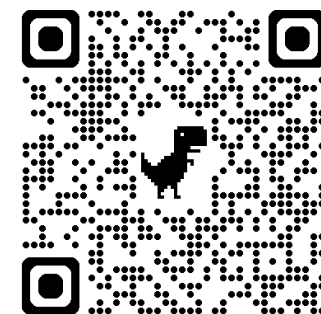







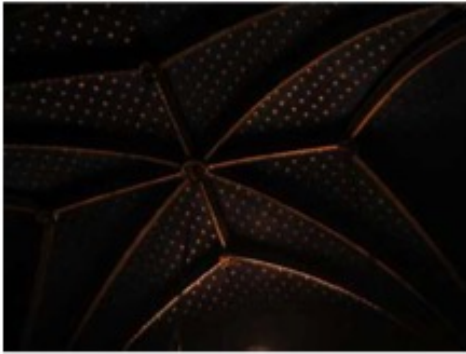
Towards Open-ended Visual Quality Comparison

HaoningWu*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, Weisi Lin¹

¹Nanyang Technological University, ²City University of Hong Kong,
³Shanghai Jiao Tong University, ⁴Sensetime Research
(*Equal Contribution)



- Visual Quality Comparison

Images				
Quality Evaluation (single image)	<p>How is the exposure level of the image? Please choose among "high", "medium" and "low".</p> <p>9 observers answered high 8 observers answered medium 13 observers answered low 6 observers answered medium 7 observers answered low 2 observers answered medium <input checked="" type="checkbox"/> all 15 observers answered low</p>			
Quality Comparison (multiple images)	<p>Please rank the exposure level among the four images.</p> <p><input checked="" type="checkbox"/> all 15 observers answered (1)>(2)>(3)>(4)</p>		<p>Which image has the highest exposure?</p> <p><input checked="" type="checkbox"/> all 15 observers answered (1)</p>	

Existing LMMs suffer from the ambiguity on absolute evaluations but provide consistent response with **comparative settings**

Q-Instruct-200K

Instruction Tuning Data for Quality Evaluation on Single Images (existing)

(a) Merge2Compare



(Human-labeled quality description)
<desc₀>: The focus of this image is blurred.....



(Human-labeled quality description)
<desc₁>: This image is predominantly black and white.....

The first image: <desc₀>.
The second image <desc₁>.
Which image has better quality, and why?



Single-modal LLM
(Mixtral-8x7B)

"merged"
Comparisons

The second image has better quality. Though the color

*Case of two images (pairs) as example.
Also includes groups of three/four images.

(b) Teach2Compare



**Which image has better quality, and why?
Judge and discuss briefly.

**Please generate questions with answers that
compares the quality of the two images.



General Comparisons

The first image has better quality. As for clarity, the first image is clearly visible

Q: Which image is brighter?
A: The first image.

Question-Answer Pairs

*Case of two images (pairs) as example.
Also includes groups of three/four images.

The first instruction-tuning dataset for visual quality comparison:

- **Merge2Compare:** LLM-merged comparisons from Q-Instruct-200K
- **Teach2Compare:** GPT4V pseudo-labeled comparisons

Merge2Compare

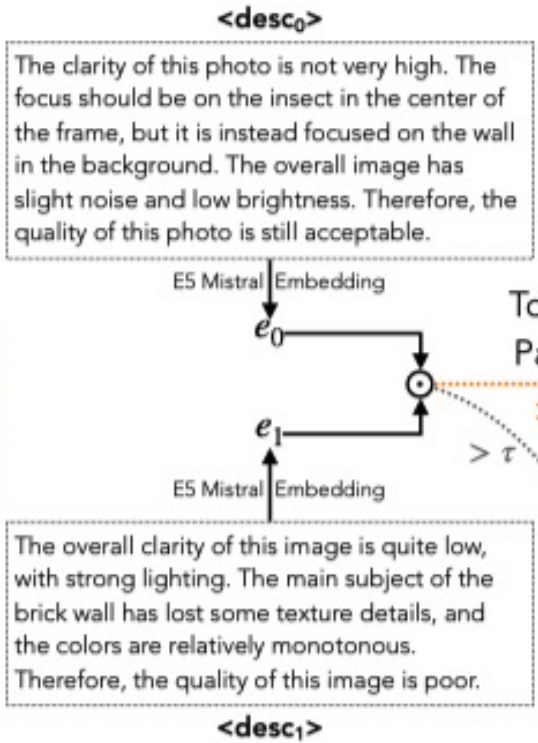
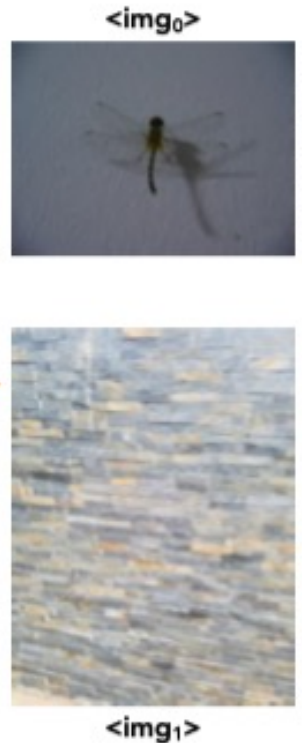


