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
{lxqin, xuweiran}@bupt.edu.cn

Abstract

Faces and humans are crucial elements in social interaction and are widely included in everyday photos and videos. Therefore, a deep understanding of faces and humans will enable multi-modal assistants to achieve improved response quality and broadened application scope. Currently, the multi-modal assistant community lacks a comprehensive and scientific evaluation of face and human understanding abilities. In this paper, we first propose a hierarchical ability taxonomy that includes three levels of abilities. Then, based on this taxonomy, we collect images and annotations from publicly available datasets in the face and human community and build a semi-automatic data pipeline to produce problems for the new benchmark. Finally, the obtained Face-Human-Bench includes a development set and a test set, each with 1800 problems, supporting both English and Chinese. We conduct evaluations over 25 mainstream multi-modal large language models (MLLMs) with our Face-Human-Bench, focusing on the correlation between abilities, the impact of the relative position of targets on performance, and the impact of Chain of Thought (CoT) prompting on performance. We also explore which abilities of MLLMs need to be supplemented by specialist models. The dataset and evaluation code have been made publicly available at `https://face-human-bench.github.io/`.

1 Introduction

Faces and humans are always the most crucial elements of photos and videos in our everyday lives. Consequently, they are also critical focuses in multi-modal AI applications. In the past two years, ChatGPT [44] and GPT-4 [45] have achieved great success with impressive instruction-following and multi-modal understanding capabilities respectively. Numerous excellent works [34, 78, 10, 3] from the open-source community have followed, collectively presenting the immense potential of multi-modal assistants. Since faces and humans are central to social interaction, a deep understanding of this information can make multi-modal assistants achieve improved response quality and broadened application scope. For instance, in movie understanding [69, 20, 57], identifying characters is a prerequisite for multi-modal assistants to describe the plot accurately. In multi-modal human-computer interaction [16], perceiving expressions and body language can help multi-modal assistants accurately understand the context, generating more personalized and humanized responses. In media forensics [36, 47, 72], determining whether deepfake artifacts exist on a face is crucial for multi-modal assistants to detect misinformation.

[*]Equal Contribution.
[†]Corresponding Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/085713-0322

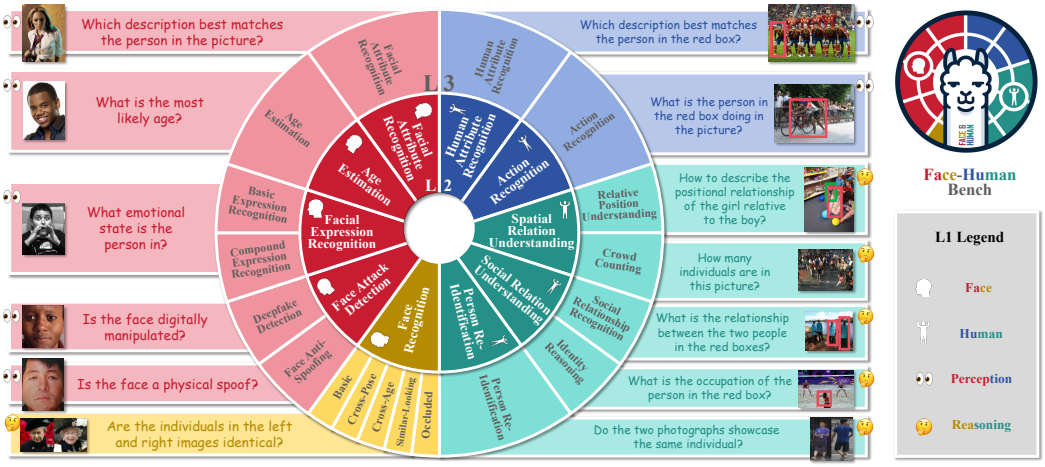


Figure 1: The three-level ability taxonomy for evaluating face and human understanding abilities. We construct the Face-Human-Bench based on this taxonomy. The proportion of the sectors represents the weight of the corresponding abilities in the overall score on the Face-Human-Bench.

Comprehensive and scientific evaluation is the foundation for researching applications of multi-modal assistants related to "faces and humans." Existing benchmarks [15, 26, 40] for large multi-modal models typically involve limited abilities of face and human understanding, such as celebrity recognition, action recognition, identity reasoning, and social relation, leaving many important abilities unexplored. On the other hand, since face and human understanding is one of the earliest research topics in artificial intelligence, there are numerous datasets available for evaluating the performance of specialist models. The images and annotations from these datasets can serve as original material to evaluate multi-modal assistants.

As the starting point of our evaluation, we propose a hierarchical ability taxonomy, as shown in Figure 1. This taxonomy consists of three levels. Level-1 (L1) has two perspectives to study: from the target perspective, L1 includes face understanding and human understanding; from the cognitive process perspective, L1 includes perception and reasoning. Subsequently, we incorporate finer-grained abilities into the taxonomy and categorize them into 10 Level-2 (L2) and 18 Level-3 (L3) ability dimensions. Then, based on this taxonomy, we collect datasets from the face and human community and use a semi-automatic data pipeline to transform original images and annotations into multi-modal QAs. The final obtained benchmark called Face-Human-Bench, including a development set and a test set, each with 1800 problems, supporting evaluations in both English and Chinese. For ease of evaluation, we adopt multiple-choice as the problem format following MMBench [40] and SEED-Bench [26].

Our study aims to provide readers with the following insights: **Q1**: *How do existing Multi-modal Large Language Models (MLLMs) perform in the face and human understanding?* Specifically, we focus on (a) the performance of 25 mainstream MLLMs, (b) the correlation between abilities at different levels, (c) the impact of the relative position of targets on performance, and (d) the impact of Chain of Thought (CoT) [74] prompting on performance. **Q2**: *In the field of face and human understanding, which tasks' specialized models achieve significantly better performance than current MLLMs?* With these insights, researchers in the multi-modal community can strategically strengthen the deficient abilities in their general-purpose models. Moreover, researchers in the face and human understanding community can use our evaluation results to select suitable MLLMs as initialization for downstream application scenarios. Knowing the specific abilities that MLLMs lack, researchers can also strategically leverage the outputs of specialized models to construct multi-modal agents [62, 65] that enhance responses.

In response to Q1, our main findings are as follows: (a) The Face-Human-Bench can effectively distinguish the abilities of MLLMs in faces and human understanding. Under the zero-shot setting, the best-performing closed-source model, GPT-4o [46], does not perform as well as the best open-source model, InternVL-Chat-v1.2-Plus [7]. (b) The correlation coefficients can reveal correlations between abilities at different levels. At L2 and L3, there are some ability groups in which the ability dimensions exhibit significant positive correlations between each pair. (c) Many models show

https://doi.org/10.52202/085713-0322

substantial performance differences on the same task with different relative positions of targets. We use a new metric called the relative position sensitivity score (RPSS) to measure this phenomenon. On this metric, InternLM-XComposer2-VL-7B [13] performs the best, indicating that its performance is almost unaffected by the relative position of targets. (d) Introducing hints and CoT instructions into the prompts significantly improves the performance of the closed-source model GPT-4o, but has no effect on the open-source model, InternVL-Chat-v1.2-Plus. In response to Q2, we find that in tasks of deepfake detection, crowd counting, and face recognition (under challenging scenarios), the performance of MLLMs is significantly inferior to that of corresponding specialist models.

Our contributions can be summarized as follows:

- We propose the Face-Human-Bench, the first benchmark dedicated to evaluating multi-modal assistants' face and human understanding abilities. It is based on a three-level ability taxonomy and supports both English and Chinese.
- Utilizing the Face-Human-Bench, we conduct a comprehensive evaluation of mainstream MLLMs, revealing the correlation between abilities, and exploring the impact of the relative position of targets and CoT prompting on the performance of MLLMs.
- We explore which face and human understanding tasks see specialist models significantly outperform MLLMs, providing recommendations for downstream application scenarios.

2 Face-Human-Bench

2.1 Hierarchical Ability Taxonomy

As shown in Figure 1, the proposed ability taxonomy includes three levels. Level 1 (L1) has two research perspectives. From the target perspective, L1 includes face understanding and human understanding. From the cognitive process perspective, L1 includes perception and reasoning. In our evaluation, perception involves direct comprehension of only one target, while reasoning requires synthesizing information from multiple targets and environments to conclude. There are ten abilities in total at Level 2 (L2). Five are focused on faces: facial attribute recognition, age estimation, facial expression recognition, face attack detection, and face recognition, and five are focused on humans: human attribute recognition, action recognition, spatial relation understanding, social relation understanding, and person re-identification. It should be noted that at L2, there are 6 abilities under perception and 4 abilities under reasoning. Level 3 (L3) further refines the ability dimensions at L2. Facial expression recognition can be categorized into basic and compound types. Face attack detection includes deepfake detection and face anti-spoofing. Face recognition involves five scenarios: basic, cross-pose, cross-age, similar-looking, and occluded. Spatial relation understanding concerns relative position and count. Social relation understanding includes social relationship recognition and identity reasoning. Please refer to Section A.1 for detailed definitions and examples of these abilities.

2.2 Semi-Automatic Data Pipeline

Based on the hierarchical ability taxonomy defined in Section 2.1, we collect 16 public datasets from the face and human community, covering each L3 ability. Then, we employ a semi-automatic data pipeline to produce problems for the Face-Human-Bench.

An original sample S_i from public datasets can be represented as a binary tuple (I_i, L_i), where I_i denotes an original image set and L_i denotes an original label set. Note that we use "image set" and "label set" to describe the composition of one sample because, in some datasets, a single sample may consist of multiple images or labels. For instance, in face recognition, a sample includes a pair of face images to verify identity, and in facial attribute recognition, a sample may involve 40 attribute labels.

