NEURAL INFORMATION PROCESSING SYSTEMS

FACE-HUMAN-BENCH

A Comprehensive Benchmark of Face and Human Understanding for Multi-modal Assistants

Lixiong Qin[1,*], Shilong Ou[1,*], Miaoxuan Zhang[1,*], Jiangning Wei[1,*], Yuhang Zhang[1,*],
Xiaoshuai Song[1], Yuchen Liu[1], Mei Wang[2], Weiran Xu[1]

[1]Beijing University of Posts and Telecommunications

[2]Beijing Normal University

* Equal Contribution

Hierarchical Taxonomy

Face-Human Bench

L1 Legend

- Face
- Human
- Perception
- Reasoning

Face-Human-Bench: A Comprehensive Benchmark of Face and Human Understanding for Multi-modal Assistants

Hierarchical Taxonomy

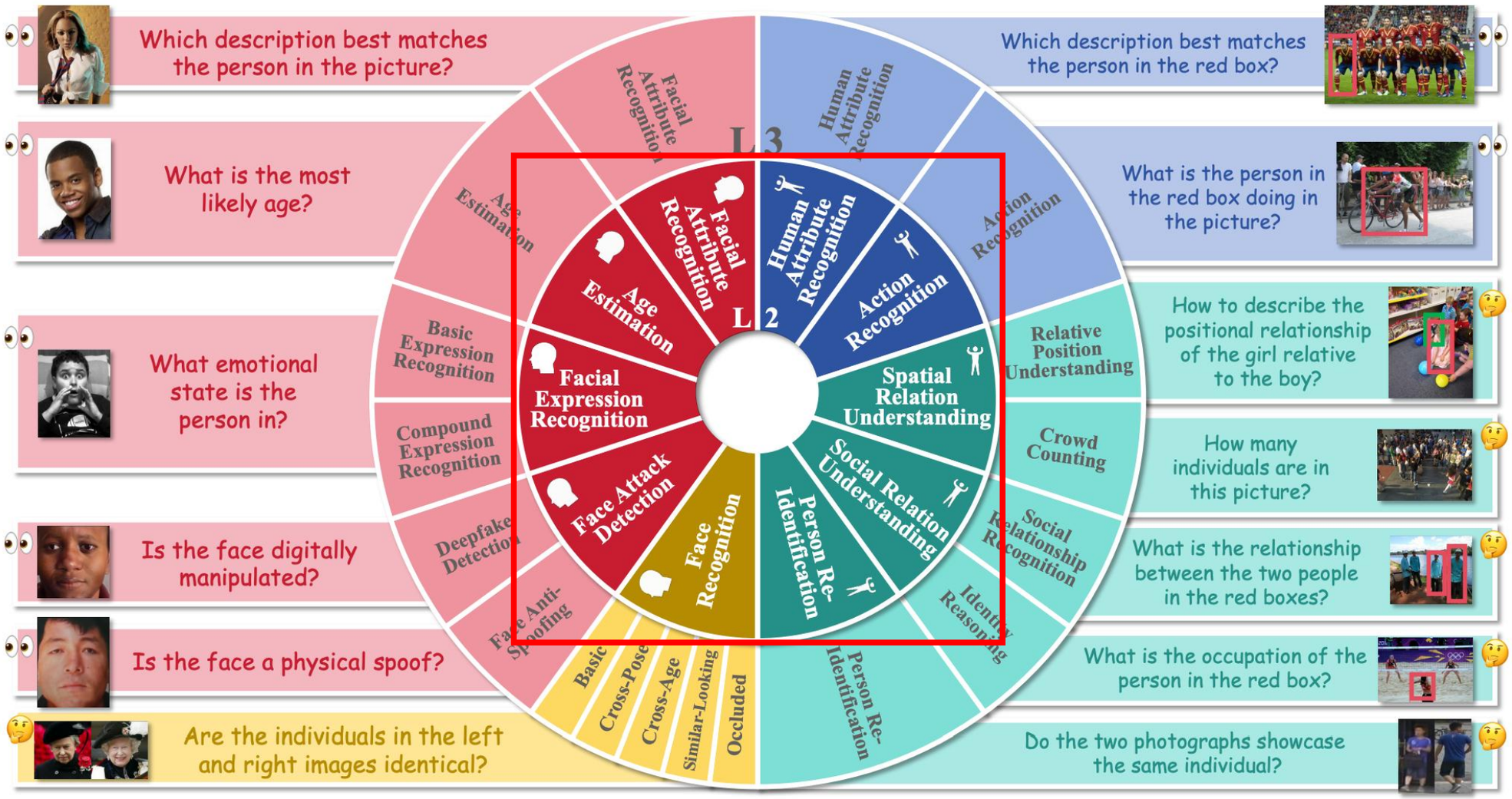
Face-Human Bench

L1 Legend

- Face
- Human
- Perception
- Reasoning

Face-Human-Bench: A Comprehensive Benchmark of Face and Human Understanding for Multi-modal Assistants

