Towards Open-ended Visual Quality Comparison

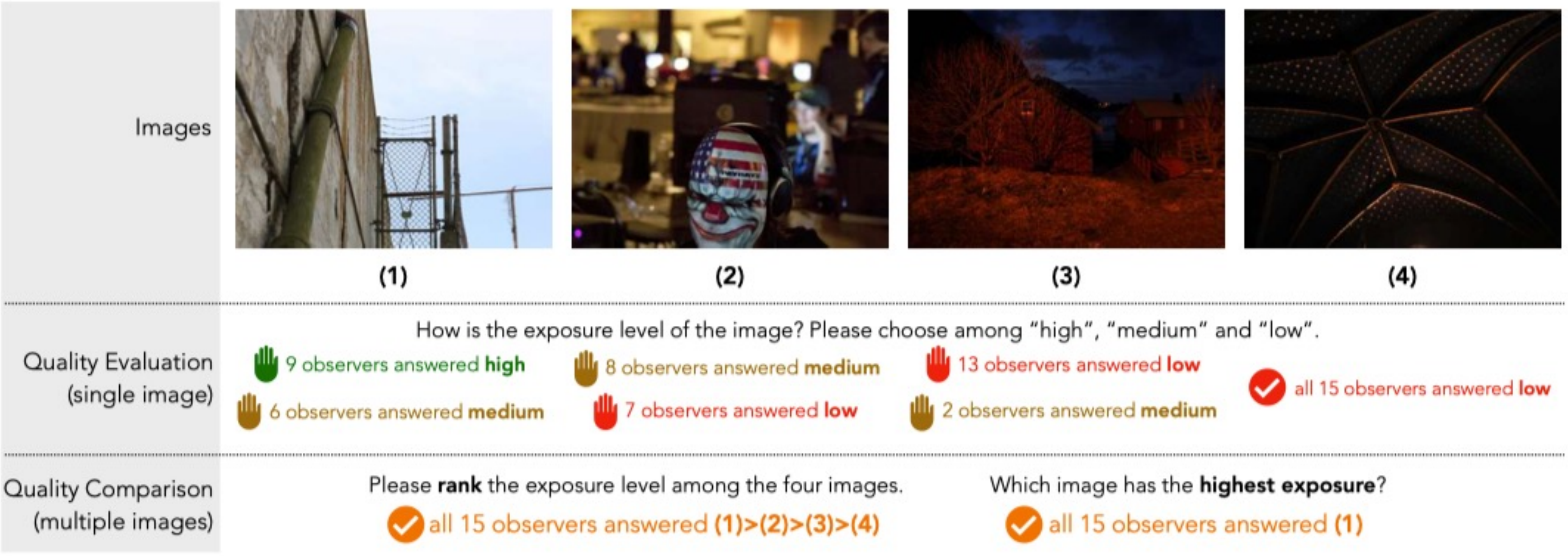
HaoningWu*[1], Hanwei Zhu*[2], Zicheng Zhang*[3], Erli Zhang[1], Chaofeng Chen[1], Liang Liao[1], Chunyi Li[3], Annan Wang[1], Wenxiu Sun[4], Qiong Yan[4], Xiaohong Liu[3], Guangtao Zhai[3] , Shiqi Wang[2], Weisi Lin[1]