For ease of evaluation, we adopt multiple-choice as the problem format in our Face-Human-Bench. Each problem P_i corresponds to a quadruple (V_i, Q_i, O_i, A_i). Here, V_i refers to the images obtained via the image processing pipeline $p_{image} : \mathbb{I} \rightarrow \mathbb{V}$. p_{image} performs an operation such as cropping, concatenating, adding boxes, or leaving the original images unchanged, depending on the ability to test. Q_i denotes the question. Each L3 ability includes a set of pre-written questions that share the same semantics but exhibit diversity. When producing samples, a question Q_i is randomly selected from this question set. O_i is the set of n options $(o_1, o_2, ..., o_n)$, where $2 \leq n \leq 4$. These options are obtained through the text processing pipeline $p_{text} : \mathbb{L} \rightarrow \mathbb{O}$. p_{text} converts the original labels into

 https://doi.org/10.52202/085713-0322

one correct option and $n-1$ incorrect options. For some tasks, ChatGPT [44] is used within p_{text} to assist in generating incorrect options or adjusting options at the sentence level (fixing grammar or re-wording sentences for fluency). A_i is the correct answer to the problem. The produced P_i will be checked by data reviewers to ensure that options are unambiguous and there is one and only one correct answer. Problems that do not meet the requirements will be removed.

In summary, our semi-automatic data pipeline leverages image and text processing pipelines, p_{image} and p_{text}, to transform original samples into multiple-choice format problems. These problems are then manually checked to ensure quality. We obtain a benchmark with a development set of 1800 problems for the MLLM community to evaluate during training and a test set of 1800 problems for the formal evaluation in our paper. Additionally, the English problems are translated into Chinese to create a Chinese version of the benchmark. For more details on data sources, statistics, and the semi-automatic data pipeline, please refer to Sections A.2 and A.3.

3 Experiment

3.1 Experimental Setup

We use the weighted accuracy of multiple-choice problems as the evaluation score. As shown in Figure 1, the proportion of the sectors represents the weight of the corresponding abilities in the overall score on the Face-Human-Bench. Note that we set equal weights for each L2 ability.[3] To prevent models from favoring certain option letters over others, we shuffle the options to ensure the correct answers are evenly distributed across all option letters. During the testing, we add some constraint instructions to ensure MLLMs output only option letters as much as possible.[4] After obtaining the MLLM's response, we use regular expressions to extract the option letters. If this fails, we follow the implementation of MMBench [40] using ChatGPT [44] to extract the choices.[5] We evaluate 25 MLLMs (as shown in Table 1) in different sizes from 13 model families. For more details on these models, please refer to Section B.1.

3.2 Main Results

Table 1 shows the performance of all evaluated MLLMs at different levels of abilities on the Human-Face-Bench (English)[6] under the zero-shot setting. Overall scores range from 27.9% to 76.4%, demonstrating the effectiveness of the Face-Human-Bench in distinguishing the abilities of MLLMs in face and human understanding. We visualize the overall scores of MLLMs in Figure 2. The findings can be summarized as follows: (1) The top 3 performing open-source models in terms of the overall score are InternvL-Chat-v1.2-Plus, LLaVA-Next-34B, and InternVL-Chat-v1.5. These models' LLMs have the largest number of parameters among all open-source models we evaluate. (2) Generally, open-source models within the same series tend to show improved performance with increasing parameter scale. However, there are exceptions; for instance, the 13B version of LLaVA-1.5 and LLaVA-Next perform slightly worse than their 7B counterparts. (3) Under the zero-shot setting, the best closed-source model, GPT-4o, does not surpass the performance of the top-performing open-source models. We believe this is because GPT-4o does not fully realize its potential under the zero-shot setting. The experiments in

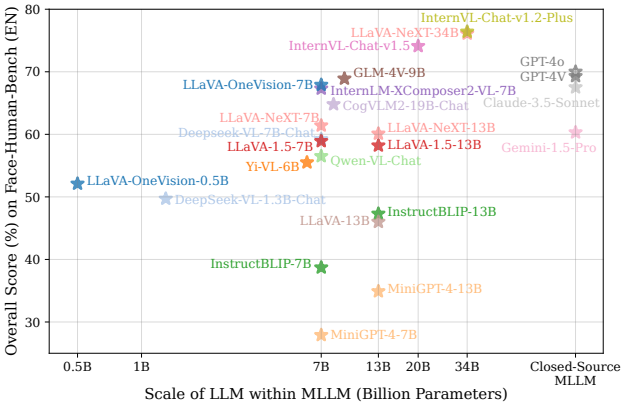


Figure 2: The leaderboard of MLLMs on our proposed Face-Human-Bench (English).

[3]For detailed weights of each subset, please refer to Section A.2.

[4]For the prompt template, please refer to Section B.2.1.

[5]For the prompt for choice extraction, please refer to Section B.2.2.

[6]For the results of the Chinese version, please refer to Section C.2.

https://doi.org/10.52202/085713-0322

Table 1: Zero-shot scores of MLLMs on the hierarchical Face-Human-Bench (EN). The highest scores for open-source and closed-source MLLMs are marked in blue and green respectively. The scores in the "random" row are theoretical values. For convenience, the names of the various tasks are abbreviated, where the abbreviations are formed by retaining the initial letters of phrases or the first few letters of key terms; their full names can be referred to in Figure 1.

Model	Face Understanding													
	Attr.	Age	Expression			Attack Detection			Face Recognition					
			Basic	Comp.	Mean	DFD	FAS	mean	Basic	C.P.	C.A.	S.L.	Occ.	Mean
Random	25.0	25.0	25.0	25.0	25.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
LLaVA -OneVision-0.5B [25]	36.0	43.0	71.0	60.0	65.5	46.0	55.0	50.5	50.0	42.0	44.0	50.0	38.0	44.8
DeepSeek -VL-1.3B-Chat [42]	36.5	49.0	57.0	50.0	53.5	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Yi-VL-6B [68]	75.5	51.7	65.0	52.0	58.5	34.0	43.0	38.5	50.0	48.0	48.0	50.0	44.0	48.0
MiniGPT-4-7B [78]	24.0	17.7	26.0	24.0	25.0	31.5	40.5	36.0	38.0	56.0	44.0	48.0	34.0	44.0
InstructBLIP-7B [10]	39.5	36.7	38.0	40.0	39.0	50.5	53.0	51.8	52.0	58.0	48.0	52.0	54.0	52.8
Qwen-VL-Chat [3]	55.5	49.7	65.0	50.0	57.5	51.0	54.0	52.5	66.0	52.0	54.0	58.0	54.0	56.8
DeepSeek -VL-7B-Chat [42]	57.5	52.3	68.0	58.0	63.0	46.0	53.0	49.5	54.0	52.0	50.0	48.0	50.0	50.8
LLaVA-1.5-7B [32]	61.0	49.3	62.0	58.0	60.0	55.5	55.0	55.3	54.0	52.0	50.0	56.0	50.0	52.4
LLaVA-NeXT-7B [33]	69.5	50.0	72.0	62.0	67.0	59.5	58.5	59.0	62.0	50.0	48.0	56.0	50.0	53.2
InternLM -XComposer2-VL-7B [13]	92.0	53.0	76.0	68.0	72.0	41.0	54.0	47.5	54.0	54.0	50.0	56.0	36.0	50.0
LLaVA -OneVision-7B [25]	90.5	60.3	74.0	62.0	68.0	35.0	56.0	45.5	58.0	42.0	34.0	42.0	34.0	42.0
CogVLM2-19B-Chat [23]	75.0	57.3	71.0	70.0	70.5	37.0	51.0	44.0	66.0	36.0	44.0	46.0	48.0	48.0
GLM-4V-9B [23]	79.5	55.7	79.0	74.0	76.5	46.0	50.0	48.0	68.0	54.0	54.0	62.0	52.0	58.0
MiniGPT-4-13B [78]	20.5	24.3	35.0	26.0	30.5	49.5	37.5	43.5	52.0	46.0	42.0	46.0	48.0	46.8
InstructBLIP-13B [10]	25.5	38.3	50.0	42.0	46.0	57.5	52.0	54.8	48.0	52.0	52.0	50.0	52.0	50.8
LLaVA-13B [34]	32.0	40.7	56.0	30.0	43.0	55.0	54.0	54.5	52.0	60.0	52.0	40.0	52.0	51.2
LLaVA-1.5-13B [32]	75.5	58.7	72.0	54.0	63.0	51.0	54.0	52.5	54.0	48.0	54.0	48.0	50.0	50.8
LLaVA-NeXT-13B [33]	77.5	46.7	71.0	52.0	61.5	50.0	54.0	52.0	58.0	54.0	54.0	56.0	56.0	55.6
InternVL-Chat-v1.5 [7]	92.0	61.7	72.0	68.0	70.0	71.5	67.0	69.2	90.0	60.0	60.0	60.0	52.0	64.4
LLaVA-NeXT-34B [33]	95.0	58.7	80.0	62.0	71.0	63.5	60.5	62.0	92.0	70.0	70.0	72.0	56.0	72.0
InternVL -Chat-v1.2-Plus [7]	86.0	59.7	74.0	60.0	67.0	65.5	65.0	65.3	94.0	74.0	62.0	72.0	52.0	70.8
Gemini-1.5-Pro [49]	66.0	40.0	72.0	48.0	60.0	31.0	21.0	26.0	98.0	82.0	86.0	90.0	72.0	85.6
Claude-3.5-Sonnet [2]	83.5	54.0	73.0	32.0	52.5	55.0	45.0	50.0	92.0	64.0	76.0	74.0	66.0	74.4
GPT-4V [45]	77.5	53.7	75.0	48.0	61.5	50.5	58.5	54.5	96.0	72.0	92.0	82.0	64.0	81.2
GPT-4o [46]	77.0	61.0	83.0	62.0	72.5	53.0	64.0	58.5	96.0	72.0	74.0	76.0	50.0	73.6

