HUMAN-MME: A Holistic Evaluation Benchmark for Human-Centric Multimodal Large Language Models

Yuansen Liu[1*] , Haiming Tang[1*] , Jinlong Peng[2*] , Jiangning Zhang[2], Xiaozhong Ji[3]

Qingdong He[2], Donghao Luo[2], Zhenye Gan[2], Junwei Zhu[2], Yunhang Shen[2]

Chaoyou Fu[3], Chengjie Wang[2], Xiaobin Hu[1✉], Shuicheng Yan[1]

1 NUS National University of Singapore

2 Tencent Youtu Lab

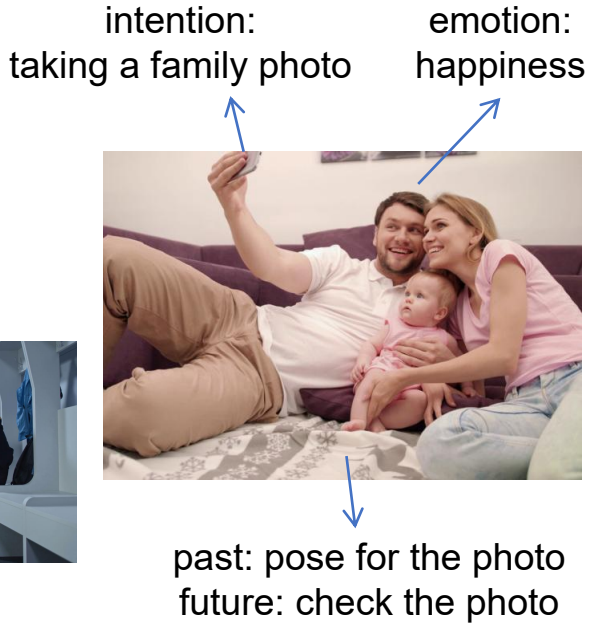
3 南京大学 NANJING UNIVERSITY

Human-MME

Motivation

Missing fine-grained details

Confusion in multi-person or multi-image questions



Limited capability in high-level intent and causality

Face Understanding

Please provide bounding box of *left eye*

Answer: [862, 250, 993, 315]

Please select facial feature of the person.

A. High cheekbones B. Double chin
C. Young face D. Mature face

Body Understanding

Please provide bounding box of *right hand*

Answer: [200, 500, 400, 600]

Give name and color of the *top* wearing

Answer: Judo uniform white

Select name and color of the *top* wearing

A. Mustard-Yellow sweater B. Deep Grey Judo uniform
C. Judo uniform white D. White shirt

Human-Object Interaction Understanding

Which one best describe human-object interaction in image

A. Right hand, holding, green marker B. Right hand, holding, pencil
C. Left hand, holding, green marker D. Left hand, holding, pencil

Please provide bounding box of the object with interaction: *left hand, holding*

Answer: [200, 500, 400, 600]

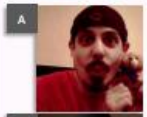
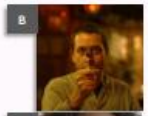
Please name the object that has interaction: *left hand, holding*, with the person

Answer: Pencil

Multi-Image Understanding

Provide a sequence of images provided based on the count of feature present

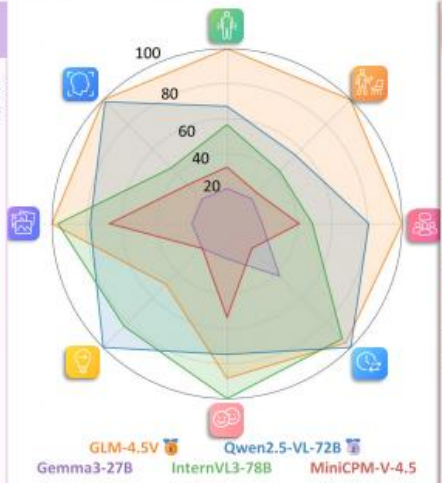
- *Has mustache*
- *Big lips*
- *Wearing hat*

Answer: A – C – D – B

Please select the image fits the description of human-object interaction *face, sip, wine glass*


Answer: A



Multi-Person Reasoning

What is the *age group* of the person who is *wearing beige jumpsuit*

Answer: Child

Please give the bounding box of *whole visible body* for the person who has *no receding hairline*

Answer: [428, 224, 842, 986]

If there is someone who *is Asian* and *wearing light blue shirt* please give the *name and color of the person's bottom wearing*, else, give "unknown"

