

Towards Open-ended Visual Quality Comparison Supplementary Material

Haoning Wu^{*1}, Hanwei Zhu^{*2}, Zicheng Zhang^{*3}, Erli Zhang¹
Chaofeng Chen¹, Liang Liao¹, Chunyi Li³, Annan Wang¹, Wenxiu Sun⁴
Qiong Yan⁴, Xiaohong Liu³, Guangtao Zhai³, Shiqi Wang², and Weisi Lin¹

¹ Nanyang Technological University

² City University of Hong Kong

³ Shanghai Jiao Tong University

⁴ SenseTime Research

1 Construction Details of Co-Instruct-562K

1.1 Prompt Template for Merge2Compare

Denote the description for image $\langle \text{img}_i \rangle$ as $\langle \text{desc}_i \rangle$, the user query (*i.e.*, text input) for the Mixtral-8x7B [1] *merging* is formulated as follows:

1. The prompts for two images:

The first image: $\langle \text{desc}_0 \rangle$ The second image: $\langle \text{desc}_1 \rangle$
Which image has better quality, and why?

2. The prompts for three/four images:

{The K_{i+1} image: $\langle \text{desc}_i \rangle$ $\big|_{i=0}^{N-1}$ }

Please rank the quality of the images and justify your rankings.
where K_i represents the ordinal form of $i+1$, *e.g.* K_1 is first, K_2 is second.

1.2 Prompt Template for Teach2Compare

Teach2Compare-General. The prompts for GPT-4V to generate general quality comparison are listed as follows. To improve diversity, one random prompt between the two is chosen to query GPT-4V.

1. The prompts for two images:

- "Which image has better quality among the two images, and why? Judge and discuss concisely. Please directly provide your judgement and discussion."
- "Compare the quality-related attributes and overall quality between the two images briefly. Your reference can be related to specific objects in the two images. Be short and precise."

2. The prompts for three images:

* These authors contribute equally to this paper.

- "Which image has better quality among the three images, and why? Judge and discuss concisely. Please directly provide your judgement and discussion."
- "Compare the quality-related attributes and overall quality among the three images briefly. Your reference can be related to specific objects in the three images. Be short and precise."

3. The prompts for four images:

- "Which image has better quality among the four images, and why? Judge and discuss concisely. Please directly provide your judgement and discussion."
- "Compare the quality-related attributes and overall quality among the four images briefly. Your reference can be related to specific objects in the three images. Be short and precise."

Teach2Compare-Q&A. The prompts for GPT-4V to generate questions, answers, and false candidates are listed as follows. The responses of GPT-4V are cleaned and post-processed into the Teach2Compare-Q&A.

1. The prompts of two images:

```
"Please generate some yes/no, what, how, and which questions in the multi-choice question (MCQ) format and corresponding responses on comparing the quality between the first image and the second image in the JSON format. Your reply should only contain the dumped list that can be directly loaded via 'json.loads'. Please pay more attention to the format of the JSON string. [{
  "question": str
  "options": List[str]
  "correct_option": str (should be one element of the options)
},] Generate 10 MCQs if you can."
```

2. The prompts of three images:

```
"Please generate some yes/no, what, how, and which questions in the multi-choice question (MCQ) format and corresponding responses on comparing the quality among the first, second, and third image in the JSON format. Your reply should only contain the dumped list that can be directly loaded via 'json.loads'. Please pay more attention to the format of the JSON string. [{
  "question": str
  "options": List[str]
  "correct_option": str (should be one element of the options)
},] Generate 10 MCQs if you can."
```

3. The prompts of four images:

"Please generate some yes/no, what, how, and which questions in the multi-choice question (MCQ) format and corresponding responses on comparing the quality among the first, second, third, and fourth image in the JSON format. Your reply should only contain the dumped list that can be directly loaded via 'json.loads'.

Please pay more attention to the format of the JSON string. [{"question": str, "options": List[str], "correct_option": str (should be one element of the options)},] Generate 10 MCQs if you can."