Model	Human Understanding									Face	Human	Per.	Rea.	Overall
	Attr.	Action	Spatial Relation			Social Relation			Re-ID					
			RPU	CC	Mean	SRR	IR	Mean						
Random	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	50.0	35.0	30.0	29.2	37.5	32.5
LLaVA -OneVision-0.5B [25]	47.0	78.0	44.0	22.7	33.3	62.0	94.0	78.0	45.0	48.0	56.3	53.3	50.3	52.1
DeepSeek -VL-1.3B-Chat [42]	40.5	66.0	40.0	26.0	33.0	64.0	72.0	68.0	50.0	47.8	51.5	49.3	50.3	49.7
Yi-VL-6B [68]	67.0	73.0	54.0	24.0	39.0	48.0	66.0	57.0	47.0	54.4	56.6	60.7	47.8	55.5
MiniGPT-4-7B [78]	15.5	27.0	18.0	16.7	17.3	24.0	34.0	29.0	44.0	29.3	26.6	24.2	33.6	27.9
InstructBLIP-7B [10]	31.0	46.0	34.0	0.7	17.3	16.0	28.0	22.0	51.0	43.9	33.5	40.7	35.8	38.7
Qwen-VL-Chat [3]	49.5	83.0	54.0	34.0	44.0	64.0	70.0	67.0	50.0	54.4	58.7	57.9	54.5	56.5
DeepSeek -VL-7B-Chat [42]	64.0	78.0	52.0	35.3	43.7	70.0	76.0	73.0	57.0	54.6	63.1	60.7	56.1	58.9
LLaVA-1.5-7B [32]	62.0	71.0	54.0	30.0	42.0	68.0	78.0	73.0	63.0	55.6	62.2	59.8	57.6	58.9
LLaVA-NeXT-7B [33]	62.0	80.0	62.0	24.7	43.3	62.0	86.0	74.0	56.0	59.7	63.1	64.6	56.6	61.4
InternLM -XComposer2-VL-7B [13]	87.5	87.0	58.0	41.3	49.7	64.0	86.0	75.0	59.0	62.9	71.6	73.2	58.4	67.3
LLaVA -OneVision-7B [25]	90.5	92.0	58.0	48.0	53.0	66.0	86.0	76.0	61.0	61.3	74.5	74.5	58.0	67.9
CogVLM2-19B-Chat [23]	70.5	93.0	68.0	33.3	50.7	74.0	92.0	83.0	56.0	59.0	70.6	68.4	59.4	64.8
GLM-4V-9B [23]	85.5	94.0	62.0	32.0	47.0	68.0	88.0	78.0	67.0	63.5	74.3	73.2	62.5	68.9
MiniGPT-4-13B [78]	19.5	46.0	42.0	17.3	29.7	30.0	50.0	40.0	48.0	33.1	36.6	30.7	41.1	34.9
InstructBLIP-13B [10]	33.5	71.0	38.0	28.0	33.0	52.0	86.0	69.0	51.0	43.1	51.5	44.9	51.0	47.3
LLaVA-13B [34]	27.0	66.0	36.0	30.7	33.3	38.0	76.0	57.0	55.0	44.3	47.7	43.9	49.1	46.0
LLaVA-1.5-13B [32]	60.5	72.0	44.0	26.0	35.0	60.0	60.0	60.0	54.0	60.1	56.3	63.7	50.0	58.2
LLaVA-NeXT-13B [33]	69.5	74.0	46.0	28.0	37.0	58.0	70.0	64.0	63.0	58.7	61.5	63.5	54.9	60.1
InternVL-Chat-v1.5 [7]	89.5	89.0	62.0	50.7	56.3	70.0	74.0	72.0	77.0	71.5	76.8	78.6	67.4	74.1
LLaVA-NeXT-34B [33]	91.5	88.0	64.0	59.3	61.7	64.0	86.0	75.0	88.0	71.7	80.8	77.7	74.2	76.3
InternVL -Chat-v1.2-Plus [7]	90.0	92.0	66.0	58.7	62.3	76.0	96.0	86.0	85.0	69.7	83.1	76.7	76.0	76.4
Gemini-1.5-Pro [49]	50.0	75.0	52.0	25.3	38.7	74.0	84.0	79.0	82.0	55.6	64.9	52.8	71.3	60.3
Claude-3.5-Sonnet [2]	71.5	90.0	54.0	42.7	48.3	74.0	80.0	77.0	74.0	62.9	72.2	70.0	68.4	67.5
GPT-4V [45]	73.0	78.0	38.0	71.3	54.7	68.0	84.0	76.0	83.0	65.7	72.9	66.4	73.7	69.3
GPT-4o [46]	63.5	81.0	50.0	58.7	54.3	66.0	94.0	80.0	79.0	68.5	71.6	68.9	71.7	70.0

Section 3.5 confirm our hypothesis. (4) Newer models show significant improvements compared to earlier models. Among MLLMs with 7B parameters within LLM, LLaVA-OneVision-7B performs best. Impressively, LLaVA-OneVision-0.5B, with only 0.5B parameters within LLM, outperforms the earlier InstructBLIP-13B.

L2 and L3 Performance[7] (1) At L2 and L3, the best performance among open-source models is usually achieved by one of InternvL-Chat-v1.2-Plus, LLaVA-Next-34B, and InternVL-Chat-v1.5.

[7]For the visualization of L2 and L3 results, please refer to the Section C.1.

https://doi.org/10.52202/085713-0322

Specifically, GLM-4V-9B achieves the best results in compound expression recognition (L3), facial expression recognition (L2), and action recognition (L2) and CogVLM2-19B-Chat achieves the best result in relative position understanding (L3). (2) At L2 and L3, the best performance among closed-source models is usually achieved by GPT-4o or GPT-4V. Notably, Gemini-1.5-Pro demonstrates outstanding face recognition ability (L2), achieving the best performance among all models with a score of 85.6%.

3.3 Correlation Between Abilities

In this section, we examine whether improving one ability in a model will enhance another by calculating the Pearson Correlation Coefficient between abilities at different levels, using the evaluation scores from Section 3.2.

At L1, the correlation coefficient of face and human understanding is 0.94 and the correlation coefficient of perception and reasoning is 0.79, both indicating significant positive correlations, as shown in Figure 3(a) and (b). We further investigate the correlations between L2 abilities, resulting in the correlation coefficient matrix shown in Figure 3(c). For clarity, we have drawn this as a lower triangular matrix. Our findings can be summarized as follows: (1) For the three face understanding abilities—facial attribute recognition, age estimation, and facial expression recognition—there are high positive correlations between each pair. (2) For the four human understanding abilities—human attribute recognition, action recognition, spatial relation understanding, and social relation understanding—there are high positive correlations between each pair. (3) For the three face understanding abilities and four human understanding abilities mentioned above, there are high positive correlations between each pair. (4) The two identity recognition tasks—face recognition and person re-identification—show a high positive correlation. (5) The correlation between face attack detection and any other ability is low. In Section C.3, we further present the correlations between L3 abilities.

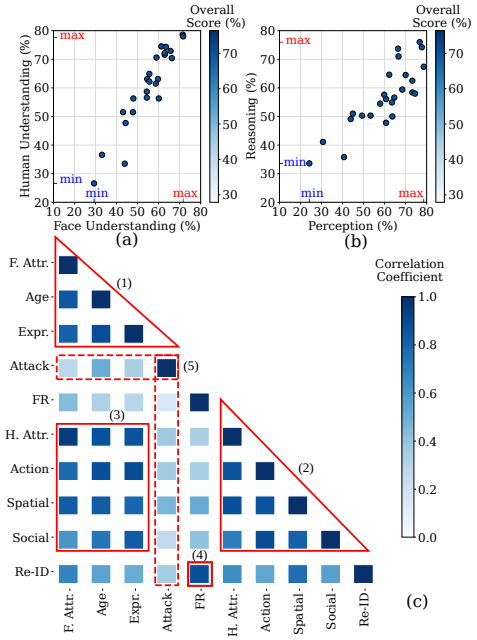


Figure 3: Correlation between abilities.

3.4 Relative Position of Targets

We investigate the impact of the relative position of targets on performance in four L3 abilities: facial attribute recognition, age estimation, basic expression recognition, and human attribute recognition. As shown in Figure 4, for the three face understanding abilities, we provide both the original and cropped versions, where only one person is included but the relative position varies. For human attribute recognition, we offer box-added and cropped versions. In the box-added version, multiple people are included, with the target to be discussed indicated by a red box. Figure 5 illustrates the performance differences between the two versions across various models. Our findings can be summarized as follows.

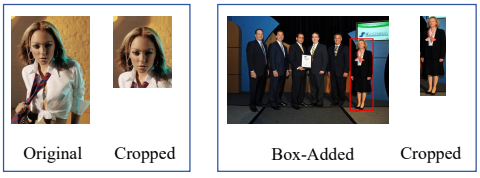


Figure 4: (a) The versions used for the three face understanding abilities. (b) The versions used for human attribute recognition.