Answer: unknown

If there is someone who *is male* and *have no oval face* please give the bounding box of the one's *mouth*, else, give [-1,-1,-1,-1]

Answer: [511, 514, 500, 575]

The person *looking downward* has interaction: *right hand, holding* with an object. Name it and give its bounding box

Answer: Spoon, [306, 516, 460, 783]

Select the option that fits everyone in the image

A. wearing blue shirt
B. happy
C. child
D. male

Intention Discrimination

Please select the best analysis of intention for someone in the image

A. ... expressing enthusiasm ... in a celebratory ... activity through deliberate posing and vibrant attire
B. The individual is preparing for an outdoor adventure with companions and pets ...
C. The individual is expressing ... relaxation while engaging with a natural outdoor setting ...
D. The individual is expressing joy and contentment while embracing a relaxed and natural ...

Emotion Discrimination

Please select the best analysis of emotion for someone in the image

A. A deep sense of joy and contentment ...
B. A deep sense of concern and focused determination ... navigating an unexpected challenge ...
C. Deep focus and intense concentration on complex information requiring careful analysis ...
D. Feeling overwhelmed by unexpected information experiencing a mix of concern and frustration while trying to process and regain control

Causal Discrimination

Please select the best analysis of what happened in the past and what will happen in the future

A. The individual sets down the instrument, takes a deliberate breath, then flips through the pages ...
B. The figure moved steadily along a winding trail ...
C. The individual had carefully arranged the performance setup, positioning the instrument ...
D. The hand moves with deliberate calm, tracing lines ...

Answer: Past is C, Future is A

Question Component
Evaluation Metric

Choice — Accuracy

Short-Answer — BERT+Embedding+Keyword

Bounding Box — IoU

Ranking — Kendall's τ

Judgment — F1 Score

NUS National University of Singapore

Tencent Youtu Lab

南京大學 NANJING UNIVERSITY

Data Curation Pipeline

Human-MME

2.1 Data Collection

Free stock media sites (pexels, pixabay)

Opensource Dataset: HICO-DET → More than one person

YoloV11 Head Detection / Body Detection → Image hash De-duplicate

Image set to Be processed

2.2 Automated Annotation Pipeline

Step 1: Body Box

YOLOv11

Face Bounding Box F_i

Body Bounding Box B_i

DWPose Pose Estimation

Body Part Bounding Box $P_i^{left\ hand}, P_i^{right\ foot}, \ldots$

Step 2: General Attribute, Wearing and Object

Qwen2.5-VL

Q: Information about person in the bounding box? {template}

General

Age	Adult
Emotion	Neutral
Gender	Female
Race	Asian

General Attributes $A_i^{age}, A_i^{gender}, \ldots$

Wearing type, color, name

Top	Black white	Apron
Top	Light grey	Shirt

Wearing $W_i^{(j)}$

HOI body part, action, object

Single hand	holding	Flour bag

Object Interactions $O_i^{(j)}$

Step 3: Object Box

Prompt: "Flour bag" → Grounding DINO → Bounding Box of HOI Object

Step 4: Face Attribute and Box

Cropped → FaceXFormer

Face Attributes

Young	True
Mouth Open	True
Mustache	True
Bald	False
...	

Facial Attributes $FA_i^{mustache}, FA_i^{notbald}, FA_i^{young}, \ldots$

Face Landmarks

Face Part Bounding Box $FP_i^{lefteye}, FP_i^{nose}, \ldots$

Step 5: Emotion, Intention and Cause

Person i in image → Qwen2.5-VL: Comprehensive description from various aspects → Qwen3: Identity-neutral descriptions