Hierarchical Taxonomy



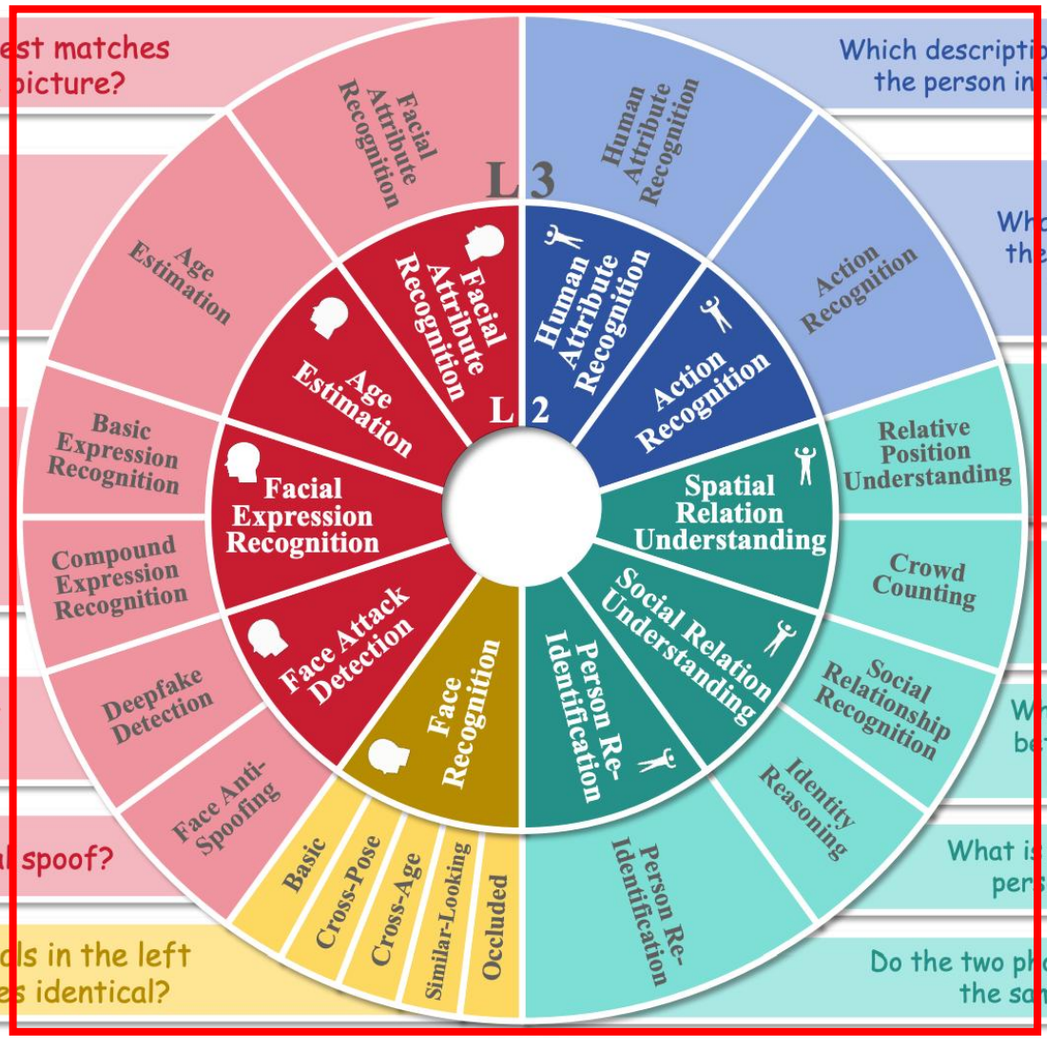
Face-Human Bench

L1 Legend

- Face
- Human
- Perception
- Reasoning

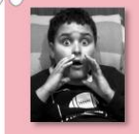
Face-Human-Bench: A Comprehensive Benchmark of Face and Human Understanding for Multi-modal Assistants

Hierarchical Taxonomy




Which description best matches the person in the picture?


What is the most likely age?


What emotional state is the person in?


Is the face digitally manipulated?


Is the face a physical spoof?


Are the individuals in the left and right images identical?

Which description best matches the person in the red box?

What is the person in the red box doing in the picture?

How to describe the positional relationship of the girl relative to the boy?

How many individuals are in this picture?

What is the relationship between the two people in the red boxes?

What is the occupation of the person in the red box?

Do the two photographs showcase the same individual?

Face-Human Bench

L1 Legend

- Face
- Human
- Perception
- Reasoning

Hierarchical Taxonomy

Face-Human Bench

L1 Legend

- Face
- Human
- Perception
- Reasoning

Face-Human-Bench: A Comprehensive Benchmark of Face and Human Understanding for Multi-modal Assistants

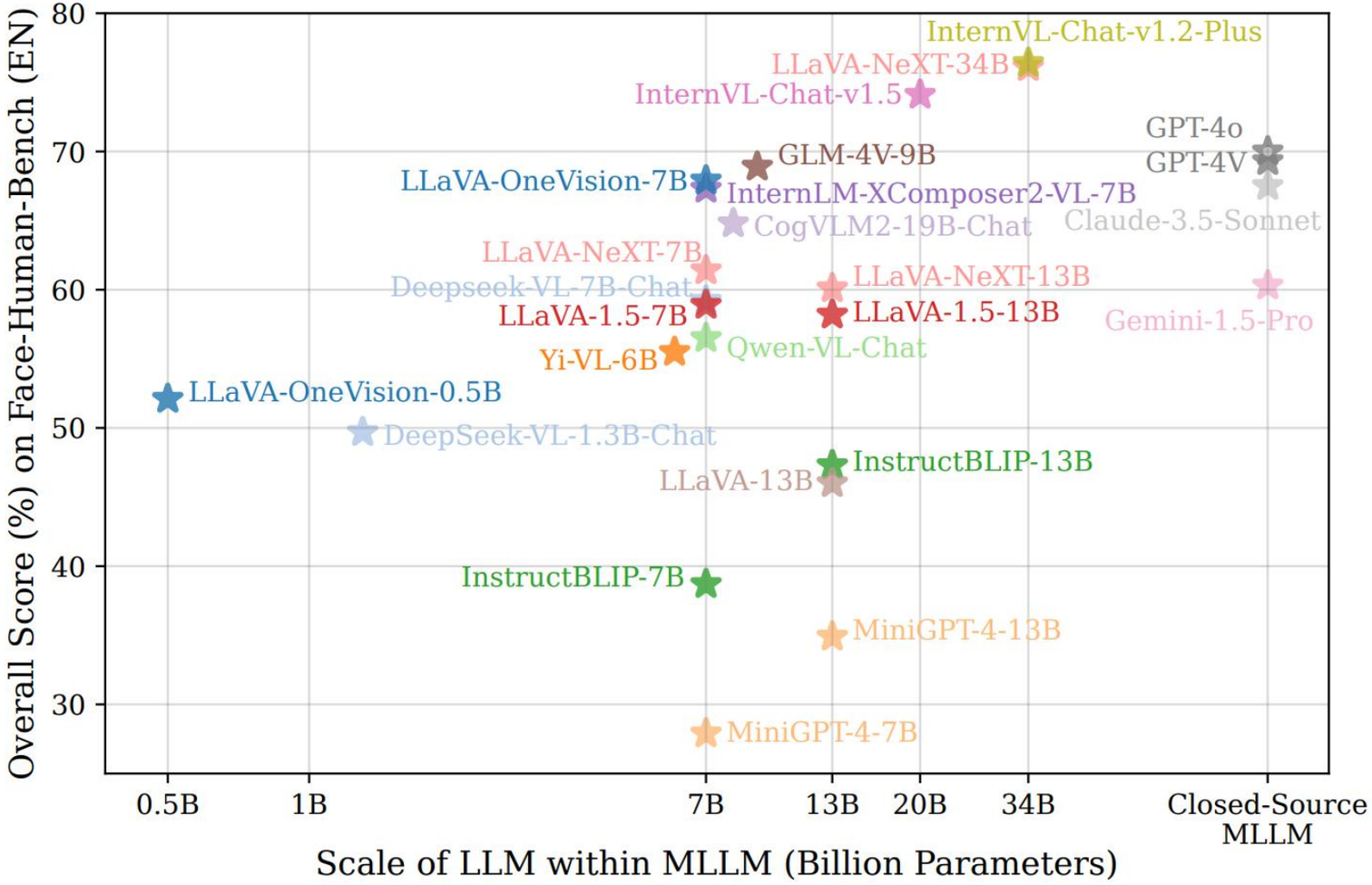
Our Research Questions

- Q1: How do existing MLLMs perform in face and human understanding?
  - 25 MLLMs' overall performance
  - Ability correlations across levels
  - Target relative position impacts
  - CoT prompting effects.

- Q2: Which tasks do specialist models outperform MLLMs at?