Face Understanding Abilities. (1) Preferences for either version depend on the model and the ability, with no overarching trend observed. (2) A model's preference can vary across different face understanding abilities. For example, Yi-VL-6B shows no significant preference for facial attribute recognition, prefers the original images for age estimation, and favors cropped images for basic

https://doi.org/10.52202/085713-0322

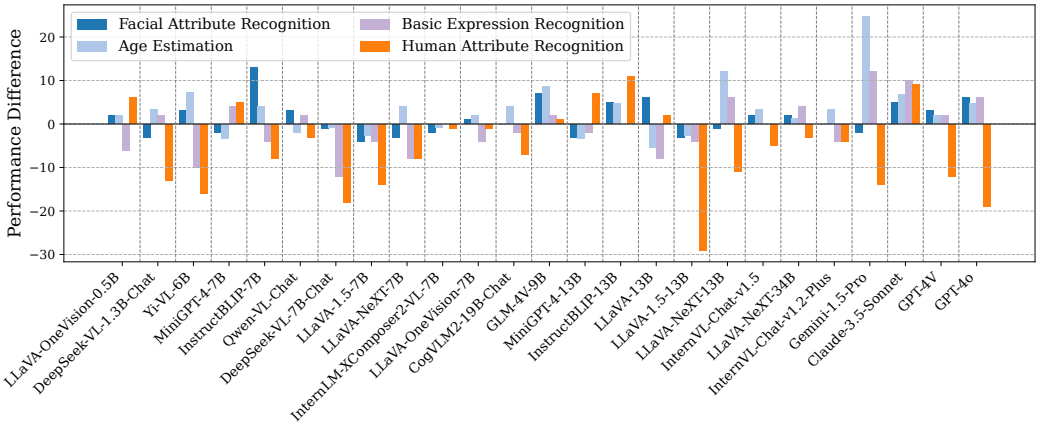


Figure 5: The performance differences between the two versions across various models. For the three face understanding abilities, we show the performance of the original version minus that of the cropped version. For human attribute recognition, we show the performance of the box-added version minus that of the cropped version.

expression recognition. We think that this phenomenon may occur because MLLMs have been trained using images with different target relative positions when aligning visual information for different facial features.

Human Attribute Recognition. The majority of models perform better on the cropped version. This indicates that these models still struggle to accurately understand a specific individual when there are multiple people in the image.

We define the relative position sensitivity score (RPSS) as the sum of the absolute differences in scores between the two versions across the four tasks. This metric can serve as an effective reference for training MLLMs with more robust visual alignment for face and human understanding. We observe that InternLM-XComposer2-VL-7B, LLaVA-OneVision-7B, InternVL-Chat-v1.5, LLaVA-NeXT-34B, and InternVL-Chat-v1.2-Plus not only perform well in the four tasks but also exhibit low sensitivity scores. Among them, InternLM-XComposer2-VL-7B has the lowest sensitivity score of only 3.7%.[8]

3.5 CoT prompting

In this section, we select InternVL-Chat-v1.2-Plus and GPT-4o to explore whether incorporating hints and Chain-of-Thought (CoT) instructions in the prompts can enhance the MLLMs' performance. These two models have achieved the best overall performance in the main experiment among open-source models and closed-source models respectively. A hint involves tips on how to answer the question. For example, the hint for person re-identification is "if two people have significant differences in posture and their faces are relatively blurry, the main basis for determining whether they are the same person is their clothing characteristics." CoT instructions, on the other hand, guide MLLMs in articulating the reasoning process that leads to the answer. The vanilla CoT instruction simply requires the model to "analyze the question and options step by step", whereas task-specific CoT instructions

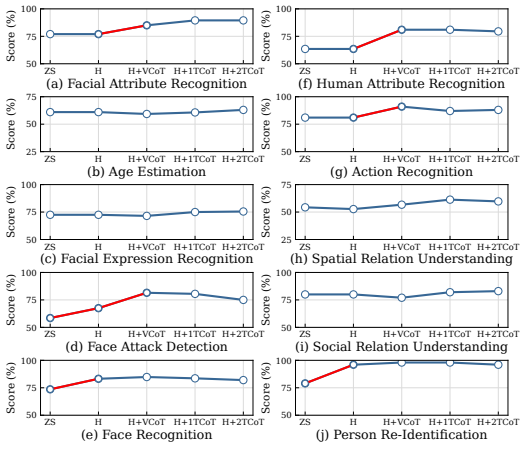


Figure 6: Main reasons of performance improvements for each L2 ability are highlighted in red.

[8]For more models' RPSS, please refer to the Section C.4

https://doi.org/10.52202/085713-0322

Table 2: Scores of InternVL-Chat-v1.2-Plus and GPT-4o under different settings. ZS is short for Zero-Shot, H is short for Hints, VCoT is short for Vanilla CoT, 1TCoT is short for 1-stage Task-specific CoT. 2TCoT is short for 2-stage Task-specific CoT. Q is short for Question. O is short for Options. A is short for Answer. R is short for Relevant Analysis. The highest scores for open-source and closed-source MLLMs are marked in blue and green respectively.

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

provide more tailored guidance based on the task. For example, for the deepfake detection task, the prompt might instruct the model to "analyze whether there are any artifacts in the facial image." Following Multi-modal CoT [74], we also conduct ablation experiments with both 1-stage and 2-stage frameworks. In the 1-stage framework, MLLMs are required to sequentially output the relevant analysis (rationale) and the answer in one round of dialogue. In the 2-stage framework, MLLMs first output the relevant analysis (rationale) in the first round and then provide the answer in the second round. Hints and task-specific CoT instructions for each L3 ability can be found in Section B.2.3.

Table 2 presents the performance of InternVL-Chat-v1.2-Plus and GPT-4o after incorporating hints and three different CoT settings. The results indicate that including hints and CoT instructions does not improve the performance of the open-source model; in fact, it may even cause a slight performance decline. By analyzing the outputs, we find that the open-source model does not provide rationales in its responses after adding CoT instructions to prompts. We believe this could be due to the model's insufficient generalization capabilities, preventing it from understanding the CoT instructions. Specifically, the reasons may include the following aspects: (1) The scarcity of vision-language training samples aligned with the evaluation tasks of face and human understanding. (2) The insufficient exposure to samples of the complex reasoning paradigm during the training process. (3) The inadequate model capacity, which stems from substantially smaller parameter sizes compared to those of closed-source MLLMs. In contrast, the closed-source GPT-4o shows significant performance improvements. Adding hints leads to a 3.4% improvement compared to the zero-shot setting. Building upon this, vanilla CoT, 1-stage task-specific CoT, and 2-stage task-specific CoT further improve performance by 5.2%, 6.5%, and 5.7%, respectively. Ultimately, the combination of hints and 1-stage task-specific CoT instructions emerge as the best setting for overall performance.

In Figure 6, we further explore the main reasons for the performance improvements of GPT-4o in each ability at L2. Hints significantly improve performance in face attack detection, face recognition, and person re-identification, while CoT instructions significantly improve performance in facial attribute recognition, face attack detection, human attribute recognition, and action recognition. For the reasons behind the performance improvements in each ability at L3, please refer to Section C.5.

3.6 Specialist Models Significantly Outperforming MLLMs

In this section, we explore whether specialist models can significantly outperform MLLMs for the evaluation of 13 L3 abilities.[9] We directly test the performance of MLLMs using original datasets from the face and human community to facilitate comparison with specialist models. We design a set of prompt templates to transform the classification problems into multiple-choice problems and the regression problems (age estimation and crowd counting) into fill-in-the-blank problems.[10] Specialist models are generally trained and tested on data from the same distribution. They can achieve high performance even if the test labels contain noise. However, the visual information learned by MLLMs and the original datasets used for testing may exhibit data distribution bias. To enable an effective comparison, we utilize early specialist models (which emerged after the widespread adoption of deep learning) as a reference to judge the performance of MLLMs on these tasks.[11]

[9] We explain the reasons for not conducting experiments on the remaining 5 L3 abilities in Section B.3.1

[10] For prompt templates, please refer to Section B.3.2

[11] For the early specialist models used, refer to Section C.6

Table 3: Comparison between MLLMs and specialist models on 13 L3 abilities. The best-performing MLLMs are highlighted in blue, while abilities where MLLMs perform significantly worse than specialist models are marked in orange.

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.

We further define the relative performance score S to normalize performances across different tasks: $S = (P_m - P_r)/(P_s - P_r)$, where P_m is the performance of the MLLM. Here, we take the highest-performing model among InternVL-Chat-v1.2-Plus, LLaVA-Next-34B , and InternVL-Chat-v1.5 (the top three models in the main experiment). P_r is the performance of random responses, and P_s is the performance of the early specialist model. This metric typically ranges from 0 to 1, where a higher relative score indicates stronger abilities of MLLMs on the corresponding task. A relative score below 0 stands for even worse results than random responses, whereas a score above 1 indicates the performance surpassing the corresponding specialist models for reference.

As shown in Table 3, MLLMs perform well in age estimation, facial expression recognition, face anti-spoofing, action recognition, and person re-identification, eliminating the need to introduce specialist models to enhance response quality. In contrast, for deepfake detection and crowd counting tasks, the MLLM significantly underperforms specialist models. Moreover, for face recognition, MLLMs can approach the specialist model under the basic scenario but indicate poor performance under more challenging scenarios, such as cross-pose, cross-age, similar-looking, and occluded. We recommend incorporating the corresponding specialist models into multi-modal assistants for applications where deepfake detection, crowd counting, and accurate face recognition are required. Section F provides a demonstration of how to enhance multi-modal assistant responses with specialist models.

4 Related Work

Evaluation of MLLMs for Face and Human Understanding. Currently, there is no dedicated benchmark evaluating the face and human understanding abilities of MLLMs. Some efforts aim at comprehensively benchmarking MLLMs, containing some ability dimensions about face and human understanding. LAMM [67] evaluates 9 different 2D vision tasks using 11 existing public datasets. Among these, the facial classification task utilizes the CelebA [41] dataset to evaluate the accuracy of smile detection and hair attribute classification. MME [15] includes the celebrity recognition ability, requiring MLLMs to respond with Yes/No answers. SEED-Bench [26] includes the action recognition ability, where the inputs consist of multiple frames taken from a video, and MLLMs are required to choose the correct answer from four descriptions. MMBench [40] includes the most extensive set of abilities related to faces and humans: celebrity recognition, action recognition, identity reasoning, and

[12]The original metric for Market1501 is mAP. For easier comparison, we create a new testing protocol consisting of 750 positive pairs and 750 negative pairs. The ACC can be calculated in the same way as for LFW. We re-evaluate the early specialist model for Re-ID using the new protocol.

https://doi.org/10.52202/085713-0322

social relation, all of which are tested using multiple-choice problems. Considering the importance of faces and humans in multimedia, these evaluations are insufficient.