Emotion E_i

Intention I_i

Cause C^{past}, C^{future}

2.3 Manual Adjustment

De-duplication

best automatic annotation: HOI annotated / No HOI found

highest readability: Clear hands / Blurry hands

highest complexity: 2 persons / 1 person

Correction

Add Annotation: No HOI → Right hand, Hold, Sign

Modify Annotation: Black hat → Brown hat

Delete Annotation: Persons on photos on the wall are removed

NUS National University of Singapore

Tencent Youtu Lab

南京大學 NANJING UNIVERSITY

Benchmarking Results

Human-MME

Model	FU	BU	HU	MIU	MPR	ID	CD	ED	Avg.	Bounding Box	Choice	Short-Answer	Ranking	Judgment
GLM-4.5V	**61.6**	**77.4**	**82.5**	79.2	**71.5**	83.9	85.4	66.6	**76.0**	**66.3**	70.8	**83.5**	86.2	68.3
GLM-4.1V-9B	55.2	74.1	69.5	71.8	64.3	82.7	76.0	58.8	69.1	49.7	68.0	80.7	82.5	66.3
Qwen2.5-VL-72B	61.1	70.2	70.6	75.4	65.2	**88.1**	**86.3**	65.3	72.8	50.8	70.4	81.7	83.9	**71.3**
Qwen2.5-VL-32B	56.2	73.3	65.3	70.7	58.2	82.9	81.1	64.9	69.1	44.9	67.9	72.7	82.4	67.0
Qwen2.5-VL-7B	49.4	68.4	61.4	61.0	46.3	84.1	72.1	60.9	63.0	31.7	60.1	71.0	70.7	56.5
Intern-S1	41.0	65.2	65.5	**79.8**	59.3	82.9	83.2	**68.3**	68.2	22.1	**72.6**	82.0	**86.6**	68.9
InternVL3.5-241B	50.7	74.6	71.4	76.4	59.4	83.7	82.5	66.4	70.6	46.3	68.9	81.3	84.0	57.3
InternVL3-78B	43.4	67.9	67.2	78.6	54.6	86.7	84.7	67.7	68.9	26.6	70.9	82.9	85.2	61.6
InternVL3.5-38B	44.6	72.6	64.6	75.0	53.8	86.9	78.0	65.6	67.6	30.6	67.9	80.7	82.6	62.0
Llama-4-Scout	27.3	50.6	49.4	48.9	33.9	66.5	57.1	50.4	48.0	6.4	47.9	69.5	71.0	38.6
LLaVA-NeXT-72B	38.0	66.8	65.1	54.8	47.2	77.0	70.5	54.6	59.3	29.8	58.2	72.8	61.1	52.3
Aya-vision-32B	30.9	57.2	57.1	67.9	42.8	76.2	71.8	57.4	57.7	8.9	57.7	78.5	75.7	53.8
Gemma3-27B	35.1	59.9	61.2	65.3	45.1	81.5	73.0	60.1	60.2	13.8	59.4	78.4	75.4	54.5
Kimi-VL-A3B	37.3	63.1	50.8	27.3	42.6	81.0	63.1	55.3	52.6	17.2	56.7	72.1	63.2	50.4
MiniCPM-V-4.5	38.9	62.6	62.4	73.5	52.1	81.5	67.8	63.3	62.8	20.2	65.0	80.7	84.0	57.9
Phi-4	29.5	48.1	48.6	39.6	29.6	62.9	38.1	46.4	42.9	5.5	45.5	62.4	54.4	19.8
Gemini-2.5-Pro	**42.4**	66.5	70.0	**83.6**	**48.6**	79.4	86.1	**64.5**	**67.6**	23.5	**72.4**	**83.9**	**90.9**	**72.0**
GPT-4o	28.8	58.8	59.8	74.7	34.4	79.2	76.2	52.7	58.1	11.5	57.6	78.3	83.8	48.6
GPT-5	34.4	**67.8**	**71.1**	75.8	43.1	**82.3**	**89.2**	42.6	63.3	**25.2**	65.8	77.5	85.8	41.7
GPT-5-mini	30.6	66.3	67.4	76.4	41.3	79.4	81.7	39.9	60.4	21.1	63.4	76.9	87.4	49.1