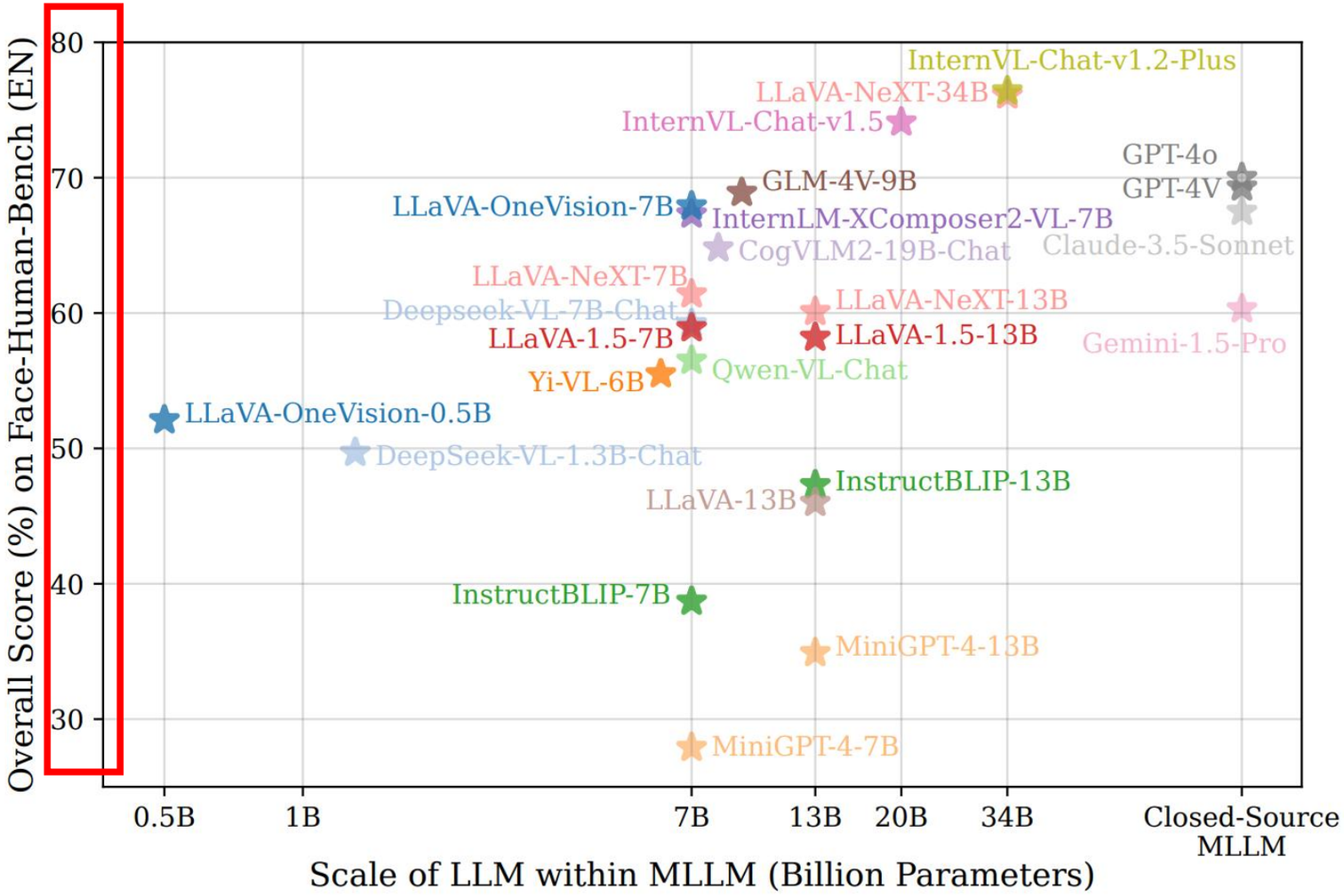
Answering Q1: MLLMs' Performance

Overall Performance



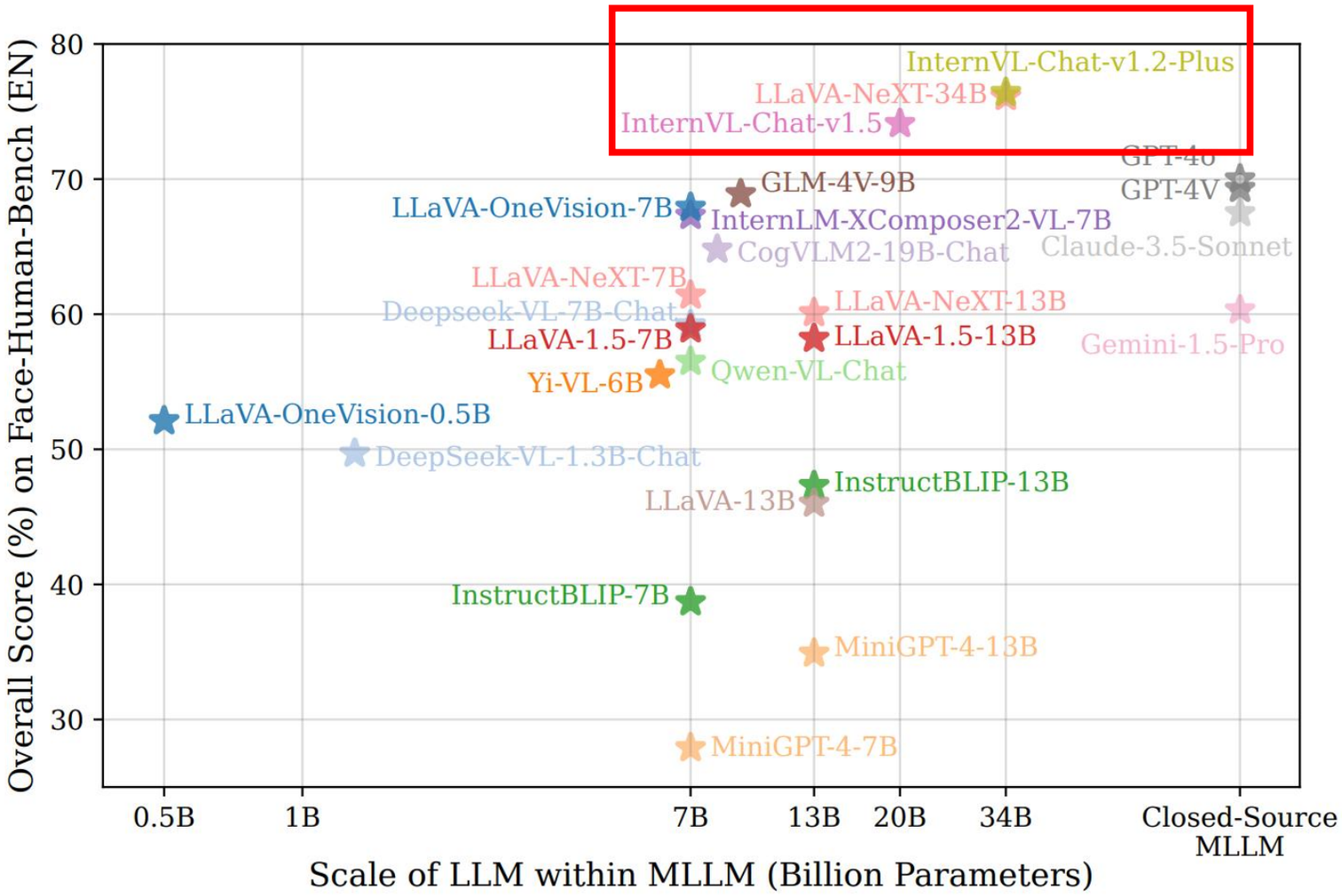
Answering Q1: MLLMs' Performance

Overall Performance



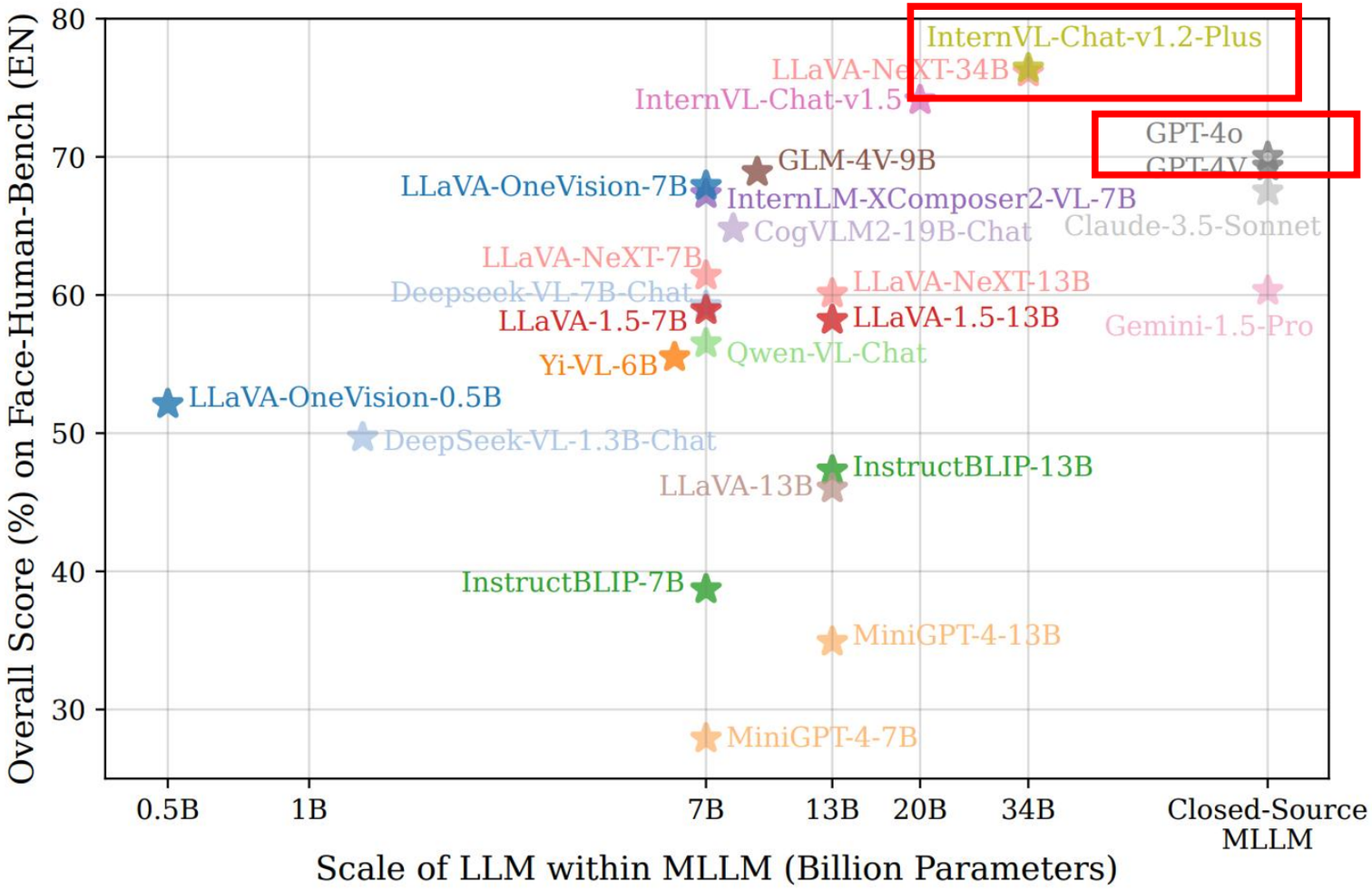
Answering Q1: MLLMs' Performance

Overall Performance



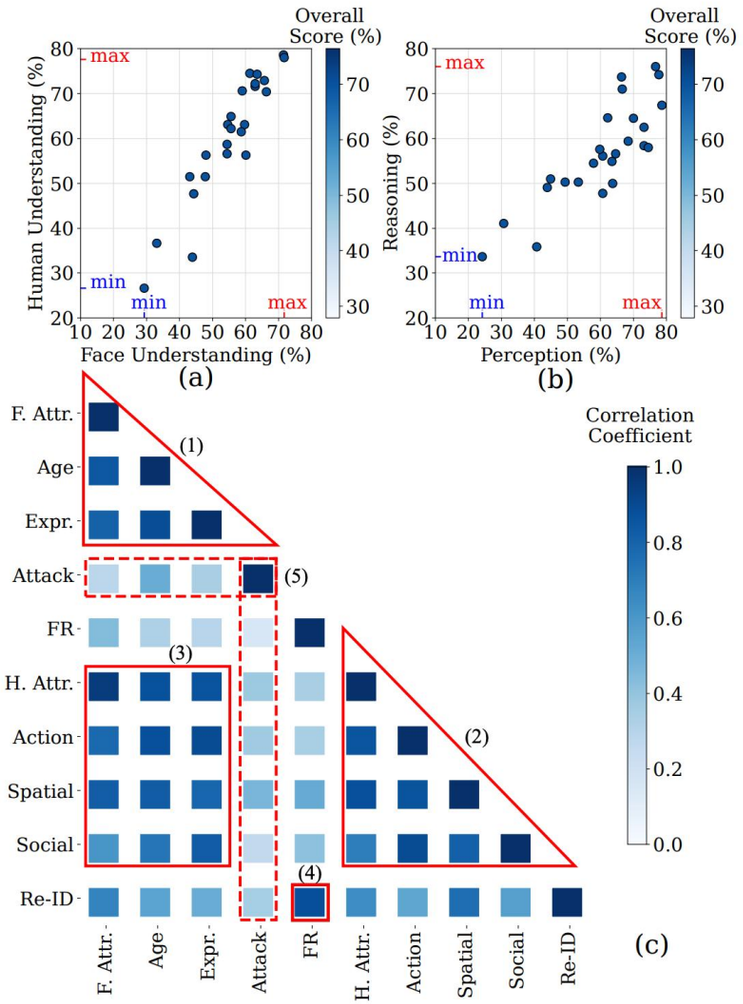
Answering Q1: MLLMs' Performance

Overall Performance



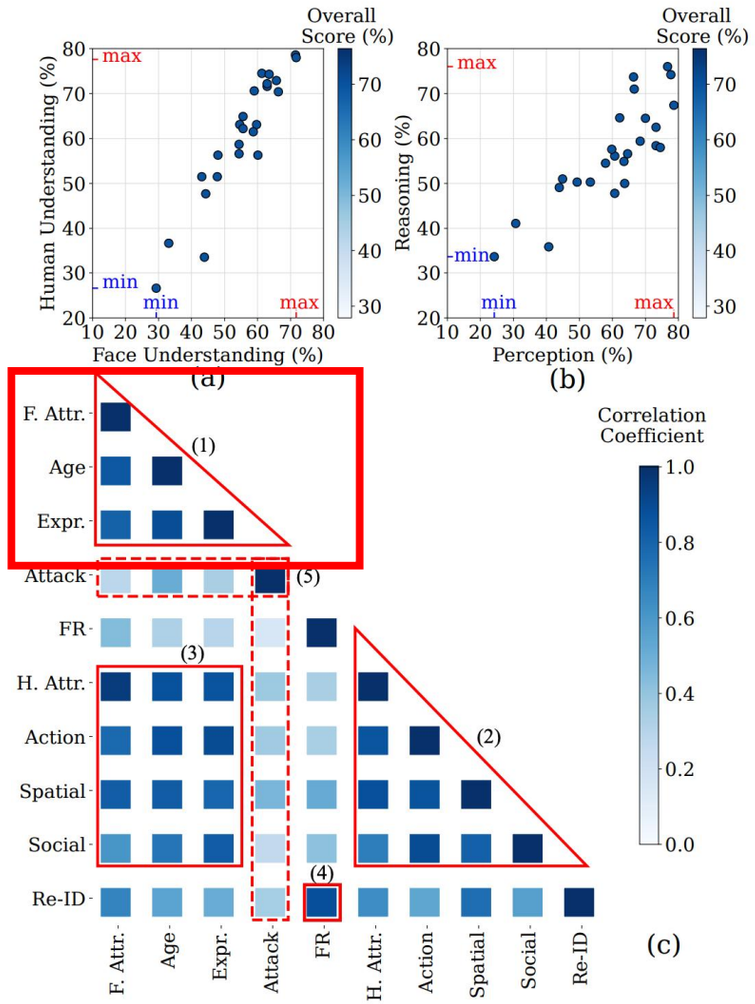
Answering Q1: MLLMs' Performance

Ability Correlations



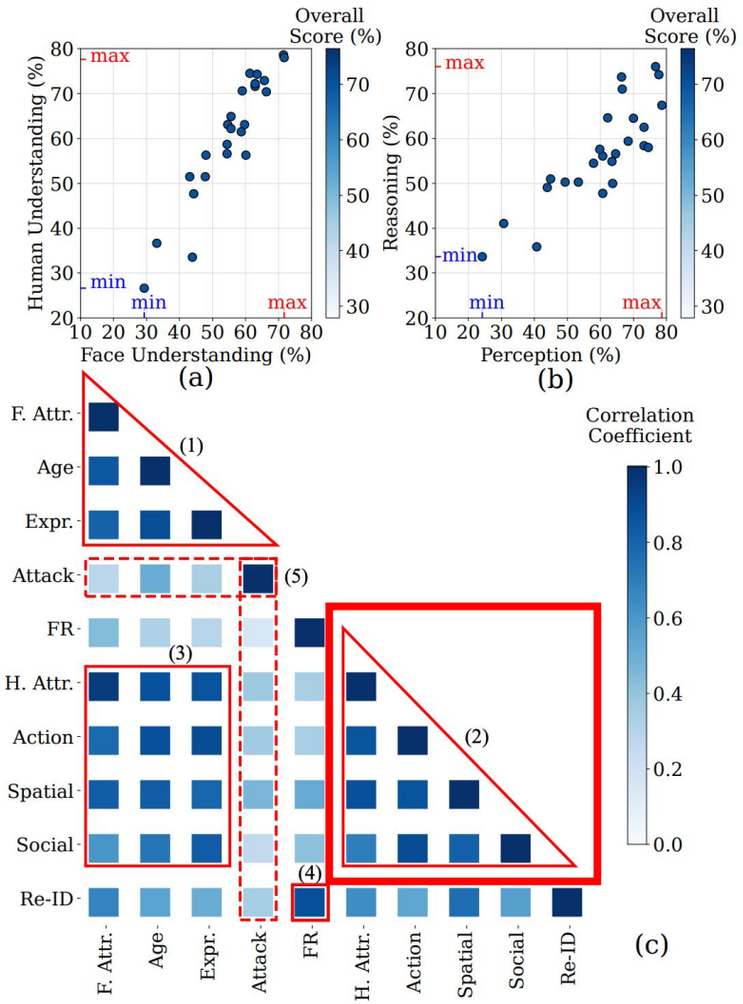
