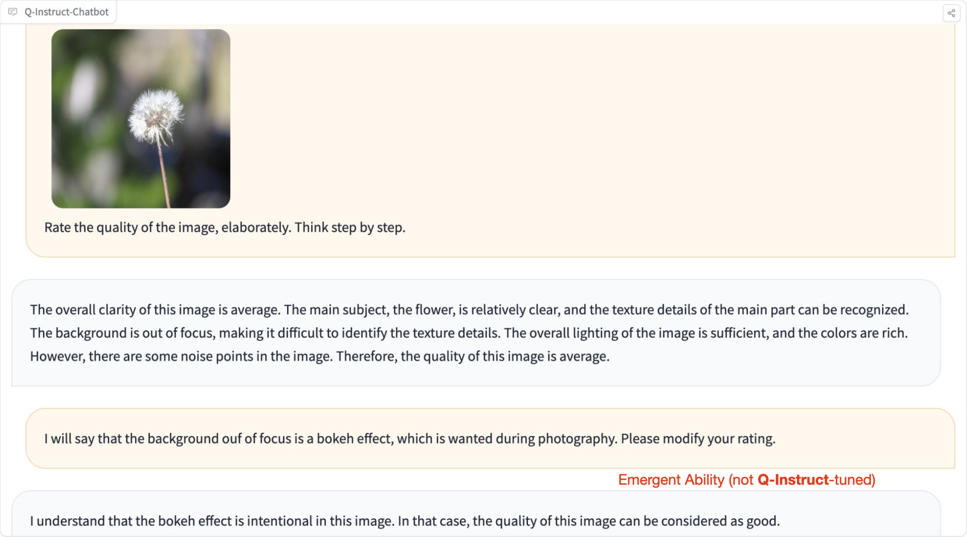
Very poor

Rate in 1, 2, 3, 4, 5.

Emergent Ability (not **Q-Instruct**-tuned)

1

Figure 3. **Qualitative Analysis (II)**: The **Q-Instruct**-*tuned* MLLM can serve as a multi-purpose (*overall quality* or *specific distortion*) and multi-format (*text, good/average/poor* or *numerical, e.g. 1/2/3/4/5*) quality evaluator.



Rate the quality of the image, elaborately. Think step by step.

The overall clarity of this image is average. The main subject, the flower, is relatively clear, and the texture details of the main part can be recognized. The background is out of focus, making it difficult to identify the texture details. The overall lighting of the image is sufficient, and the colors are rich. However, there are some noise points in the image. Therefore, the quality of this image is average.

I will say that the background ouf of focus is a bokeh effect, which is wanted during photography. Please modify your rating.

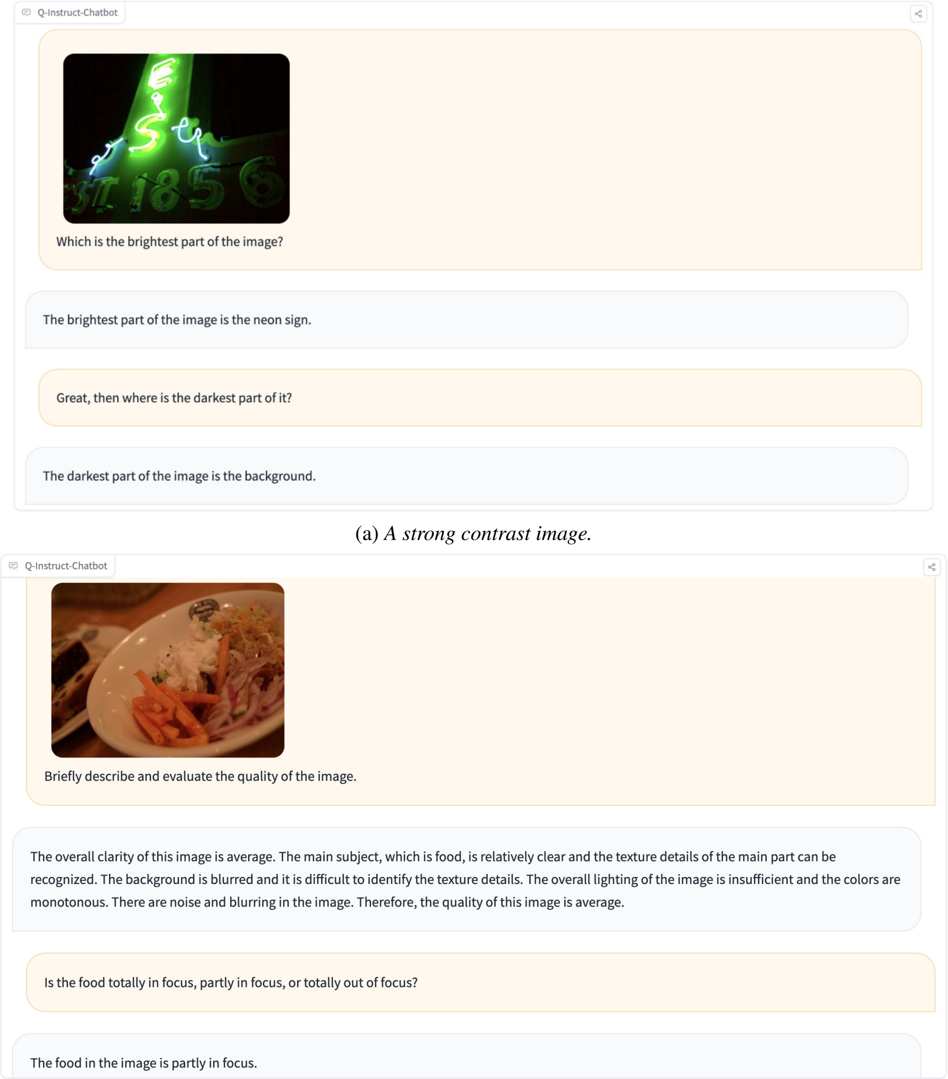
Emergent Ability (not **Q-Instruct**-tuned)