*Case of two images as example.



Q-Pathway
Individual Images
with Quality Descriptions

(a)
Pair/Group
Matching
 $O(n \log n)$ Scale



(c)
LLM Merging
Mixtral-8x7B



The first image has better quality than the second image. Although the first image has some issues such as low brightness, slight noise, and a slightly out-of-focus subject, it still retains more detail and color than the second image. The second image has poor clarity, with a loss of texture details in the main subject and monotonous colors, making it less visually appealing and informative than the first image.

"merged" Comparison

Merge2Compare
LLM-merged General Quality Comparisons

Images are first matched into groups (a), and then filtered via top-similarity removal (b). After filtering, the single image quality descriptions are merged (c) into comparisons by the LLM

*Case of two images as example.



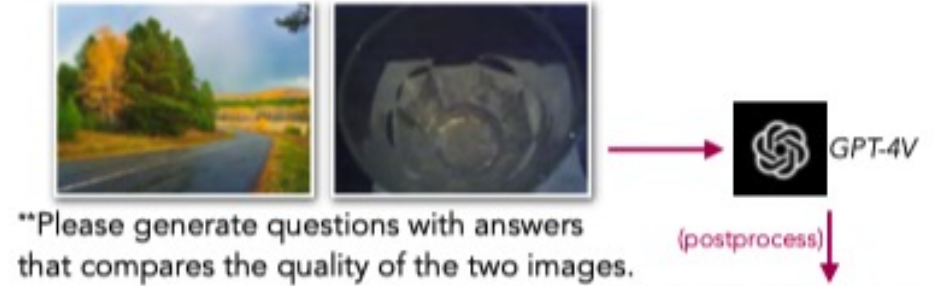
9K Images { 5.4K in-the-wild
1.8K artificially-distorted
1.8K AI-generated

(a) Collection of Images



The first image, showing a bathroom interior, has better quality compared to the second image. It has fewer issues with overexposure and clarity, making the content more recognizable. The second image is heavily overexposed and lacks clear content, making it difficult to discern the details of the scene. The "Friendly Assistance" emblem is visible, but the surrounding elements are washed out, lacking definition and contrast. In contrast, the bathroom scene, despite some visual distortions, maintains a level of detail that allows for the identification of objects and their relationships within the space.

(b) Teach2Compare-general



**Direct
Question-Answering**

Q: Is the first image more colorful than the second?
A: Yes.

**Multi-choice
Questions
(MCQ)**

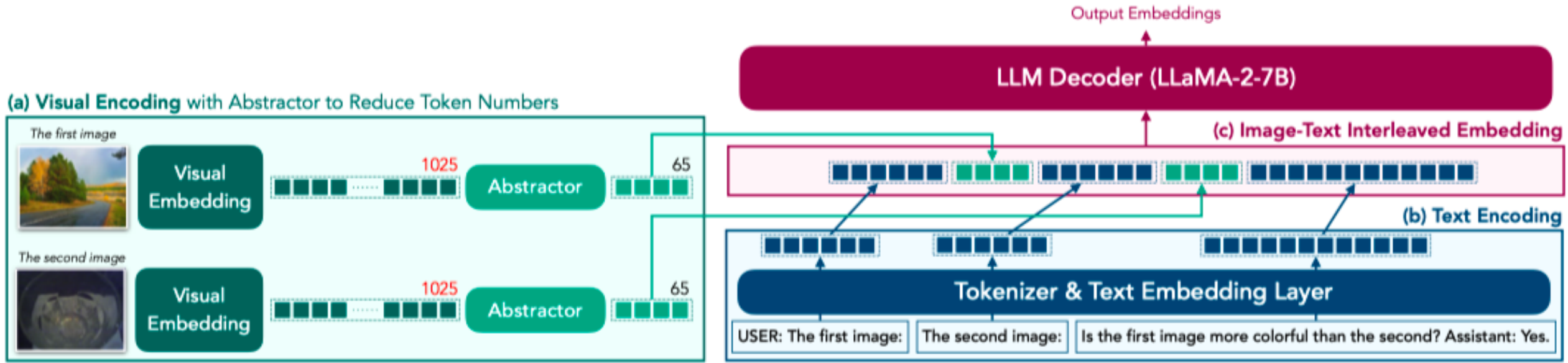
Q: Does the second image show a clear subject like the first image? A. Yes; B. No.
Answer with the option's letter from the given choices directly.
A: Yes.