Answering Q1: MLLMs' Performance

Ability Correlations



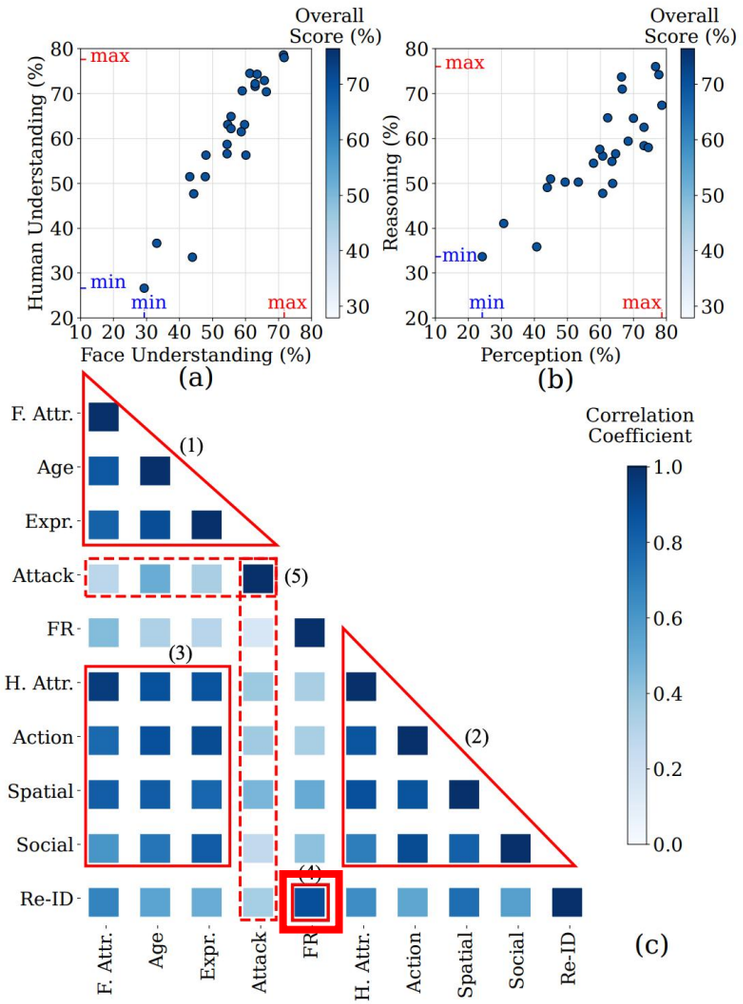
Answering Q1: MLLMs' Performance

Ability Correlations



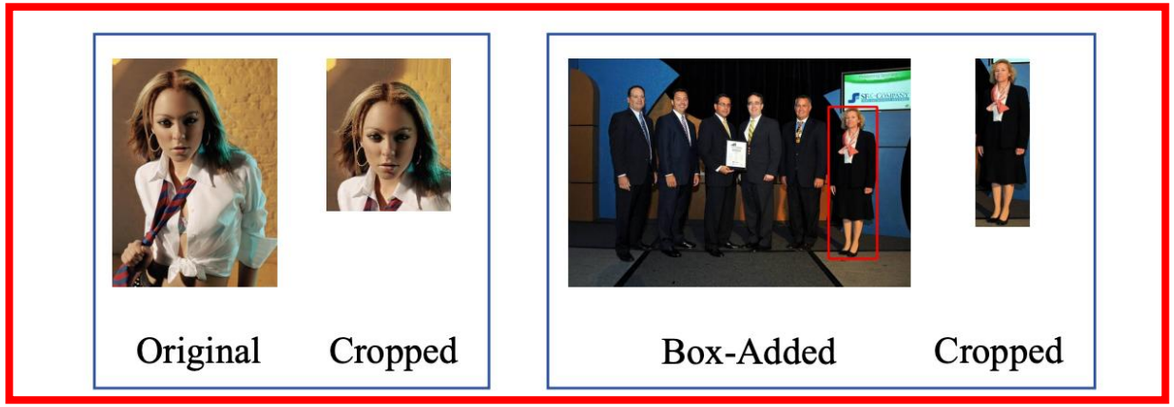
Answering Q1: MLLMs' Performance

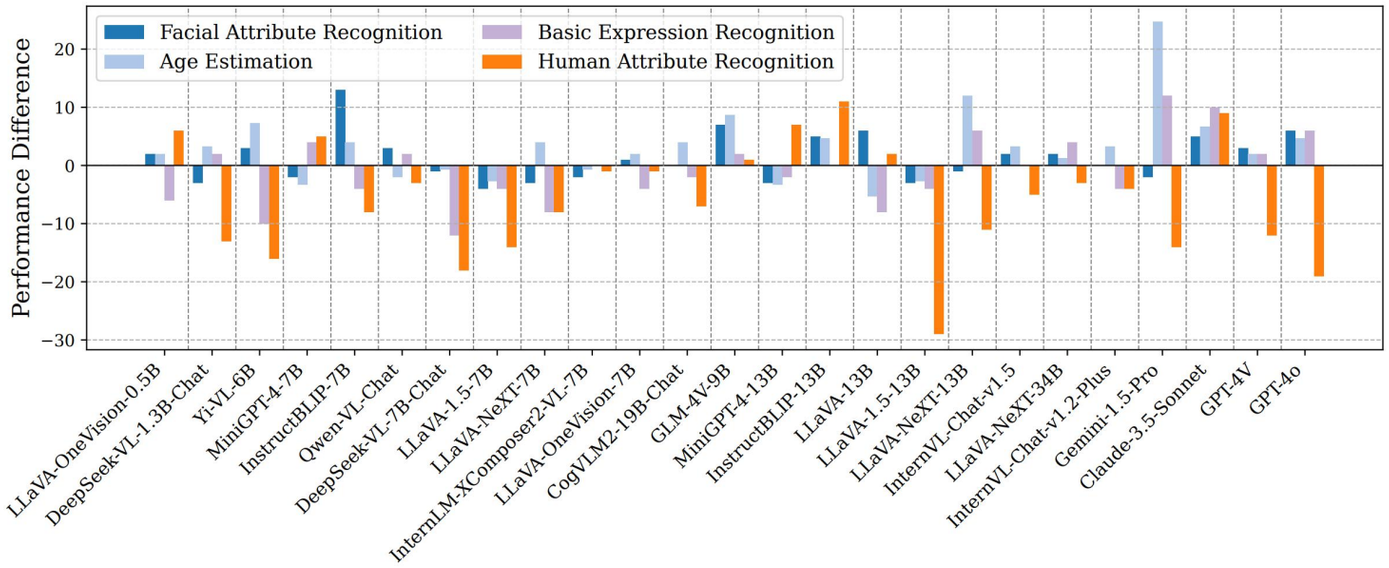
Ability Correlations



Answering Q1: MLLMs' Performance

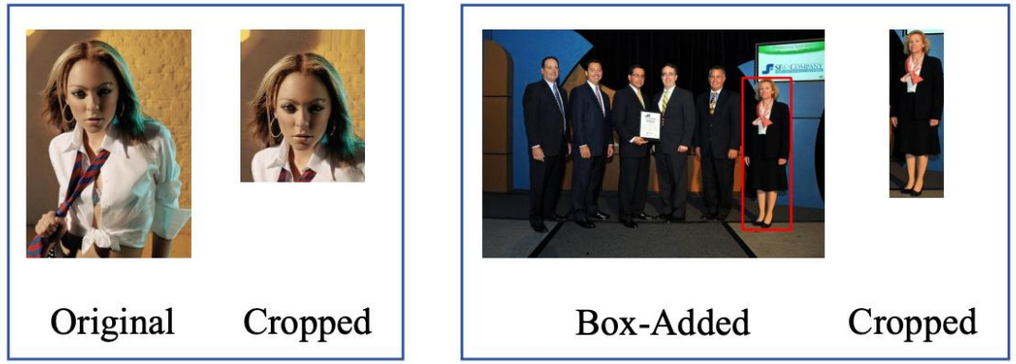
Relative Position of Targets

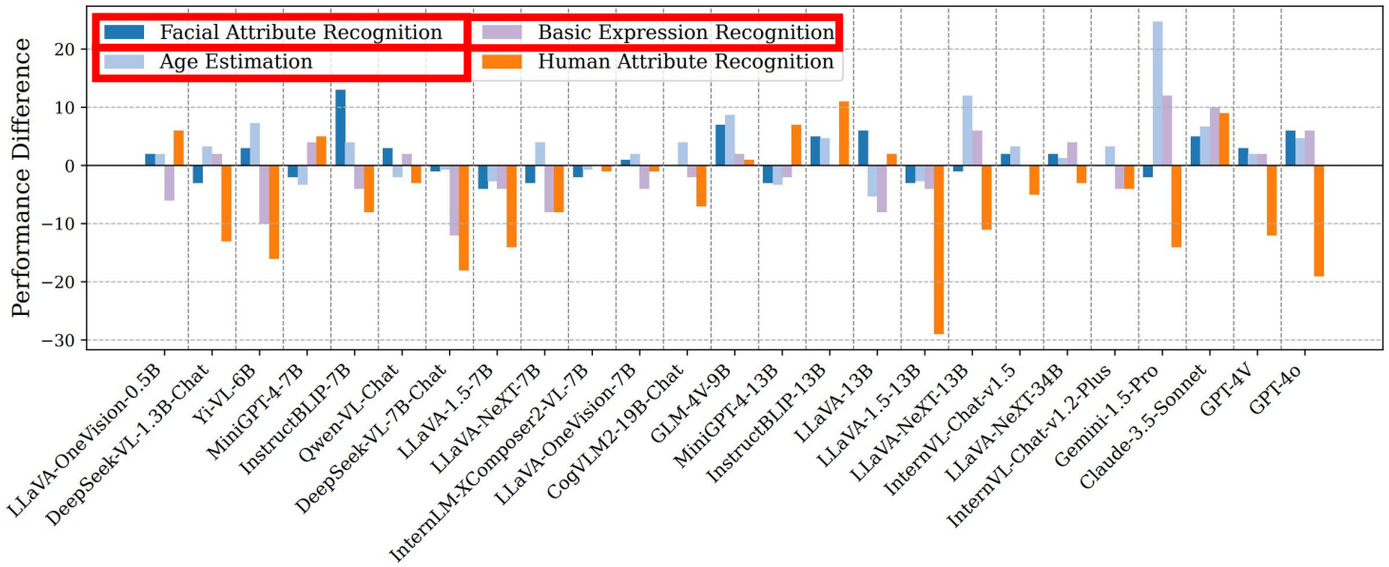




Answering Q1: MLLMs' Performance

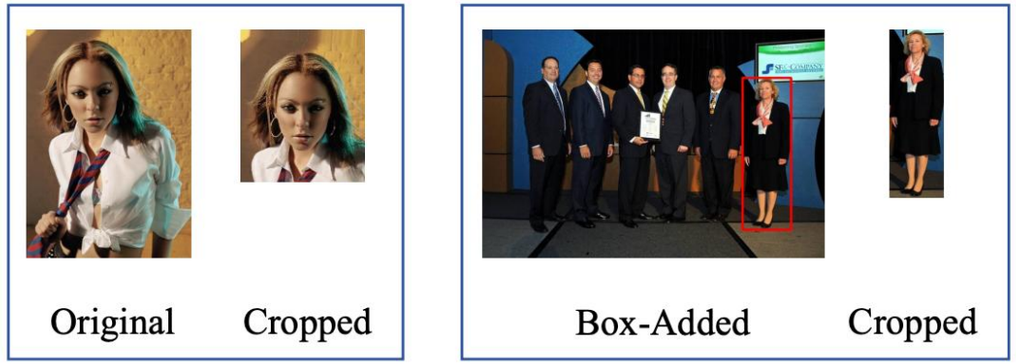
Relative Position of Targets