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

Project Page

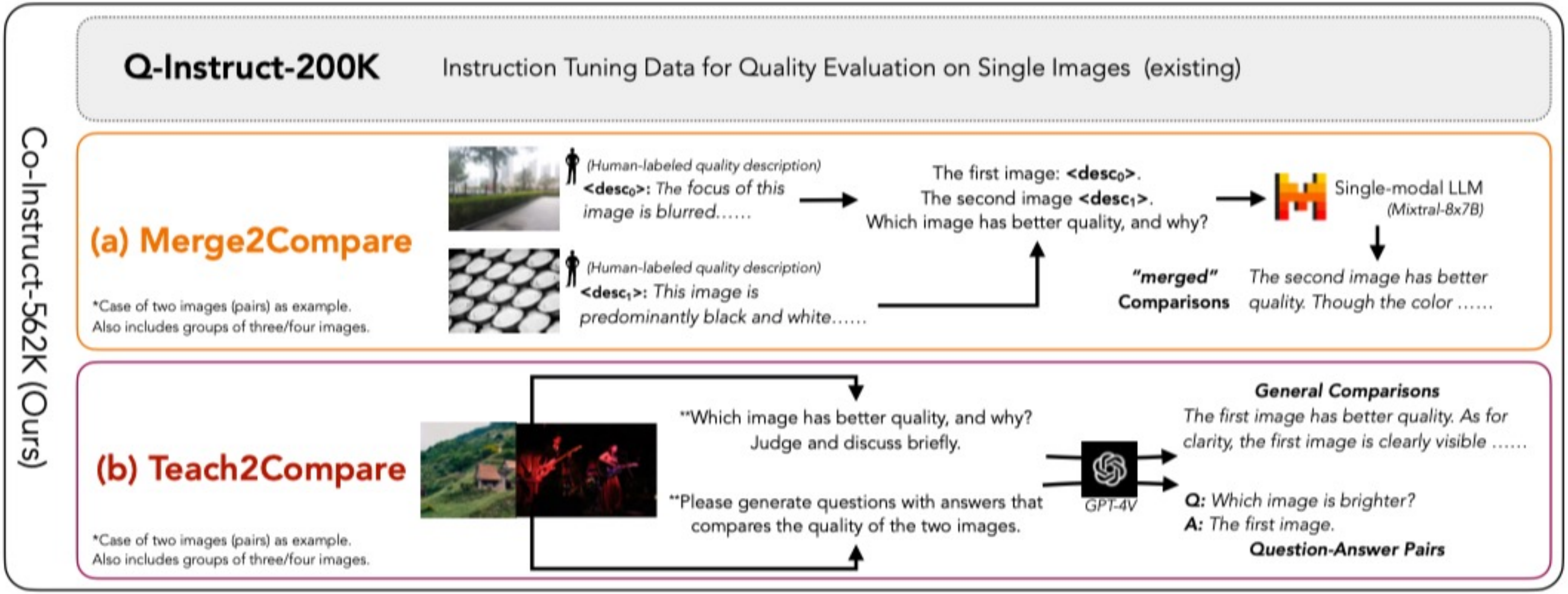
Motivation

- Visual Quality Comparison



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

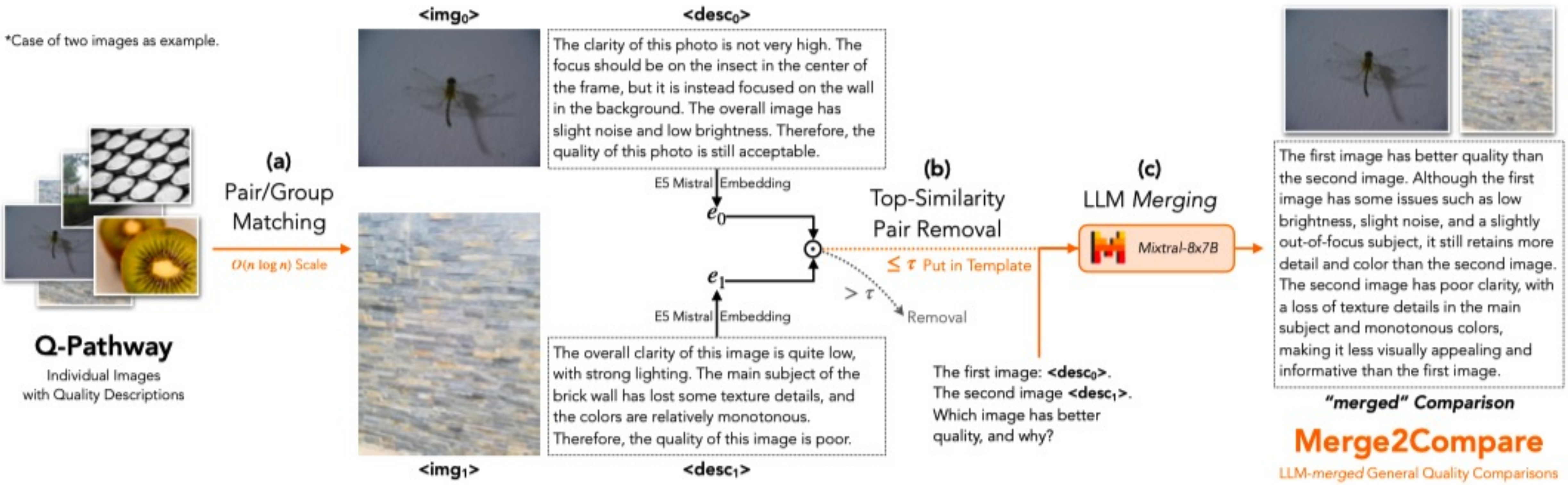