**Actual prompts are slightly more complicated. See supplementary for details.

(c) Teach2Compare-Q&A

9K diverse images are collected and matched into 30K groups (a). The groups are then fed to GPT-4V to obtain *general* quality comparisons (b) and question-answering (c) related to quality comparisons.

The structure of Co-Instruct



(a) Images are encoded by visual embedding layers and then passed through an abstractor module to reduce token numbers, and then (c) fused with text embeddings into under the image-text interleaved format.

Prompts

User: The first image: `<img0>` The second image: `<img1>` (...) `<query>`
 Assistant: `<response>`

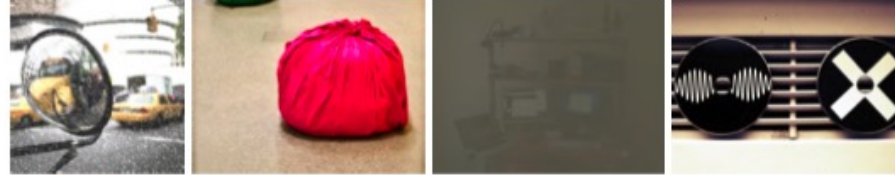
The MICBench



Which (60%)



Question:
Among all the images, which image has the most vivid color?
Candidates: A. The first; B. The third; C. The second.
Correct Answer: B. The third



Question:
Which image has the most noise?
Candidates: A. The first one; B. The second one; C. The third one; D. The fourth one
Correct Answer: A. The first one

Yes-or-No (22%)



Question:
Is the second image the clearest among the three?
Candidates: A. Yes; B. No.
Correct Answer: B. No



Question:
Is there a noticeable difference in clarity among these four photos?
Candidates: A. Yes; B. No.
Correct Answer: B. No

Others (18%)



Question:
In terms of clarity, how does the second image compare to the first one?
Candidates: A. The first has better clarity; B. The second has better clarity;
C. Both have good clarity; D. Both have poor clarity
Correct Answer: A. The first has better clarity



Question: Compared to the second image, how is the color of the fourth image?
Candidates: A. The fourth image has better color;
B. The fourth image has worse color; C. Two images have similar color
Correct Answer: A. The fourth image has better color

(a) **Which** questions (60%), (b) **Yes-or-No** questions (22%), and (c) **Other** types of questions (18%) on three/four images.

We introduce the MICBench to cover the open-ended evaluation settings on groups of **three or four** images, as a complementary of existing evaluation settings

Experiments: Q-Bench^{PAIR}-A1 (1,999 MCQs)