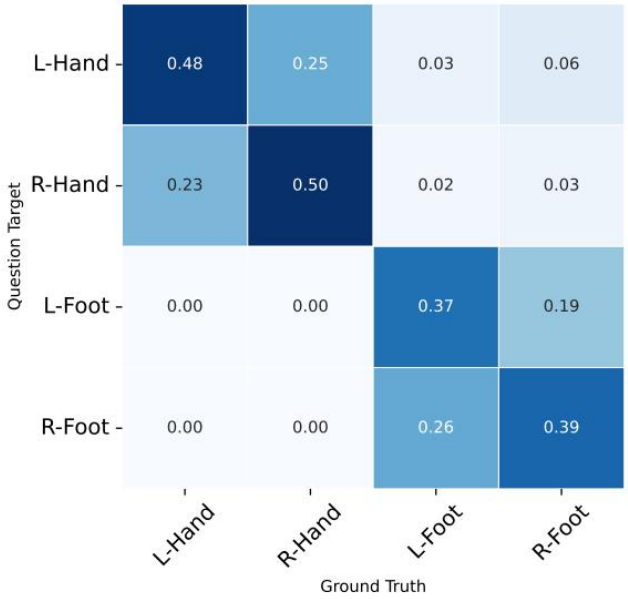
NUS National University of Singapore

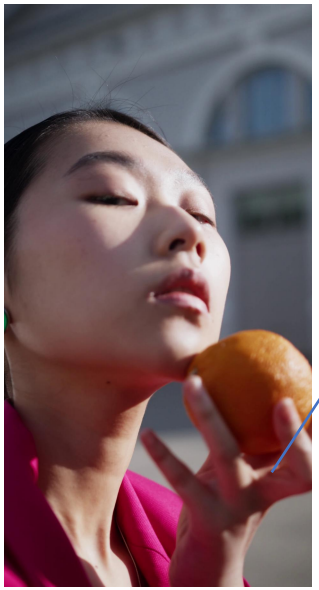
Tencent Youtu Lab

南京大學 NANJING UNIVERSITY

Human-MME

Key Findings



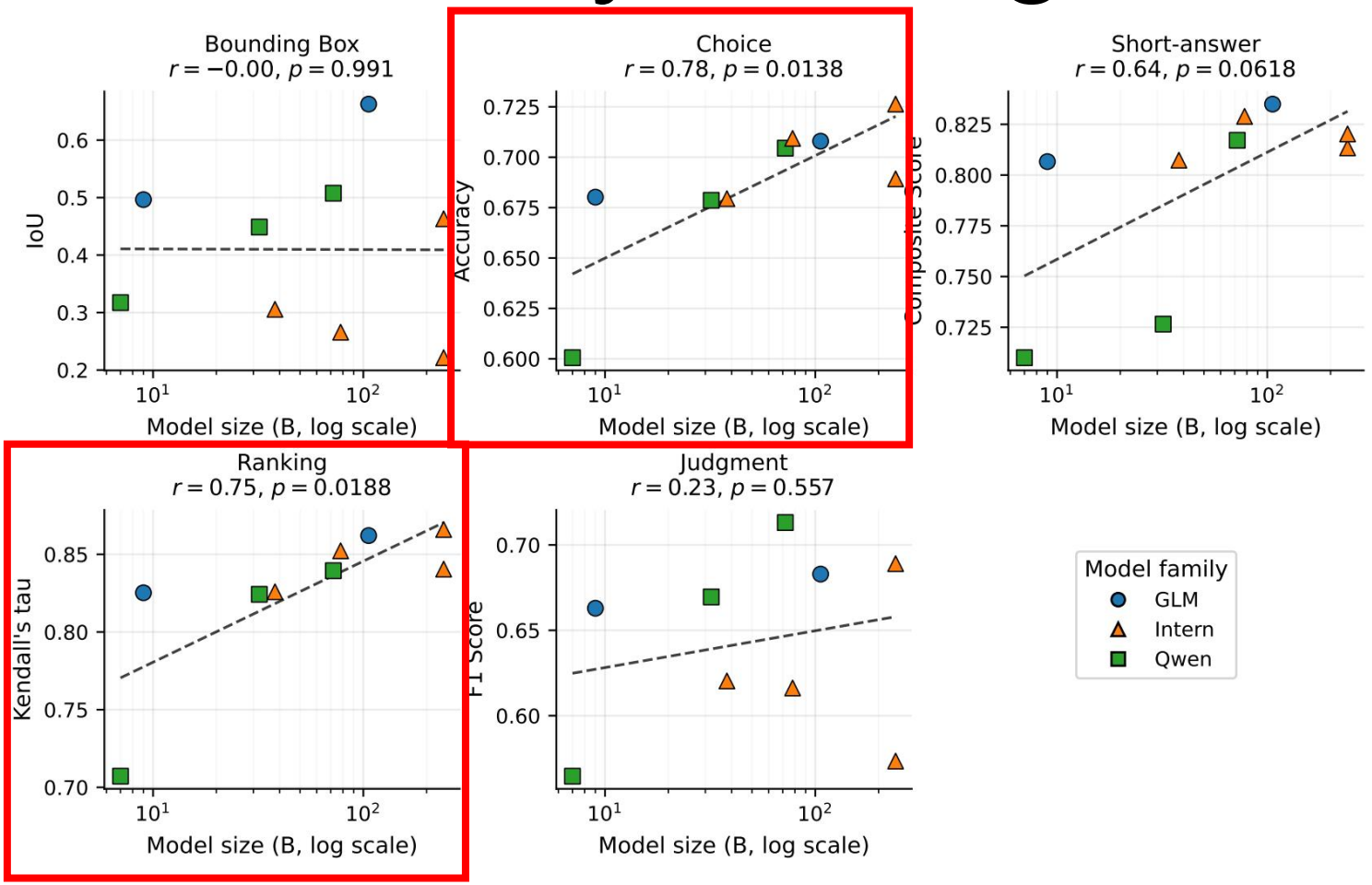


Right hand?
Left hand!