Face and Human Understanding. Face and human understanding is among the earliest research topics in artificial intelligence with successful applications. Numerous high-quality datasets were proposed for training and evaluating tasks of face attribute recognition [41], age estimation [52, 14, 73], facial expression recognition [4, 29, 43], deepfake detection [51, 12], face anti-spoofing [37, 38], face recognition [66, 18, 77, 11, 76], human attribute recognition [31, 35], human-object interaction detection [19, 63], crowd counting [71], social relationship recognition [53, 28] and person re-ideitification [30, 75]. Entering the 2020s, a new paradigm emerged, which initially pre-trains a task-agnostic backbone and then based on this, trains a unified face or human model [9, 61, 48] to simultaneously handle multiple face and human understanding tasks within a unified structure. In our evaluation, we observe that in certain tasks, MLLMs do not perform as well as specialist models. Utilizing these unified face or human models as specialist models to help MLLMs can greatly enhance responses.

5 Conclusion

In this work, we propose the hierarchical Face-Human-Bench, the first benchmark specifically designed to evaluate MLLMs' face and human understanding abilities. We comprehensively and scientifically assess the performance of 25 mainstream MLLMs with our benchmark. We reveal the correlations between abilities and explore the impact of the relative position of targets and CoT prompting on the performance of MLLMs. Inspired by multimodal agents, we investigate which abilities of MLLMs need to be supplemented by specialist models. Our work will provide the face and human community valuable insights on how to more effectively leverage multi-modal assistants in applications related to "faces and humans."

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grants 62306043, 62076031, and 62076036.

References

[1] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2023.

[2] Anthropic. Claude 3.5 sonnet, 2024a.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, pages 279–283. ACM, 2016.

[5] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognit. Lett.*, 140:325–331, 2020.

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238, 2023.

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1800–1807. IEEE Computer Society, 2017.

[9] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *CVPR*, pages 17840–17852. IEEE, 2023.

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

[11] Weihong Deng, Jiani Hu, Nanhai Zhang, Binghui Chen, and Jun Guo. Fine-grained face verification: FGLFW database, baselines, and human-dcmn partnership. *Pattern Recognit.*, 66:63–73, 2017.

[12] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *CoRR*, abs/1910.08854, 2019.

[13] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[14] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo Jair Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *ICCV Workshops*, pages 243–251. IEEE Computer Society, 2015.

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.

[16] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. VITA: towards open-source interactive omni multimodal LLM. *CoRR*, abs/2408.05211, 2024.

[17] Xiao Guo, Yaojie Liu, Anil K. Jain, and Xiaoming Liu. Multi-domain learning for updating face anti-spoofing models. In *ECCV (13)*, volume 13673 of *Lecture Notes in Computer Science*, pages 230–249. Springer, 2022.

[18] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV (3)*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2016.

[19] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *CoRR*, abs/1505.04474, 2015.

[20] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad: Movie description in context. In *CVPR*, pages 18930–18940. IEEE, 2023.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[22] Fabian Herzog, Xunbo Ji, Torben Teepe, Stefan Hörmann, Johannes Gilg, and Gerhard Rigoll. Lightweight multi-branch network for person re-identification. In *2021 IEEE international conference on image processing (ICIP)*, pages 1129–1133. IEEE, 2021.

[23] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[24] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[25] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

 https://doi.org/10.52202/085713-0322

[26] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125, 2023.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[28] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Dual-glance model for deciphering social relationships. In *ICCV*, pages 2669–2678. IEEE Computer Society, 2017.

[29] Shan Li, Weihong Deng, and Junping Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2584–2593. IEEE Computer Society, 2017.

[30] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159. IEEE Computer Society, 2014.

[31] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV (6)*, volume 9910 of *Lecture Notes in Computer Science*, pages 684–700. Springer, 2016.

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023.

[33] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[35] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, pages 350–359. IEEE Computer Society, 2017.

[36] Xuannan Liu, Pei Pei Li, Huaibo Huang, Zekun Li, Xing Cui, Weihong Deng, Zhaofeng He, et al. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *ACM Multimedia 2024*, 2024.

[37] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398. Computer Vision Foundation / IEEE Computer Society, 2018.

[38] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *CVPR*, pages 4680–4689. Computer Vision Foundation / IEEE, 2019.

[39] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACM Multimedia*, pages 4235–4243. ACM, 2020.

[40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *CoRR*, abs/2307.06281, 2023.

[41] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE Computer Society, 2015.

[42] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, abs/2403.05525, 2024.

[43] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, 2019.

[44] OpenAI. Chatgpt. https://openai.com/blog/chatgpt/, 2023.

[45] OpenAI. Gpt-4v(ision) system card, 2023.

[46] OpenAI. Hello gpt-4o, 2024.

[47] Lixiong Qin, Ning Jiang, Yang Zhang, Yuhan Qiu, Dingheng Zeng, Jiani Hu, and Weihong Deng. Towards interactive deepfake analysis. *CoRR*, abs/2501.01164, 2025.

[48] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. *CoRR*, abs/2403.09500, 2024.

[49] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[50] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 0–1, 2020.

[51] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11. IEEE, 2019.

[52] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: deep expectation of apparent age from a single image. In *ICCV Workshops*, pages 252–257. IEEE Computer Society, 2015.

[53] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *CVPR*, pages 435–444. IEEE Computer Society, 2017.

[54] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[55] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

[56] Chengrui Wang, Han Fang, Yaoyao Zhong, and Weihong Deng. MLFW: A database for face recognition on masked faces. In *CCBR*, volume 13628 of *Lecture Notes in Computer Science*, pages 180–188. Springer, 2022.

[57] Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. Contextual AD narration with interleaved multimodal sequence. *CoRR*, abs/2403.12922, 2024.

[58] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274. Computer Vision Foundation / IEEE Computer Society, 2018.

[59] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 692–702, 2019.

[60] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *CoRR*, abs/2311.03079, 2023.

[61] Yizhou Wang, Yixuan Wu, Shixiang Tang, Weizhen He, Xun Guo, Feng Zhu, Lei Bai, Rui Zhao, Jian Wu, Tong He, and Wanli Ouyang. Hulk: A universal knowledge translator for human-centric tasks. *CoRR*, abs/2312.01697, 2023.

[62] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023.

 https://doi.org/10.52202/085713-0322

[63] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, pages 2019–2028. Computer Vision Foundation / IEEE, 2019.

[64] Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *ICCV*, pages 2051–2060. IEEE, 2019.

[65] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: prompting chatgpt for multimodal reasoning and action. *CoRR*, abs/2303.11381, 2023.

[66] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[67] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, Jing Shao, and Wanli Ouyang. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In *NeurIPS*, 2023.

[68] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652, 2024.

[69] Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. Movie101: A new movie understanding benchmark. In *ACL (1)*, pages 4669–4684. Association for Computational Linguistics, 2023.

[70] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *CoRR*, abs/2309.15112, 2023.

[71] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597. IEEE Computer Society, 2016.

[72] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *ECCV (88)*, volume 15146 of *Lecture Notes in Computer Science*, pages 399–415. Springer, 2024.

[73] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, pages 4352–4360. IEEE Computer Society, 2017.

[74] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024, 2024.

[75] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124. IEEE Computer Society, 2015.

[76] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5:7, 2018.

[77] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017.

[78] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*. OpenReview.net, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction can accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the proposed benchmark in Appendix G of the supplementary materials.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/085713-0322

Answer: [NA]

Justification: The main content of our paper is benchmarking and evaluation, and it does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In the main text, we describe the evaluation protocol and provide further details in Appendices A and B. Additionally, we have made the dataset and evaluation code available, along with a user guide.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have made the dataset and evaluation code available, along with a user guide, by providing the URLs of the dataset and code repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experimental details in Appendices A and B. Additionally, we have made the dataset and evaluation code available, along with a user guide.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please refer to Appendix C.7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/085713-0322

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the relevant information in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research complies with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impacts in Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our data pipeline uses raw images sourced from 16 public datasets from the face and human community, thereby eliminating such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For more details, please refer to Appendix H.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

https://doi.org/10.52202/085713-0322

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have made the dataset and evaluation code available, along with a user guide, by providing the URLs of the dataset and code repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Our study does not include human subjects; however, we have used crowdsourcing methods for data checking in our semi-automated data pipeline. For more details, please refer to Appendices A.3.3 and H.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our study does not include human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-0322

Appendix

Contents

https://doi.org/10.52202/085713-0322

 https://doi.org/10.52202/085713-0322

A More Details on Face-Human-Bench

A.1 Definition about Each Leaf Ability

We will sequentially describe the definitions of L2 abilities and the L3 abilities they encompass. We provide examples of problems in Face-Human-Bench in Tables 4 to 11.

Facial Attribute Recognition: Recognize various characteristics and traits from facial images.

Age Estimation: Estimate the age of the person in the image based on facial information.

Facial Expression Recognition: Recognize the emotions of the person in the image, categorized into basic and compound types. Basic expressions include surprised, fearful, disgusted, happy, sad, angry, and neutral. Compound expressions provide more nuanced emotional descriptions, including: happily surprised, happily disgusted, sadly fearful, sadly angry, sadly surprised, sadly disgusted, fearfully angry, fearfully surprised, angrily surprised, angrily disgusted, and disgustedly surprised.