Sub-categories	Question Types			Low-level Concerns		Pairwise Settings		Overall↑
	Yes-or-No↑	What↑	How↑	Distortion↑	Other↑	Compare↑	Joint↑	
<i>random guess accuracy</i>	50.00%	32.03%	33.16%	38.95%	41.95%	38.69%	43.70%	39.82%
(Sep/2023) LLaVA-v1.5-13B	57.34%	47.45%	49.13%	49.01%	59.51%	52.06%	52.00%	52.05%
(Oct/2023) BakLLava	60.09%	45.42%	50.86%	53.09%	58.82%	54.52%	55.55%	52.75%
(Nov/2023) mPLUG-Owl2 (<i>baseline of Co-Instruct</i>)	58.07%	36.61%	48.44%	47.74%	51.90%	45.73%	60.00%	48.94%
(Dec/2023) Emu2-Chat	51.94%	29.78%	53.84%	42.01%	55.71%	46.26%	49.09%	47.08%
(Feb/2024) InternLM-XComposer2-VL	71.81%	58.64%	62.28%	65.77%	63.67%	64.34%	68.00%	65.16%
Qwen-VL-Max (<i>Proprietary</i>)	67.65%	67.56%	65.35%	69.09%	61.18%	68.65%	61.29%	66.99%
Gemini-Pro (<i>Proprietary</i>)	65.78%	56.61%	56.74%	60.42%	60.55%	60.46%	60.44%	60.46%
GPT-4V (<i>Proprietary, teacher of Co-Instruct</i>)	79.75%	69.49%	84.42%	77.32%	79.93%	81.00%	68.00%	78.07%
<i>Non-expert Human</i>	78.11%	77.04%	82.33%	78.17%	77.22%	80.26%	76.39%	80.12%
<i>without Multi-image Comparative Data</i>	60.24%	47.46%	48.78%	52.81%	53.97%	51.42%	59.11%	53.15%
Co-Instruct (Ours)	86.50%	72.20%	79.23%	80.00%	80.62%	81.91%	74.22%	80.18%

Co-Instruct shows far superior accuracy than open-source LMMs: it is 64% better than its baseline (mPLUG-Owl2), 51% better than the variant without our multi-image subsets, and also 23% better than the best of them.

Experiments: Q-Bench^{PAIR}-A2 (499 Descriptions)



Dimensions ----- Model	Completeness				Precision				Relevance				Sum.↑
	P_0	P_1	P_2	score↑	P_0	P_1	P_2	score↑	P_0	P_1	P_2	score↑	
(Sep/2023) LLaVA-v1.5-13B	18.77%	73.44%	7.79%	0.89	34.66%	38.72%	26.62%	0.92	1.02%	34.59%	64.39%	1.63	3.44
(Oct/2023) BakLLava	29.46%	59.77%	10.57%	0.80	40.0%	38.08%	21.33%	0.80	2.26%	15.06%	82.04%	<u>1.79</u>	3.40
(Nov/2023) mPLUG-Owl2 (<i>baseline</i>)	19.43%	65.54%	14.45%	0.94	30.94%	43.71%	24.63%	0.92	3.79%	26.94%	68.28%	1.63	3.50
(Dec/2023) Emu2-Chat	41.25%	54.33%	4.42%	0.63	38.11%	36.41%	25.48%	0.87	4.12%	38.61%	57.27%	1.53	3.03
(Feb/2024) InternLM-XComposer2-VL	13.20%	72.17%	14.13%	1.00	31.28%	42.13%	25.77%	0.93	1.60%	24.17%	72.93%	1.70	3.64
----- Qwen-VL-Max (<i>Proprietary</i>)	11.64%	54.08%	34.08%	1.22	24.26%	39.15%	36.22%	1.11	2.533%	10.97%	85.64%	1.82	4.16
Gemini-Pro (<i>Proprietary</i>)	18.22%	44.48%	36.84%	1.18	34.13%	37.95%	27.02%	0.92	0.67%	5.91%	92.22%	1.90	4.00
GPT-4V (<i>Proprietary, teacher of Ours</i>)	4.09%	31.82%	64.09%	1.60	10.44%	45.12%	44.44%	1.34	0.18%	1.69%	96.35%	1.94	4.89
----- <i>w/o Multi-Image Comparative Data</i>	15.25%	65.76%	18.32%	1.02	39.44%	40.18%	19.62%	0.79	0.09%	9.86%	89.02%	1.87	3.69
----- Co-Instruct (Ours)	4.04%	31.55%	63.55%	1.58	13.68%	43.68%	41.37%	1.26	0.0%	0.44%	98.22%	1.96	4.82

The capability of **Co-Instruct** in reasoning-related comparisons can match that of GPT-4V, while significantly surpassing other existing LMMs

Experiments: 2AFC-LMMs



Consistency (κ), Correlation (ρ)