I understand that the bokeh effect is intentional in this image. In that case, the quality of this image can be considered as good.

Figure 4. **Qualitative Analysis (III)**: The **Q-Instruct**-*tuned* MLLM can correct itself based on further instructions. While provided with additional context (*i.e.* background bokeh is intentional), it can modify its rating to align with the context.

E. Limitations

The known limitations of our studies are as follows. First, though with improved quality assessment and low-level visual perception abilities, the **Q-Instruct**-*tuned* models have witnessed declined performance on general-purpose tasks, especially language-centric tasks, or tasks that require heavy reasoning abilities. Therefore, they may produce unwanted outputs if applied to tasks other than low-level visual perception and understanding. Second, though with improved accuracy, the **Q-Instruct**-*tuned* models still perform worse (68%-71% accuracy on LLVisionQA-*test*) than an average human (about 74%), and may not yet be able to directly replace human on low-level related tasks. Thirdly, the **Q-Instruct** dataset mainly consists of natural in-the-wild images. Though they prove excellent generalization on other types of visual contents, the performance might still be improvable if further tuned on these datasets.



Which is the brightest part of the image?

The brightest part of the image is the neon sign.

Great, then where is the darkest part of it?

The darkest part of the image is the background.

(a) *A strong contrast image.*

Briefly describe and evaluate the quality of the image.

The overall clarity of this image is average. The main subject, which is food, is relatively clear and the texture details of the main part can be recognized. The background is blurred and it is difficult to identify the texture details. The overall lighting of the image is insufficient and the colors are monotonous. There are noise and blurring in the image. Therefore, the quality of this image is average.

Is the food totally in focus, partly in focus, or totally out of focus?

The food in the image is partly in focus.

(b) *A partly in-focus image.*

Figure 5. **Qualitative Analysis (IV)**: Local in-context low-level perceptual abilities of **Q-Instruct**-*tuned* MLLMs. They can effectively discern the bright part and dark part in a *strong contrast image* (a), or the clarity of different objects/areas in a *partly in-focus image* (b).

F. Ethical Acknowledgements

In our study, all participants were fully informed about the nature and amount of the tasks involved prior to their participation. No uncomfortable content was reported during this process. We express our gratitude to the participants for their valuable contributions, which were essential to the success of our research. We commit to upholding all ethical standards to ensure the well-being of our participants, as well as the integrity of our research findings.

G. Acknowledgements

Our team would like to sincerely thank Qinghao Ye and Xiaoyi Dong for providing pre-trained weights of mPLUG-Owl-2 and InternLM-XComposer-VL after the *low-level visual instruction tuning*, including a complete training recipe on their respective high-level datasets mixed with the **Q-Instruct**. We believe these weights will significantly contribute to the open-source community working on tasks related to low-level visual perception and understanding.

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

H. License

Researchers and open-source developers are free to use the **Q-Instruct** dataset and the fine-tuned weights as provided for the four MLLMs. We also allow commercial use, while any commercial use should be pre-permitted by our team. Any usage should also comply with licenses of the original base models (*inc.* base LLMs such as Vicuna, LLaMA-2).

I. Results on General-Purpose Benchmarks

Though our main aim is to provide first-of-its-kind specialized language assistant for low-level vision tasks and have not optimized for the general benchmarks, our tuning still retains decent general ability. Take InternLM-XComposer-VL [7] (***mix***) as an example, on MMBench-test [1], after Q-Instruct tuning, it slightly drops from 74.4% to 71.4%, from 3rd (among 31) to 7th, still more competitive than LLaVA-v1.5-13B (67.8%) and Qwen-VL-Plus (67%).

References

[1] 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. 6

[2] 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

[3] 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. 1, 2

[4] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. 2

[5] 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. 2

[6] 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. 3

[7] 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. 6

[8] 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. 1