Face Attack Detection: Determine whether the face in the image involves digital manipulation or physical spoofing. The corresponding sub-abilities are referred to as Deepfake Detection and Face Anti-Spoofing, respectively.

Face Recognition Identify and verify individuals' identities in images according to facial information. In our tests, this ability is mainly to determine whether two photos showcase the same individual. Five scenarios are involved: basic, cross-pose, cross-age, similar-looking, and occluded.

Human Attribute Recognition Recognize various characteristics and traits from human images.

Action Recognition Recognize human actions, including interactions with objects.

Spatial Relation Understanding Understand the spatial positions of people in the image, including relative position understanding (comprehending the relative positions of one person to others and objects) and crowd counting (counting the number of people in the image).

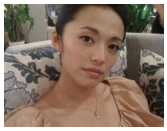
Social Relation Understanding Including social relationship recognition (inferring social relationships between people through their interactions) and identity reasoning (deducing social identity based on a person's attributes, actions, interactions with others, and environmental information).

Person Re-Identification Identify and verify individuals' identities in images based on full-body attributes (usually excluding the face, as facial features are often blurry).

Table 4: Examples of problems in Face-Human-Bench.

Ability	**Example**
Facial Attribute Recognition	**Image:**  **Question:** Please select the description that best applies to the person in the picture. A. not wearing necktie, not wearing lipstick, not wearing earrings. B. without eyeglasses, bald, with mouth slightly open. C. male, with black hair, wearing earrings. D. with eyeglasses, not wearing hat, with bangs. **Answer:** A.

Table 5: Examples of problems in Face-Human-Bench.

Ability	Example
Age Estimation (5-Year Interval)	**Image:**  **Question:** Which age do you believe is most likely for the person in the photo? A. 10. B. 15. C. 20. D. 25. **Answer:** D.
Age Estimation (10-Year Interval)	**Image:**  **Question:** Which of the following ages is the most likely for the person in the picture? A. 20. B. 30. C. 40. D. 50. **Answer:** A.
Age Estimation (15-Year Interval)	**Image:**  **Question:** Which of the following ages is the most likely for the person in the picture? A. 47. B. 62. C. 77. D. 92. **Answer:** B.
Facial Expression Recognition (Basic Expression Recognition)	**Image:**  **Question:** What is the expression of the person in this photo? A. Neutral. B. Sadness. C. Disgust. D. Fear. **Answer:** A.
Facial Expression Recognition (Compound Expression Recognition)	**Image:**  **Question:** Based on this picture, what is the person's expression? A. Happily Disgusted. B. Fearfully Surprised. C. Sadly Disgusted. D. Sadly Fearful. **Answer:** A.

https://doi.org/10.52202/085713-0322

Table 6: Examples of problems in Face-Human-Bench.

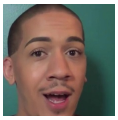
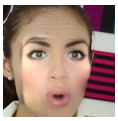
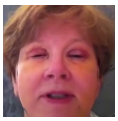
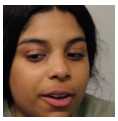
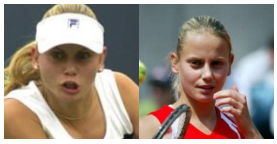
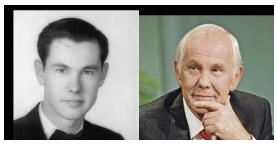
Ability	Example
Face Attack Detection (Deepfake Detection) (Bona Fide)	**Image:**  **Question:** Has the facial image undergone digital alteration? A. yes. B. no. **Answer:** B.
Face Attack Detection (Deepfake Detection) (Face-Swapping)	**Image:**  **Question:** Was the facial image digitally modified in any way? A. yes. B. no. **Answer:** A.
Face Attack Detection (Deepfake Detection) (Face-Reenactment)	**Image:**  **Question:** Was the facial appearance digitally changed? A. yes. B. no. **Answer:** A.
Face Attack Detection (Face Anti-Spoofing) (Bona Fide)	**Image:**  **Question:** Has the facial image been compromised by a presentation attack? A. yes. B. no. **Answer:** B.
Face Attack Detection (Face Anti-Spoofing) (Print)	**Image:**  **Question:** Is there a spoofing attempt visible in the facial image? A. yes. B. no. **Answer:** A.
Face Attack Detection (Face Anti-Spoofing) (Replay)	**Image:**  **Question:** Is the facial recognition being deceived by a presentation attack? A. yes. B. no. **Answer:** A.

Table 7: Examples of problems in Face-Human-Bench.

Ability	Example
Face Recognition (Basic Face Recognition)	**Image:**  **Question:** Are the people portrayed in the two pictures identical? A. yes. B. no. **Answer:** A.
Face Recognition (Basic Face Recognition)	**Image:**  **Question:** Are the individuals in both images one and the same? A. yes. B. no. **Answer:** B.
Face Recognition (Cross-Pose Face Recognition)	**Image:**  **Question:** Do the individuals appearing in the two images happen to be identical? A. yes. B. no. **Answer:** A.
Face Recognition (Cross-Pose Face Recognition)	**Image:**  **Question:** Do the people shown in both pictures happen to be one and the same person? A. yes. B. no. **Answer:** B.
Face Recognition (Cross-Age Face Recognition)	**Image:**  **Question:** Are the people portrayed in the two pictures identical? A. yes. B. no. **Answer:** A.

https://doi.org/10.52202/085713-0322

Table 8: Examples of problems in Face-Human-Bench.

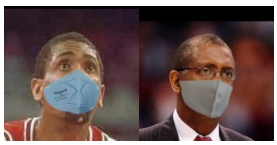
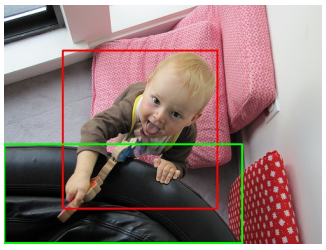
Ability	Example
Face Recognition (Cross-Age Face Recognition)	**Image:**  **Question:** Do the individuals in both images happen to be the same person? A. yes. B. no. **Answer:** B.
Face Recognition (Similar-Looking Face Recognition)	**Image:**  **Question:** Are the persons depicted in the photos on the left and right sides identical? A. yes. B. no. **Answer:** A.
Face Recognition (Similar-Looking Face Recognition)	**Image:**  **Question:** Are the persons depicted in the photos on the left and right sides identical? A. yes. B. no. **Answer:** B.
Face Recognition (Occluded Face Recognition)	**Image:**  **Question:** Is the individual captured in both the left and right photographs one and the same person? A. yes. B. no. **Answer:** A.
Face Recognition (Occluded Face Recognition)	**Image:**  **Question:** Do the individuals appearing in the two images happen to be identical? A. yes. B. no. **Answer:** B.

Table 9: Examples of problems in Face-Human-Bench.

Ability	Example
Human Attribute Recognition	**Image:**  **Question:** Which statement best describes the individual highlighted in the red box in the picture? A. She is wearing a long-sleeve shirt and is not wearing a hat or a skirt. B. She is wearing a T-shirt and a hat, but her clothes do not have any logos. C. She is dressed informally in a skirt and wearing sunglasses. D. She has long hair and is wearing a short-sleeved shirt along with a face mask. **Answer:** A.
Action Recognition	**Image:**  **Question:** Which of these options best describes what the person in the red box is doing in the picture? A. Washing the motorcycle. B. Waxing the motorcycle. C. Polishing the motorcycle. D. Repairing the motorcycle. **Answer:** A.
Spatial Relation Understanding (Relative Position Understanding)	**Image:**  **Question:** Among the following options, what is the most fitting way to characterize the subject (marked with a red box)'s location in relation to the object (marked with a green box)? A. The child is behind the sofa. B. The child is to the right of the sofa. C. The child is to the left of the sofa. D. The child is under the sofa. **Answer:** A.

https://doi.org/10.52202/085713-0322

Table 10: Examples of problems in Face-Human-Bench.

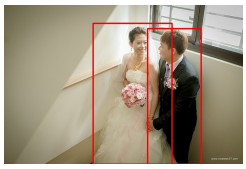
Ability	Example
Spatial Relation Understanding (Crowd Counting) (Less than 10)	**Image:**  **Question:** What's the number of individuals in this picture? A. 2. B. 3. C. 4. D. 5. **Image:** D.
Spatial Relation Understanding (Crowd Counting) (10-100)	**Image:**  **Question:** Among the options, which numeral is closest to the total count of humans in the picture? A. 10. B. 30. C. 90. D. 140. **Image:** B.
Spatial Relation Understanding (Crowd Counting) (More than 100)	**Image:**  **Question:** What is the closest numerical value among the options to the number of individuals in the image? A. 400. B. 1100. C. 3200. D. 5300. **Answer:** B.
Social Relation Understanding (Social Relationship Recognition)	**Image:**  **Question:** Which relationship do the two people in the red box in the photo most likely have? A. Couple. B. No Relation. C. Family. D. Friends. **Answer:** A.
Social Relation Understanding (Identity Reasoning)	**Image:**  What is the most likely occupation of the person highlighted in red in the picture? A. basketball player. B. basketball team manager. C. basketball coach. D. sports commentator. **Answer:** A.

Table 11: Examples of problems in Face-Human-Bench.

Ability	Example
Person Re-Identification	**Image:**  **Question:** Is the person in the first picture the same as the person in the second picture? A. yes. B. no. **Answer:** A.
Person Re-Identification	**Image:**  Is the individual captured in both the left and right photographs one and the same person? A. yes. B. no. **Image:** B.