Dataset Model	CSIQ		MM21		KADID-10k		LIVEC		KonIQ-10k		SPAQ		<i>weighted avg.</i>	
	κ	ρ	κ	ρ	κ	ρ	κ	ρ	κ	ρ	κ	ρ	κ	ρ
(Aug/2023) IDEFICS-Instruct-9B	0.206	0.570	0.337	0.338	0.202	0.552	0.323	0.492	0.251	0.479	0.330	0.474	0.286	0.470
(Sep/2023) LLaVA-v1.5-13B	0.483	0.423	0.356	0.149	0.310	0.137	0.273	0.162	0.262	0.403	0.291	0.156	0.302	0.224
(Oct/2023) BakLLava	0.356	0.235	0.337	0.244	0.245	0.166	0.296	0.159	0.185	0.217	0.274	0.146	0.261	0.185
(Nov/2023) mPLUG-Owl2 (<i>baseline</i>)	0.435	0.627	0.378	0.306	0.402	0.443	0.375	0.441	0.386	0.417	0.362	0.356	0.460	0.397
(Feb/2024) InternLM-XComposer2-VL	0.800	0.527	0.688	0.377	0.600	0.552	0.600	0.516	0.825	0.581	0.700	0.755	0.705	0.567
Qwen-VL-Max (<i>Proprietary</i>)	0.540	0.418	0.497	0.304	0.625	0.406	0.578	0.544	0.631	0.610	0.592	0.718	0.592	0.540
Gemini-Pro (<i>Proprietary</i>)	0.672	0.527	0.604	0.377	0.790	0.552	0.650	0.516	0.652	0.581	0.671	0.755	0.678	0.622
GPT-4V (<i>Proprietary, teacher of Ours</i>)	0.778	0.764	0.792	0.474	0.763	0.560	0.837	0.685	0.835	0.800	0.871	0.876	0.823	0.721
<i>w/o Multi-Image Comparative Data</i>	0.117	0.650	0.480	0.392	0.397	0.466	0.327	0.432	0.489	0.512	0.485	0.397	0.432	0.449
Co-Instruct (Ours)	0.800	0.779	0.852	0.325	0.829	0.685	0.872	0.797	0.883	0.927	0.881	0.931	0.864	0.754

- Co-Instruct outperforms all existing models in 2AFC-LMM, including GPT-4V
- Co-Instruct also shows very high consistency κ while swapping two images

Experiments: MICBench



Sub-categories Model	Question Types			Number of Images		Overall↑
	Yes-or-No↑	Which↑	Others↑	Three↑	Four↑	
#questions	220	594	182	503	493	996
<i>random guess accuracy</i>	49.55%	28.59%	28.31%	34.10%	29.17%	31.47%
(Sep/2023) LLaVA-v1.5-13B (<i>length: 2048→2560</i>)	47.51%	40.74%	52.49%	46.81%	41.90%*	44.38%
(Oct/2023) BakLLava (<i>length: 2048→2560</i>)	68.35%	35.01%	52.78%	48.51%	42.54%*	45.56%
(Nov/2023) mPLUG-Owl2 (<i>baseline of Co-Instruct</i>)	62.25%	35.70%	53.71%	44.19%	45.42%	44.80%
(Feb/2024) InternLM-XComposer2-VL (<i>length: 4096→5120</i>)	62.95%	47.29%	52.02%	55.70%	46.51%*	51.76%
Qwen-VL-Max (<i>Proprietary</i>)	62.33%	70.00%	81.54%	72.35%	68.79%	70.55%
Gemini-Pro (<i>Proprietary</i>)	75.00%	67.37%	66.92%	68.71%	70.87%	69.79%
GPT-4V (<i>Proprietary, teacher of Co-Instruct</i>)	80.32%	77.28%	78.82%	80.32%	77.28%	78.82%
<i>Non-expert Human</i>	82.27%	78.15%	74.31%	77.18%	79.55%	78.35%
<i>without Multi-image Comparative Data</i>	62.72%	37.54%	53.30%	45.33%	46.65%	45.98%
Co-Instruct (Ours)	79.55%	85.35%	81.32%	84.69%	81.94%	83.33%

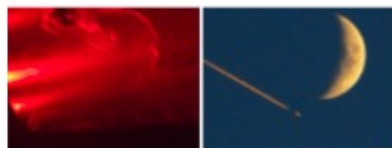
Co-Instruct provides very competitive accuracy on open-question quality comparison among three/four images, **5.7%** better than GPT-4V (best existing) **and 6.4%** more accurate than non-expert human; open-source LMMs even struggle to obtain 50%

Experiments: Overall



Compare the quality between the two images in details.