Challenges in left-right discrimination for body parts

Human-MME

Key Findings



Stronger scaling effects in Choice and Ranking tasks

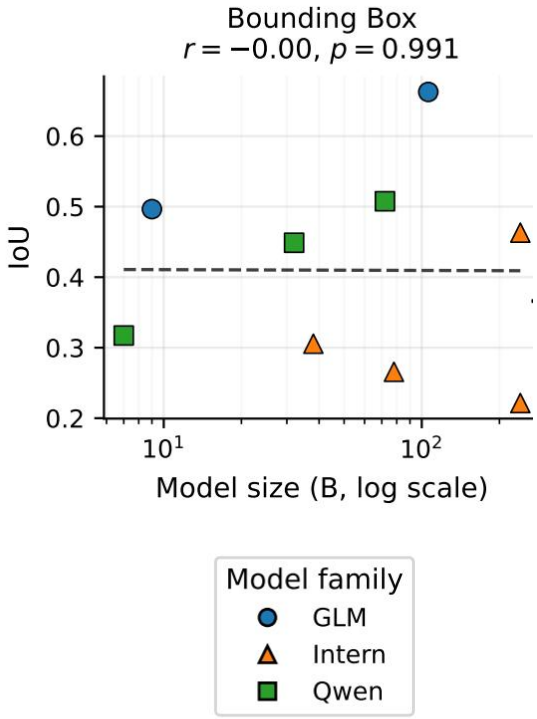
NUS National University of Singapore

Tencent Youtu Lab

南京大學 NANJING UNIVERSITY

Human-MME

Key Findings



Model	Vision Encoder	Language Model	Image Encoding	Grounding Training
GLM-4.5V (106B)	AIMv2-Huge	GLM-4.5-Air	dynamic tokenization	xyxy-format bounding box with dedicated tokens
GLM-4.1V-9B	AIMv2-Huge	GLM-4-9B	dynamic tokenization	xyxy-format bounding box with dedicated tokens
Qwen2.5-VL-72B	Redesigned ViT	Qwen2.5-72B	dynamic tokenization	xyxy-format bounding box with JSON format alignment
Qwen2.5-VL-32B	Redesigned ViT	Qwen2.5-32B	dynamic tokenization	xyxy-format bounding box with JSON format alignment
Qwen2.5-VL-7B	Redesigned ViT	Qwen2.5-7B	dynamic tokenization	xyxy-format bounding box with JSON format alignment
InternVL3-78B	InternViT-6B	Qwen2.5-72B	448×448 → 256	with image grounding training data
InternVL3.5-38B	InternViT-6B	Qwen3-32B	448×448 → 256	with image grounding training data
InternVL3.5-241B	InternViT-6B	Qwen3-235B	448×448 → 256	with image grounding training data
Intern-S1 (241B)	InternViT-6B	Qwen3-235B	448×448 → 256	None
MiniCPM-V-4.5 (8B)	SigLIP2-400M	Qwen3-8B	448×448 → 256	None
Gemma3-27B	SigLIP-400M	Dec-only Transf.	896×896 → 256 (whole image)	None
Aya-vision-32B	SigLIP2-400M	Aya Expanse 32B	364×364 → 169	None
LLaVA-NeXT-72B	CLIP-Large	Qwen1.5-72B	336×336 → 576	None
Llama-4-Scout	MetaCLIP	Llama MoE	dynamic tokenization	None
Kimi-VL-A3B (16B)	MoonViT	Moonlight MoE	dynamic tokenization	With GUI-focused grounding training data
Phi-4 (6B)	SigLIP-400M	Phi-4-Mini	384×384 → 729 (whole image)	None

Training data influence on grounding tasks

NUS National University of Singapore
Tencent Youtu Lab
南京大學 NANJING UNIVERSITY

Thank you!

HUMAN-MME: A Holistic Evaluation Benchmark for Human-Centric Multimodal Large Language Models

github.com/Yuan-Hou/Human-MME

arxiv.org/abs/2509.26165

National University of Singapore

Tencent Youtu Lab

NANJING UNIVERSITY