A.2 Data Sources and Statistics

Table 12 provides information on the data sources for Face-Human-Bench, as well as the image processing pipeline, the number of problems in the development and test sets, and the weights, for each subset.

We set the weights of all 10 L2 abilities to be equal. For L2 abilities that encompass multiple L3 abilities, each L3 ability equally shares the weight of the corresponding L2 ability. For L3 abilities that encompass multiple image versions, each image version subset equally shares the weight of the corresponding L3 ability. Finally, we obtain the detailed weights of each subset, as shown in Table 12.

We sequentially provide overviews of the public datasets we used for original samples.

CelebA [41] is a large-scale facial attributes dataset released by the Multimedia Laboratory of Chinese University of Hong Kong. It contains over 200,000 celebrity images, each annotated with 40 attributes. The dataset includes a wide range of body pose variations and complex, diverse background information. It comprises 10,177 identities, 202,599 face images, and 5 landmark positions, with 40 binary attribute annotations for each image.

UTKFace [73] dataset is a large-scale facial dataset with a wide age range, spanning from 0 to 116 years. It contains over 20,000 face images, annotated with age, gender, and ethnicity labels.

RAF-DB [29] is a large-scale facial expression database consisting of 29,672 real-world images, each accompanied by a 7-dimensional expression distribution vector. It includes two different subsets: a single-label subset with 7 basic expressions (RAF-DB Basic) and a two-tab subset with 12 compound expressions (RAF-DB Compound). Additionally, the dataset provides 5 precise landmark locations, 37 automatic landmark positions, bounding boxes, and annotations for ethnicity, age range, and gender attributes for each image.

FF++ [51] consists of 1,000 original video sequences processed using four different automated facial manipulation methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The data in FaceForensics++ comes from 977 YouTube videos, all featuring trackable frontal faces without occlusions, allowing the automated manipulation methods to generate realistic forgeries.

 https://doi.org/10.52202/085713-0322

Table 12: Data sources and statistics of the Face-Human-Bench.

Level-1	Level-2	Level-3	Data Source	p_{image}	Dev. Num.	Test Num.	Weight
Face	Facial Attribute Recognition	Facial Attribute Recognition	CelebA	Identity	100	100	5.0%
				Crop	100	100	5.0%
	Age Estimation	Age Estimation	UTKFace	Identity	150	150	5.0%
				Crop	150	150	5.0%
	Facial Expression Recognition	Basic Expression Recognition	RAF-DB (Basic)	Identity	50	50	2.5%
				Crop	50	50	2.5%
		Compound Expression Recognition	RAF-DB (Compound)	Identity	50	50	5.0%
	Face Attack Detection	Deepfake Detection	FF++	Identity	100	100	5.0%
		Face Anti-Spoofing	SiW-Mv2	Identity	100	100	5.0%
	Face Recognition	Basic Face Recognition	LFW	Cat	50	50	2.0%
		Cross-Pose Face Recognition	CPLFW	Cat	50	50	2.0%
		Cross-Age Face Recognition	CALFW	Cat	50	50	2.0%
		Similar-Looking Face Recognition	SLLFW	Cat	50	50	2.0%
		Occluded Face Recognition	MLFW	Cat	50	50	2.0%
Human	Human Attribute Recognition	Human Attribute Recognition	WIDER Attribute	AddBox	100	100	5.0%
				Crop	100	100	5.0%
	Action Recognition	Action Recognition	HICO-DET	AddBox	100	100	10.0%
	Spatial Relation Understanding	Relative Position Understanding	SpatialSense	Identity	50	50	5.0%
		Crowd Counting	PISC ShTech	Identity	150	150	5.0%
	Social Relation Understanding	Social Relationship Recognition	PISC	AddBox	50	50	5.0%
		Identity Reasoning	PISC	AddBox	50	50	5.0%
	Person Re-Identification	Person Re-Identification	Market-1501	Cat	100	100	10.0%

SiW-Mv2 [17] collects 785 videos from 493 subjects, and 915 spoof videos from 600 subjects. The dataset includes 14 types of spoofing, ranging from typical print and replay attack, to various masks, impersonation makeup and physical material coverings. SiW-Mv2 exhibits a good variance in spoofing modes, with each mode specified and validated by the IARPA project.

LFW [24] is a commonly used test set for face recognition, comprising 13,233 face images sourced from natural scenes in everyday life. Each image is associated with a name, representing 5,749 individuals, with most people having only one image. The database randomly selected 6,000 pairs of faces to create face recognition image pairs to test the accuracy of face recognition systems, with 3,000 pairs containing two images of the same person and 3,000 pairs featuring one image of different individuals.

CPLFW [76] builds upon LFW by considering the impact of pose variations. It specifically searches for and selects 3,000 pairs of positive faces with differing poses, adding pose variation to the intra-class variance. Additionally, it includes negative pairs with the same gender and ethnicity to minimize the influence of attribute differences between positive and negative pairs.

CALFW [77] builds upon LFW by considering the impact of age variations. It specifically searches for and selects 3,000 pairs of positive faces with age differences to increase the intra-class variance associated with the aging process. Negative pairs are chosen to have the same gender and ethnicity to reduce the influence of attribute differences.

SLLFW [11] intentionally selects 3,000 pairs of visually similar faces through human crowdsourcing from the original image folder, replacing the random negative pairs in LFW.

MLFW [56] dataset is created based on CALFW and focuses on masked faces. The masks generated for the faces in the dataset maintain good visual consistency with the original faces. It includes a variety of mask templates that cover most common styles encountered in everyday life, achieving diversity of the samples.

WIDER Attribute [31] is a large-scale human attributes dataset containing 13,789 images across 30 scene categories, with 57,524 human bounding boxes. Each bounding box is annotated with 14 binary attributes, including male, long hair, sunglasses, hat, long shirt, long sleeves, formal, shorts, jeans, long pants, skirt, mask, logo, and checkered or striped patterns.

HICO-DET [63] is a commonly used dataset in the Human Object Interaction (HOI) domain, consisting of 47,776 images, with 38,118 in the training set and 9,658 in the testing set. The dataset includes 117 action (verb) categories, 80 object categories, and 600 verb-object combinations.

SpatialSense [64] is a dataset for spatial relation recognition, where the task is to determine whether a specific spatial relation holds between two given objects. The dataset contains 17,498 relations on 11,569 images, involving 3,679 unique object classes, with 2,139 of these classes appearing only once, presenting a challenging long-tail distribution.

PISC [28] is focused on the task of social relationship recognition in still images. It is used to benchmark models that analyze the relationships between people based on contextual and individual features. It contains 22,670 images with 76,568 annotated samples representing 9 types of social relationships.

ShTech [71] is focused on the task of crowd counting, where the goal is to accurately estimate the number of people in an image with varying crowd density and perspective. It contains 1,198 images with approximately 330,000 annotated heads. The dataset aims to address challenges in crowd counting that were not covered by previous datasets.

Market-1501 [75] is designed for the task of person re-identification. This dataset addresses the limitations of scale and realistic conditions found in previous datasets. The large-scale data supports training and testing models effectively for person re-identification. It includes over 32,000 annotated bounding boxes and a distractor set of more than 500,000 images.

A.3 More Details on the Semi-Automatic Data Pipeline

A.3.1 Details on Image Processing Pipeline

Figure 7 illustrates four operations of the image processing pipeline: cropping, concatenating, adding boxes, or leaving the original images unchanged. For simplicity, these four operations are denoted as Crop, Cat, AddBox, and Identity, respectively. The image processing pipeline used for each L3 ability is shown in Table 12.

Figure 7: Four operations of the image processing pipeline.

A.3.2 Details on Text Processing Pipeline

We introduce the text processing pipeline for each L3 ability as follows. **Facial Attribute Recognition** Each option involves three attributes. At least two of the three attribute descriptions are incorrect in the incorrect options.

Age Estimation Add incorrect options at intervals of 5 years, 10 years, and 15 years, with each interval accounting for one-third of the total.

Basic Expression Recognition Incorrect options are randomly selected from the remaining 6 categories of expressions after removing the correct option.

https://doi.org/10.52202/085713-0322

Compound Expression Recognition Incorrect options are randomly selected from the remaining 10 categories of expressions after removing the correct option.

Deepfake Detection Set the options to “Yes" and “No". “Yes" indicates the presence of digital manipulations, while “No" indicates their absence.

Face Anti-Spoofing Set the options to “Yes" and “No". “Yes" indicates the presence of physical spoofs, while “No" indicates their absence.

Basic/Cross-Pose/Cross-Age/Similar-Looking/Occluded Face Recognition Set the options to “Yes" and “No". “Yes" indicates that the two photos are of the same person, while “No" indicates that the two photos are not of the same person.

Human Attribute Recognition Each option involves three attributes combined into a complete sentence using ChatGPT. At least two of the three attribute descriptions are incorrect in the incorrect options.

Action Recognition The incorrect options are actions generated by ChatGPT related to but not the same as the correct option.

Relative Position Understanding Each option is a sentence formed by connecting the subject and the object with a preposition. Incorrect options are generated by randomly selecting prepositions from the remaining 8 categories of relative positions after removing the correct preposition.

Crowd Counting The set includes three equally sized subsets, with the number of people in each subset being within the ranges of less than 10, 10-100, and more than 100, respectively. In the first subset, the incorrect options are also numbers within 10. In the latter two subsets, the incorrect options are numbers that are half, three times, and five times the correct option, respectively, with all options rounded to the nearest 10 and 100.

Social Relationship Recognition Incorrect options are randomly selected from the remaining 5 categories of social relations after removing the correct option.

Identity Reasoning The incorrect options are occupations generated by GPT related to but not the same as the correct option.

Person Re-Identification Set the options to “Yes" and “No". “Yes" indicates that the two photos are of the same person, while “No" indicates that the two photos are not of the same person.