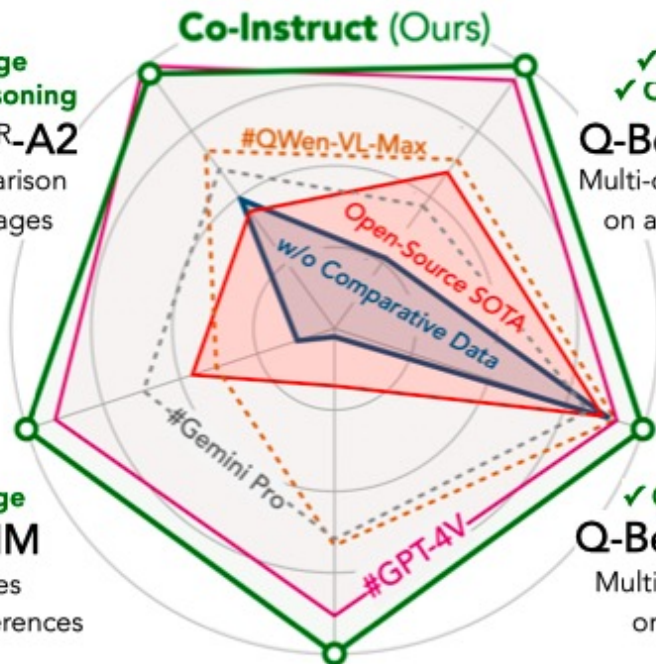
✓ Multi-Image
✓ Detailed-Reasoning
Q-Bench^{PAIR-A2}
Detailed Comparison on a Pair of Images



Which image has better quality?

↓
Regress to Quality Scores

✓ Multi-Image
2AFC-LMM
Quality Scores From Binary Preferences



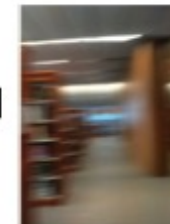
✓ Multi-Image
✓ Open-Question
Q-Bench^{PAIR-A1}
Multi-choice Questions on a Pair of Images



What makes the first image blurrier than the second image?

- A. Noise
- B. Out of focus**
- C. Low light
- D. Motion blur

✓ Open-Question
Q-Bench^{SINGLE-A1}
Multi-choice Questions on Single Images



What is the worst distortion in this image?

- A. Overexposure
- B. Noise
- C. Underexposure
- D. Motion blur**

Which image contains overexposure?



- A. First image
- B. Second image
- C. Third image**

✓ Multi-Image ✓ Open-Question
MICBench
(proposed, 2K MCQs)
Multi-choice Questions (MCQ) on Comparing Multiple Images

Which image is blurred due to motion?



- A. First image
- B. Second image
- C. Third image
- D. Fourth image**

Co-Instruct

Present by Q-Future
Open-ended Visual Quality Comparer

Co-Instruct: The first LMM outperforms GPT-4v in low-level visual quality comparison
Towards Open-ended Visual Quality Comparison (ECCV 2024)

Multiple Agents Collaboration : The outputs of Co-Instruct also support [InstructIR](#) as PLUGIN to restore image quality!

Co-Instruct Resources: [Github](#) [Code](#) [Technical Report](#) [Stars](#) 54

Image 1 (First image)

Drop Image Here
- or -
Click to Upload

Image 2 (Second image)

Drop Image Here
- or -
Click to Upload

Image 3 (Third image)

Drop Image Here
- or -
Click to Upload

Image 4 (Fourth image)

Drop Image Here
- or -
Click to Upload

Chatbot

Type a message...

[Submit](#)


[Retry](#) [Undo](#) [Clear](#)

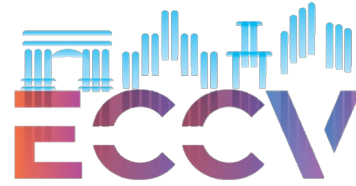
Image for Auto Restoration

Drop Image Here
- or -
Click to Upload

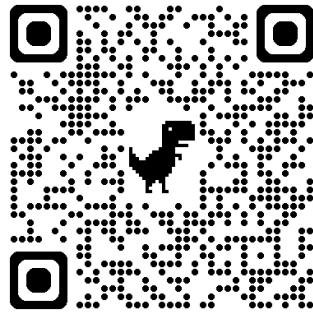
[Clear](#) [Submit](#)

Output of Auto Restoration





Thanks!



Project Page