Co-Instruct-562K



The first instruction-tuning dataset for visual quality comparison:

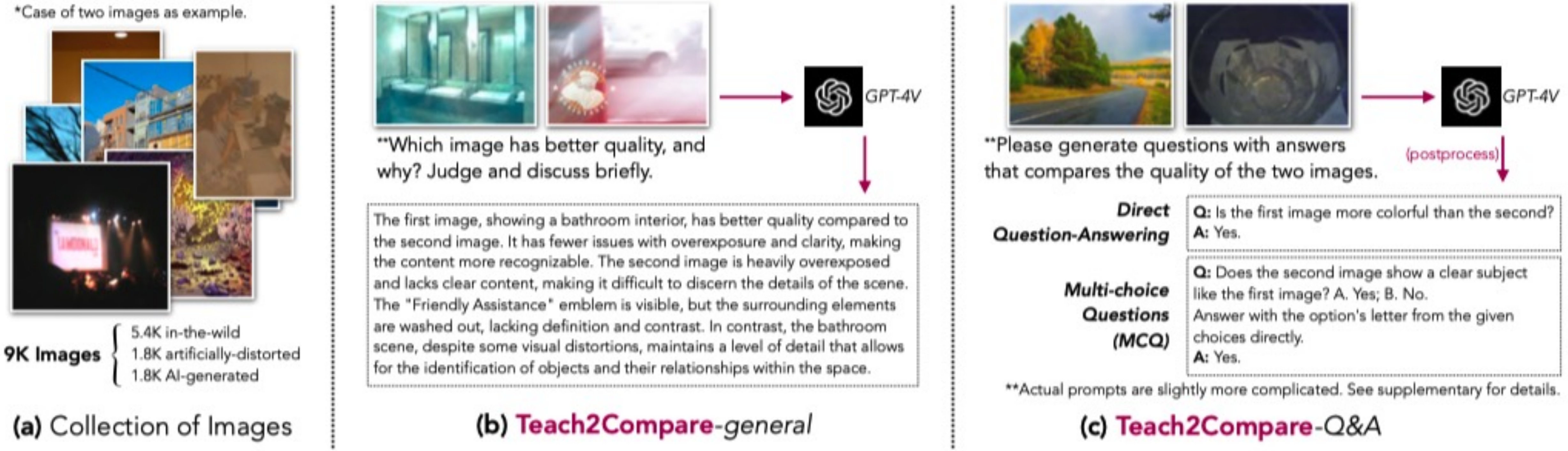
- **Merge2Compare**: LLM-merged comparisons from Q-Instruct-200K
- **Teach2Compare**: GPT4V pseudo-labeled comparisons