2 Implementation Details of Co-Instruct

2.1 Hyper-parameters

The total **Co-Instruct** consists of 8.2B parameters, which are fully updated during the training process. The hyper-parameters for training are listed as follows.

Hyper-parameter	Co-Instruct
ViT init.	Pre-train stage (updated CLIP-L/14 [3])
LLM init.	LLaMA-2 [4]
LMM init.	mPLUG-Owl2
image resolution	448 × 448
batch size	192
lr max	2e-5
lr schedule	cosine decay
lr warmup ratio	0.03
weight decay	0
gradient acc.	2
numerical precision	bfloat16
epoch	1
warm-up epochs	0.03
optimizer	AdamW
optimizer sharding	✓
activation checkpointing	✓
model parallelism	2
pipeline parallelism	1

Table 1: Hyper-parameters for training Co-Instruct.

The final checkpoint after epoch 2 (final epoch) is adopted for evaluation.

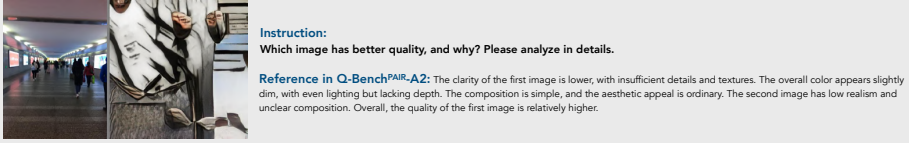
2.2 Training Code

The Co-Instruct-562K can be seamlessly trained with the Q-Align codebase (<https://github.com/Q-Future/Q-Align>) under the previous hyper-parameter settings, with the mPLUG-Owl2 format JSON data (*replacing all <image> into <|image|> from the LLaVA-format data*).

3 Further Analysis

3.1 Further Analysis on $Q\text{-Bench}^{\text{PAIR}}\text{-A2}$

In the main paper, we show that Co-Instruct still falls a little bit behind GPT-4V on the quantitative metrics of $Q\text{-Bench}^{\text{PAIR}}\text{-A2}$. This might be because outputs from GPT-4V are more than *twice as long* as **Co-Instruct** outputs, while the GPT evaluation used in this evaluation setting is observed [2] to be in favor of longer text outputs. To further analyze this potential bias, we qualitatively visualize the result of different LMMs in Fig. 1. As shown in the figure, the baseline open-source LMM even confuses the information from the two images, and Gemini-Pro makes rather poor detailed reasonings. For GPT-4V, it generates the *longest outputs* among all LMMs, which might be the reason that it gets a relatively high precision score even its outputs are not totally correct.



Instruction:
Which image has better quality, and why? Please analyze in details.

Reference in $Q\text{-Bench}^{\text{PAIR}}\text{-A2}$: The clarity of the first image is lower, with insufficient details and textures. The overall color appears slightly dim, with even lighting but lacking depth. The composition is simple, and the aesthetic appeal is ordinary. The second image has low realism and unclear composition. Overall, the quality of the first image is relatively higher.

mPLUG-Owl2 (baseline) Precision Score: 0.6/2
The first image is a black and white photo, while the second image is a colorful photo. The second image has a more vibrant and detailed appearance, with a focus on the people walking through the station.

<p>Co-Instruct (Ours) Precision Score: 1.2/2</p> <p>The first image has better quality than the second image. Although the first image has some issues such as low clarity, weak lighting, and poor composition, it still retains some texture details on the ground and has a clearer main subject compared to the second image. The second image, on the other hand, has poor clarity, making it difficult to distinguish the subject, and the content is chaotic and hard to make out. Therefore, the first image is of better quality than the second image.</p>	<p>Gemini-Pro (proprietary) Precision Score: 1.0/2</p> <p>The first image is a photo of a crowded underground walkway. The second image is a painting that was generated by an artificial intelligence (AI). The painting is based on the photo, but it is not a direct copy. The AI has added its own creative elements to the painting, such as the addition of a few extra people and the changing of the colors of some of the people's clothing. In terms of quality, the first image is a higher quality image than the second image. The first image is more detailed and has a higher resolution. The second image is more pixelated and has a lower resolution. The colors in the first image are also more vibrant and realistic than the colors in the second image. Overall, the first image is a higher quality image than the second image. However, the second image is still a very impressive piece of art. <i>The AI has done a good job of creating a painting that is both creative and realistic.</i></p>
---	---