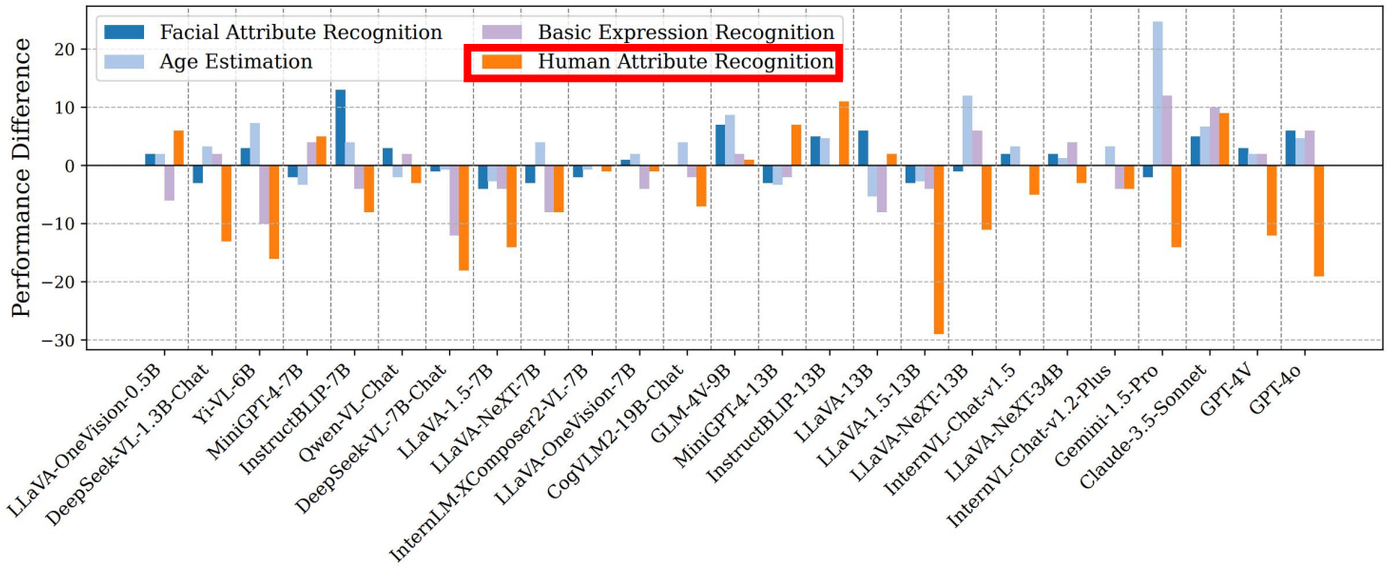




Answering Q1: MLLMs' Performance

Relative Position of Targets





Answering Q1: MLLMs' Performance

Hints & CoT Prompting

Setting	Format	Open-Source: InternVL-Chat-v1.2-Plus					Close-Source: GPT-4o				
		Face	Human	Per.	Rea.	Overall	Face	Human	Per.	Rea.	Overall
ZS	QO→A	69.7	83.1	76.7	76.0	76.4	68.5	71.6	68.9	71.7	70.0
H	QOH→A	68.4	83.2	76.4	75.9	75.9	72.2	74.6	70.4	78.0	73.4
H+VCoT	QOH→RA	69.1	82.5	75.9	74.8	75.7	76.4	80.7	78.2	77.2	78.6
H+1TCoT	QOH→RA	68.6	81.4	75.6	74.3	75.0	77.9	81.9	79.0	81.2	79.9