Merge2Compare



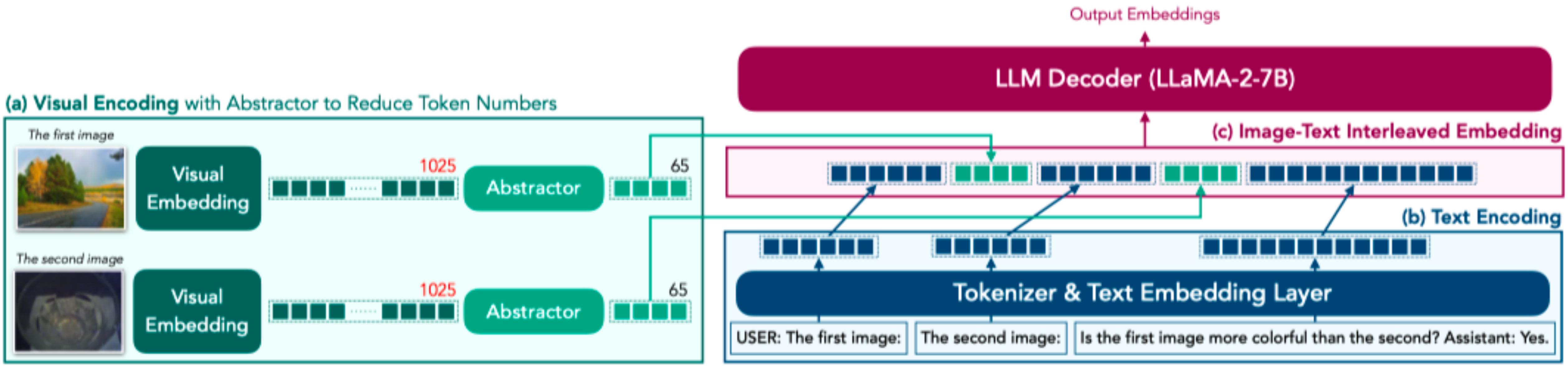
Images are first matched into groups **(a)**, and then filtered via top-similarity removal **(b)**. After filtering, the single image quality descriptions are merged **(c)** into comparisons by the LLM