GPT-4V (proprietary) Precision Score: 1.2/2 *longest outputs*

The first image is of lower quality when compared to the second image. It appears to be slightly blurry and less sharp, which could be due to camera movement, low lighting conditions, or the use of a lower-quality camera. The perspective is that of a pedestrian walkway or underpass with people moving about, and advertisement displays along the walls. The lighting is even but doesn't seem to compensate for the overall lack of sharpness. The second image, on the other hand, has been digitally manipulated or stylized to give it a more artistic look. It features high contrast with clear definition between the dark and light areas, and the details in textures are quite pronounced, giving it a sharp, possibly high-resolution appearance. It appears as though an artistic filter has been applied to it, such as a posterizing effect or similar graphic treatment, enhancing edges and contours to create a more graphic, possibly illustrated look. In summary, the first image is more of a candid, real-world scene with lower visual clarity, whereas the second image, while not necessarily a higher resolution, has been processed to create a distinct, sharp, and stylized aesthetic.

Fig. 1: Qualitative Visualization on $Q\text{-Bench}^{\text{PAIR}}\text{-A2}$. GPT-4V gives longest outputs and achieves high precision score even if it includes incorrect information.

3.2 Further Analysis on 2AFC-LMM

Herein, we further provide a more detailed analysis of direct binary judgment accuracy between GPT-4V and **Co-Instruct** for the two-alternative forced choice (2AFC) comparisons [5], as a supplementary to main paper Tab. 4. Though showing slightly inferior aggregated linear correlation with MOS on MM21 in Tab. 4, **Co-Instruct** actually demonstrates superior direct binary choice accuracy than GPT-4V **across all datasets**. This discrepancy suggests that the current aggregation strategy might be further improved to provide a fairer evaluation on the visual quality comparison ability of LMMs.

	CSIQ	KADID-10k	MM21	CLIVE	KonIQ-10k	SPAQ	<i>weighted avg.</i>
GPT-4V	0.589	0.540	0.544	0.685	0.691	0.736	0.646
Co-Instruct	0.600	0.662	0.552	0.713	0.815	0.831	0.723

Table 2: 2AFC comparisons between GPT-4V and **Co-Instruct** in terms of accuracy, computing binary judgment accuracy while swapping *first image* and *second image*.

4 Statement on Human Subjective Study

The training dataset, Co-Instruct-562K, is completely generated by models (E5-Mistral, Mixtral-8×7B & GPT-4V) and does not involve additional human participation. For the benchmark set, the **MICBench**, it is collected **in-lab** with expert annotation. During annotation, experts are shown with three or four images listed horizontally (The first image to the last image from left to right). The instructions for annotators are as follows:

Please ask questions that compare the quality
or related attributes (clarity, color, brightness, etc)
of the given images.

The leftmost image should be referred as first image (or image 1),
and then second image (or image 2), and so on.

You can also ask questions that relate to only one images,
but you need to clearly refer which image you ask about.

Despite questions, please also provide the correct answer
and at lease one false candidate(s) for each question.

During the annotation process, we set a REPORT button for annotating experts to report any inappropriate or violent content. We have not received such report during the annotation process.

5 Social Impacts

This paper presents work whose goal is to advance the field of Machine Learning as well as Machine Visual Perception. There are many potential societal consequences of our work, none which must be specifically highlighted here.

References

1. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mixtral of experts. CoRR [abs/2401.04088](https://arxiv.org/abs/2401.04088) (2024) 1
2. Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., Hashimoto, T.B.: AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval (2023) 4

3. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) [3](#)
4. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. CoRR [abs/2307.09288](#) (2023) [3](#)
5. Zhu, H., Sui, X., Chen, B., Liu, X., Chen, P., Fang, Y., Wang, S.: 2AFC prompting of large multimodal models for image quality assessment. CoRR [abs/2402.01162](#) (2024) [4](#)