H+2TCoT	QOH→R, QOHR→A	69.1	79.1	75.8	71.8	74.1	77.0	81.2	78.4	77.2	79.1

Answering Q1: MLLMs' Performance

Hints & CoT Prompting

Setting	Format	Open-Source: InternVL-Chat-v1.2-Plus					Close-Source: GPT-4o				
		Face	Human	Per.	Rea.	Overall	Face	Human	Per.	Rea.	Overall
ZS	QO→A	69.7	83.1	76.7	76.0	76.4	68.5	71.6	68.9	71.7	70.0
H	QOH→A	68.4	83.2	76.4	75.9	75.9	72.2	74.6	70.4	78.0	73.4
H+VCoT	QOH→RA	69.1	82.5	75.9	74.8	75.7	76.4	80.7	78.2	77.2	78.6
H+1TCoT	QOH→RA	68.6	81.4	75.6	74.3	75.0	77.9	81.9	79.0	81.2	79.9
H+2TCoT	QOH→R, QOHR→A	69.1	79.1	75.8	71.8	74.1	77.0	81.2	78.4	77.2	79.1

Answering Q1: MLLMs' Performance

Hints & CoT Prompting

Setting	Format	Open-Source: InternVL-Chat-v1.2-Plus					Close-Source: GPT-4o				
		Face	Human	Per.	Rea.	Overall	Face	Human	Per.	Rea.	Overall
ZS	QO→A	69.7	83.1	76.7	76.0	76.4	68.5	71.6	68.9	71.7	70.0
H	QOH→A	68.4	83.2	76.4	75.9	75.9	72.2	74.6	70.4	78.0	73.4
H+VCoT	QOH→RA	69.1	82.5	75.9	74.8	75.7	76.4	80.7	78.2	77.2	78.6
H+1TCoT	QOH→RA	68.6	81.4	75.6	74.3	75.0	77.9	81.9	79.0	81.2	79.9
H+2TCoT	QOH→R, QOHR→A	69.1	79.1	75.8	71.8	74.1	77.0	81.2	78.4	77.2	79.1