Teach2Compare



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

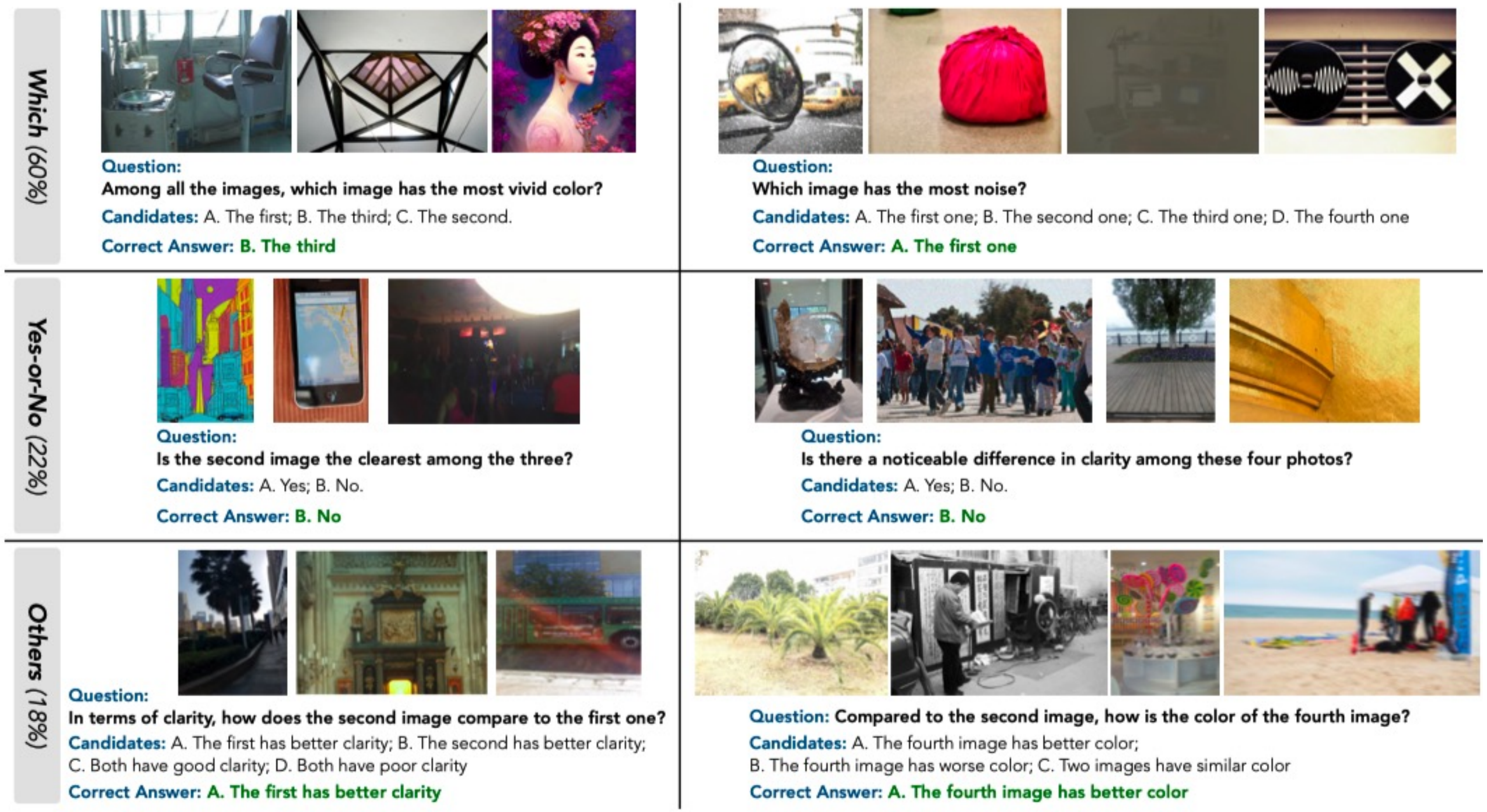


(a) Images are encoded by visual embedding layers and then passsed through an abstractor module to reduce token numbers, and then (c) fused with text embeddings into under the image-text interleaved format.

Prompts

User: The first image: <img_0> The second image: <img_1> (...) <query>
Assistant: <response>

The MICBench



(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		
Model	*Yes-or-No*↑	*What*↑	*How*↑	*Distortion*↑	*Other*↑	*Compare*↑	*Joint*↑	*Overall*↑
random guess accuracy	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 *(baseline of **Co-Instruct**)*	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 *(Proprietary)*	67.65%	67.56%	65.35%	69.09%	61.18%	68.65%	61.29%	66.99%
Gemini-Pro *(Proprietary)*	65.78%	56.61%	56.74%	60.42%	60.55%	60.46%	60.44%	60.46%
GPT-4V *(Proprietary, teacher of **Co-Instruct**)*	**79.75%**	**69.49%**	**84.42%**	**77.32%**	**79.93%**	**81.00%**	**68.00%**	**78.07%**
Non-expert Human	78.11%	77.04%	82.33%	78.17%	77.22%	80.26%	76.39%	80.12%
without Multi-image Comparative Data	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	Completeness				Precision				Relevance				*Sum.*↑
Model	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%	1.79	3.40
(Nov/2023) mPLUG-Owl2 *(baseline)*	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 *(Proprietary)*	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 *(Proprietary)*	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 *(Proprietary, teacher of **Ours**)*	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**
w/o Multi-Image Comparative Data	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	CSIQ		MM21		KADID-10k		LIVEC		KonIQ-10k		SPAQ		*weighted avg.*	
Model	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 *(baseline)*	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 *(Proprietary)*	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 *(Proprietary)*	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 *(Proprietary, teacher of **Ours**)*	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
w/o Multi-Image Comparative Data	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**