Answering Q2: Specialist Models vs. MLLMs

Hints & CoT Prompting

L3 Ability	Age	Expression		Deepfake	Spoofing	Action	Counting
Dataset	UTKFace	RAF-DB (Basic)	RAF-DB (Compound)	FF++	SiW-Mv2	HICO-DET	ShTech-A
Matric	MAE ↓	ACC ↑	ACC ↑	ACC ↑	ACER ↓	mAP ↑	MAE ↓
Random	27.89	13.85	8.08	50.84	50.05	9.32	1512.65
InternVL-Chat-v1.5	6.43	72.23	42.93	56.21	14.84	22.29	2195.69
LLaVA-NeXT-34B	6.01	77.71	41.04	53.42	22.38	13.74	1592.55
InternVL-Chat-v1.2-Plus	5.21	76.40	30.56	52.89	19.92	12.25	2518.25
Best of The Above 3	5.21	77.71	42.93	56.21	14.84	22.29	1592.55
Early Specialist Model	5.47	74.20	44.55	82.01	9.40	19.81	110.20
Relative Score	1.01	1.06	0.96	0.17	0.87	1.24	-0.06
Need Specialist Model?	No.	No.	No.	Yes.	No.	No.	Yes.
L3 Ability	Basic FR	C.P. FR	C.A. FR	S.L. FR	Occ. FR		Re-ID
Dataset	LFW	CPLFW	CALFW	SLLFW	MLFW		Market1501
Matric	ACC ↑	ACC ↑	ACC ↑	ACC ↑	ACC ↑		ACC ↑[12]
Random	50.05	49.75	50.12	50.18	50.05		49.47
InternVL-Chat-v1.5	83.68	58.13	61.40	56.72	52.15		77.53
LLaVA-NeXT-34B	91.32	65.87	62.07	70.25	53.73		85.67
InternVL-Chat-v1.2-Plus	92.57	67.98	66.50	68.50	58.65		88.73
Best of The Above 3	92.57	67.98	66.50	70.25	58.65		88.73
Early Specialist Model	99.50	87.47	92.43	98.40	82.87		95.26
Relative Score	0.86	0.48	0.39	0.42	0.26		0.86
Need Specialist Model?	No.	Yes.	Yes.	Yes.	Yes.		No.

Answering Q2: Specialist Models vs. MLLMs

Hints & CoT Prompting

L3 Ability	Age	Expression		Deepfake	Spoofing	Action	Counting
Dataset	UTKFace	RAF-DB (Basic)	RAF-DB (Compound)	FF++	SiW-Mv2	HICO-DET	ShTech-A
Matric	MAE ↓	ACC ↑	ACC ↑	ACC ↑	ACER ↓	mAP ↑	MAE ↓
Random	27.89	13.85	8.08	50.84	50.05	9.32	1512.65
InternVL-Chat-v1.5	6.43	72.23	42.93	56.21	14.84	22.29	2195.69
LLaVA-NeXT-34B	6.01	77.71	41.04	53.42	22.38	13.74	1592.55
InternVL-Chat-v1.2-Plus	5.21	76.40	30.56	52.89	19.92	12.25	2518.25
Best of The Above 3	5.21	77.71	42.93	56.21	14.84	22.29	1592.55
Early Specialist Model	5.47	74.20	44.55	82.01	9.40	19.81	110.20
Relative Score	1.01	1.06	0.96	0.17	0.87	1.24	-0.06
Need Specialist Model?	No.	No.	No.	Yes.	No.	No.	Yes.

L3 Ability	Basic FR	C.P. FR	C.A. FR	S.L. FR	Occ. FR		Re-ID
Dataset	LFW	CPLFW	CALFW	SLLFW	MLFW		Market1501
Matric	ACC ↑	ACC ↑	ACC ↑	ACC ↑	ACC ↑		ACC ↑[12]
Random	50.05	49.75	50.12	50.18	50.05		49.47
InternVL-Chat-v1.5	83.68	58.13	61.40	56.72	52.15		77.53
LLaVA-NeXT-34B	91.32	65.87	62.07	70.25	53.73		85.67
InternVL-Chat-v1.2-Plus	92.57	67.98	66.50	68.50	58.65		88.73
Best of The Above 3	92.57	67.98	66.50	70.25	58.65		88.73
Early Specialist Model	99.50	87.47	92.43	98.40	82.87		95.26
Relative Score	0.86	0.48	0.39	0.42	0.26		0.86
Need Specialist Model?	No.	Yes.	Yes.	Yes.	Yes.		No.

Answering Q2: Specialist Models vs. MLLMs

Hints & CoT Prompting

L3 Ability	Age	Expression		Deepfake	Spoofing	Action	Counting
Dataset	UTKFace	RAF-DB (Basic)	RAF-DB (Compound)	FF++	SiW-Mv2	HICO-DET	ShTech-A
Matric	MAE ↓	ACC ↑	ACC ↑	ACC ↑	ACER ↓	mAP ↑	MAE ↓
Random	27.89	13.85	8.08	50.84	50.05	9.32	1512.65
InternVL-Chat-v1.5	6.43	72.23	42.93	56.21	14.84	22.29	2195.69
LLaVA-NeXT-34B	6.01	77.71	41.04	53.42	22.38	13.74	1592.55
InternVL-Chat-v1.2-Plus	5.21	76.40	30.56	52.89	19.92	12.25	2518.25
Best of The Above 3	5.21	77.71	42.93	56.21	14.84	22.29	1592.55
Early Specialist Model	5.47	74.20	44.55	82.01	9.40	19.81	110.20
Relative Score	1.01	1.06	0.96	0.17	0.87	1.24	-0.06
Need Specialist Model?	No.	No.	No.	Yes.	No.	No.	Yes.

L3 Ability	Basic FR	C.P. FR	C.A. FR	S.L. FR	Occ. FR		Re-ID
Dataset	LFW	CPLFW	CALFW	SLLFW	MLFW		Market1501
Matric	ACC ↑	ACC ↑	ACC ↑	ACC ↑	ACC ↑		ACC ↑[12]
Random	50.05	49.75	50.12	50.18	50.05		49.47
InternVL-Chat-v1.5	83.68	58.13	61.40	56.72	52.15		77.53
LLaVA-NeXT-34B	91.32	65.87	62.07	70.25	53.73		85.67
InternVL-Chat-v1.2-Plus	92.57	67.98	66.50	68.50	58.65		88.73
Best of The Above 3	92.57	67.98	66.50	70.25	58.65		88.73
Early Specialist Model	99.50	87.47	92.43	98.40	82.87		95.26
Relative Score	0.86	0.48	0.39	0.42	0.26		0.86
Need Specialist Model?	No.	Yes.	Yes.	Yes.	Yes.		No.

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks!

Project Homepage: https://face-human-bench.github.io/
You can also scan the QR code to visit it!