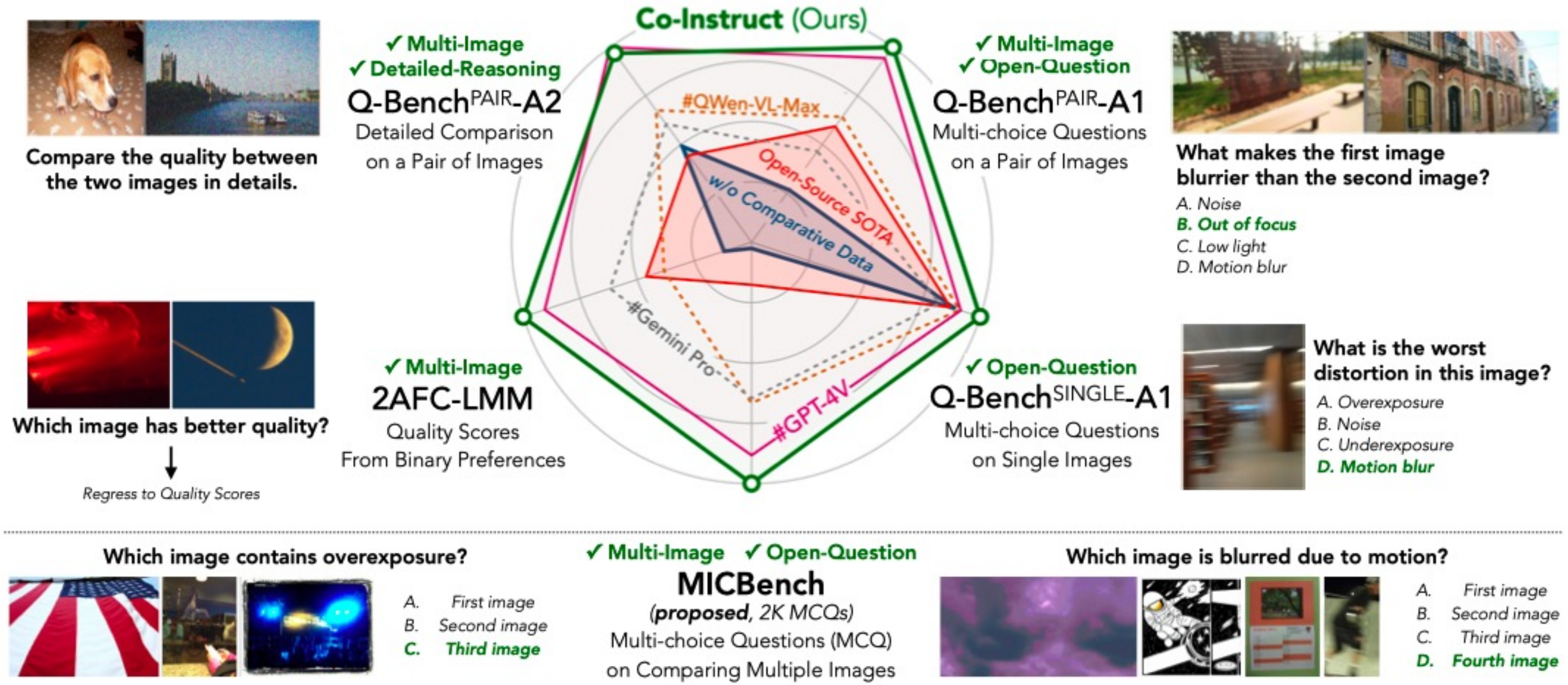
- Co-Instruct outperforms all existing models in 2AFC-LMM, including GPT-4V
- Co-Instruct also shows very high consistency κ while swapping two images

Experiments: MICBench

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

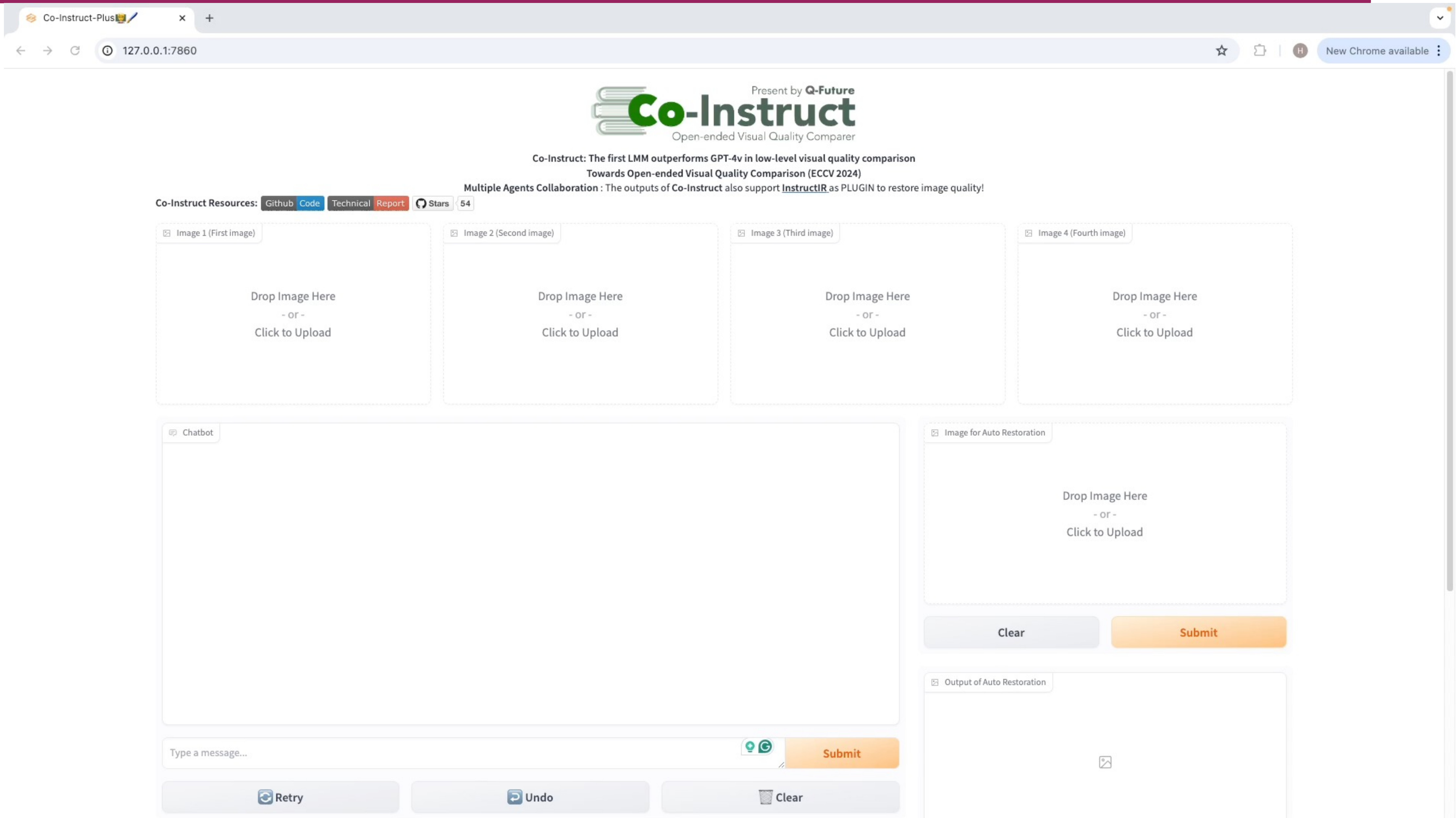
Co-Instruct provides very competitive accuracy on open-question quality comparison among three/four images, **5.7%** better than GPT-4V (best existing) **and 6.4%** more accurate than non-expert human; open-source LMMs even struggle to obtain 50%

Experiments: Overall



H. Wu, H. Zhu, **S. Wang,** and *et al.*, "Towards Open-ended Visual Quality Comparison," in *arXiv preprint arXiv:2402.16641*, 2024

Demo


Co-Instruct-Plus
127.0.0.1:7860
New Chrome available
Present by Q-Future
Co-Instruct
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)
Image 2 (Second image)
Image 3 (Third image)
Image 4 (Fourth image)
Drop Image Here
- or -
Click to Upload
Chatbot
Image for Auto Restoration
Clear
Submit
Output of Auto Restoration
Type a message...
Retry
Undo
Clear

Thanks!

Project Page