A.3.3 Details on Data Checking

At the end of our data pipeline, the produced problems are checked by data reviewers. Each problem is read by three reviewers who are provided with an instruction as shown in Table 13. A problem is retained only if all three reviewers deem it acceptable.

Table 13: The instruction provided to data reviewers.

Please review the image and read the question with options for the problem. The problem is considered acceptable if the following conditions are met: 1. The wording of the question and options is unambiguous. 2. There is one and only one correct option. Is the question acceptable? Please choose: [Yes. It is acceptable.] [No. It is unacceptable.] [I’m not sure.]

B More Details on Experiment Setup

All experiments for the open-source models were conducted on four NVIDIA A800 80G GPUs.

B.1 Overviews of Involved MLLMs

GPT-4V and GPT-4o: GPT-4V [45], released by OpenAI in September 2023, is a vision-enabled variant of the GPT-4 model, utilizing the same training process as GPT-4 for its visual capabilities. It is first trained on a large dataset of text and images, followed by fine-tuning through Reinforcement Learning with Human Feedback (RLHF). GPT-4V demonstrates the exceptional performance of a language-only system augmented with new modalities. The API we applied in our experiments is "gpt-4-turbo-2024-04-09". GPT-4o [46] is released by OpenAI in May 2024. It accepts any combination of text, image, audio and video as input and generates any combination of text, image, and audio output. GPT-4o attains GPT-4 Turbo-level performance in text, inference, and code, while also demonstrating strong capabilities in multilingual, audio, and visual tasks. The API we applied in our experiments is "gpt-4o-2024-05-13".

Gemini [54]: Gemini is a multimodal large model developed by Google, available in three scales: Ultra, Pro, and Nano. From its inception, Gemini was designed with a multimodal focus, excelling in tasks across image, audio, video, and text domains. In February 2024, Google released Gemini 1.5 [49], which includes Gemini 1.5 Pro and the more lightweight Gemini 1.5 Flash. In our work, we employ Gemini 1.5 Pro to conduct experiments.

Claude [1]: The Claude model is developed by Anthropic and is intended to be a useful, honest and harmless assistant. The version we applied in this paper, Claude 3.5 Sonnet [2], was released on June 2024. It is the most powerful visual model in the Claude series to date.

LLaVA [34]: LLaVA is an open-source large multimodal model that leverages multimodal language-image instruction-following data for instruction tuning. It was released in April 2023. LLaVA-1.5 [32], released in October 2023, introduced the following key improvements: the use of MLP as a vision-language connector, the use of prompt data with explicitly specified output formats, and the addition of task-specific datasets for training. Following that, LLaVA-1.6 (LLaVA-NeXT) [33] was released in January 2024, featuring improved input image resolution and enhanced visual reasoning and OCR capabilities. The model also supports better visual conversation on different scenarios and applications. SGLang was utilized for efficient deployment and inference. We apply LLaVA-13B, LLaVA-1.5-7B, LLaVA-1.5-13B, LLaVA-NeXT-7B, LLaVA-NeXT-13B, and LLaVA-NeXT-34B in our experiments.

MiniGPT-4 [78]: MiniGPT-4, released in April 2023, uses a projection layer to align a frozen vision encoder with the frozen LLM Vicuna. The authors trained MiniGPT-4 in two stages: the first stage involved using a low-level dataset, and in the second stage, they curated a detailed image description dataset to fine-tune the model. In our experiments, we use MiniGPT-4-7B and MiniGPT-4-13B.

InstructBLIP [10]: InstructBLIP, released in May 2023, applies its instruction-tuning paradigm to the BLIP-2 [27] model. To be specific, InstructBLIP performs instruction fine-tuning on visual tasks to enhance model performance. In our experiments, InstructBLIP-7B and InstructBLIP-13B are used.

Qwen-VL [3]: Qwen-VL, released in August 2023, accepts images, text, and bounding boxes as inputs, and outputs text and bounding boxes. It supports multilingual and multi-image interleaved dialogue, as well as open-domain localization in Chinese. Qwen-VL is also capable of relatively fine-grained recognition and understanding. We adapt Qwen-VL-Chat in our experiments.

InternLM-XComposer2-VL [70]: InternLM-XComposer-VL, released in September 2023, is a multimodal large language model built with InternLM [55] as the language model. Later, in January 2024, InternLM-XComposer2-VL [70] was released, supporting free-form text and image composition. The authors proposed the Partial LoRA (PLoRA) method, which balances precise visual understanding with literary-inspired text generation. InternLM-XComposer2-VL-7B is used in our experiments.

Yi-VL [68]: Yi-VL, released in May 2024, excels in image-text understanding and chat generation, supporting multi-turn image-text conversations, bilingual text, and fine-grained image comprehension. Yi-VL adopts the LLaVA architecture and employs a three-stage training process to align visual information with the semantic space of Yi LLM [68].

 https://doi.org/10.52202/085713-0322

InternVL [7]: InternVL, released in December 2023, extends its visual model to 6 billion parameters. It progressively aligns with the LLM using web-scale image-text data. InternVL-Chat-V1.2 was released in February 2024, expanding the LLM to 34 billion parameters. Shortly after, InternVL-Chat-v1.2-Plus was introduced, utilizing more supervised fine-tuning (SFT) data to further enhance its performance. Subsequently, InternVL-Chat-v1.5 [6] was released in April 2024, with improvements primarily focused on a stronger visual encoder, dynamic high-resolution capability, and a high-quality bilingual dataset. The model we use in the experiments includes InternVL-Chat-v1.2-Plus and InternVL-Chat-v1.5.

DeepSeek-VL [42]: DeepSeek-VL, released in March 2024, is designed for general multimodal understanding. It is built for real-world applications in visual and language comprehension, capable of handling tasks such as logical diagrams, web pages, formula recognition, scientific literature, natural images, etc. In the experiments, we apply DeepSeek-VL-1.3B and DeepSeek-VL-7B.

CogVLM2 and GLM-4V [60, 23]: CogVLM, released in October 2023, enables deep fusion of visual and language features without sacrificing performance on NLP tasks. In May 2024, the next generation, CogVLM2, was introduced. It inherited the visual expert architecture and improved training recipes in the pre-training and post-training stages, supporting high input resolutions. Shortly after, in June 2024, GLM-4V was released. It used the same data and training recipes as CogVLM2 but employed GLM-4-9B as the language models and removed the visual expert to reduce the model size. In our experiments, we utilize CogVLM2-19B-Chat and GLM-4V-9B.

LLaVA-OneVision [25]: LLaVA-OneVision, released in August 2024, supports three major computer vision scenarios: single image, multi-image, and video scenes. It also exhibits strong transfer learning capabilities across different modalities and scenarios. We use LLaVA-OneVision-0.5B and LLaVA-OneVision-7B in our experiments.

Table 14 summarizes the LLMs and vision encoders used in involved MLLMs.

Table 14: The LLMs and vision encoders used in involved MLLMs.

Model	LLM	Params.	Vision Encoder	Params.
LLaVA-OneVision-0.5B	Qwen2-0.5B	0.5B	SigLIP ViT-L/16	400M
DeepSeek-VL-1.3B-Chat	DeepSeek-LLM-1.3B-Base	1.3B	SigLIP ViT-L/16	400M
Yi-VL-6B	Yi-6B	6B	CLIP ViT-H/14	632M
MiniGPT-4-7B	Vicuna-7B	7B	EVA-CLIP-g/14	1.0B
InstructBLIP-7B	Vicunad-7B	7B	EVA-CLIP-g/14	1.0B
Qwen-VL-Chat	Qwen-7B	7B	Open CLIP-G/14	1.8B
DeepSeek-VL-7B-Chat	DeepSeek-LLM-7B-Base	7B	SigLIP ViT-L/16 + SAM ViT-B	400M + 86M
LLaVA-1.5-7B	Vicuna-v1.5-7B	7B	CLIP-L/14	304M
LLaVA-NeXT-7B	Vicuna-v1.5-7B	7B	CLIP-L/14	304M
InternLM-XComposer2-VL-7B	InternLM-7B	7B	EVA-CLIP-g/14	1.0B
LLaVA-OneVision-7B	Qwen2-7B	7B	SigLIP ViT-L/16	400M
CogVLM2-19B-Chat	Llama-3-8B-Instruct	8B	EVA-02-CLIP-E/14	4.4B
GLM-4V-9B	GLM-4-9B	9B	EVA-02-CLIP-E/14	4.4B
MiniGPT-4-13B	Vicuna-13B	13B	EVA-CLIP-g/14	1.0B
InstructBLIP-13B	Vicuna-13B	13B	EVA-CLIP-g/14	1.0B
LLaVA-13B	LLaMA-2-13B-Chat	13B	CLIP-L/14	304M
LLaVA-1.5-13B	Vicuna-v1.5-13B	13B	CLIP-L/14	304M
LLaVA-NeXT-13B	Vicuna-v1.5-13B	13B	CLIP-L/14	304M
InternVL-Chat-v1.5	InternLM2-20B	20B	InternViT-6B	6B
LLaVA-NeXT-34B	Yi-34B	34B	CLIP-L/14	304M
InternVL-Chat-v1.2-Plus	Nous-Hermes-2-Yi-34B	34B	InternViT-6B	6B

B.2 More Details on the Experiments for Q1

B.2.1 Prompt Templates for Different Settings

Zero-Shot (ZS) The prompt template used for the zero-shot setting is shown in Table 15.

Hints (H) The prompt template for experiments with hints is shown in Table 16.

Hints and Vanilla CoT Instructions (H+VCoT) The prompt template for experiments with hints and vanilla CoT instructions is shown in Table 17.

Hints and Task-Specific Instructions With One-Stage Framework (H+1TCoT) The prompt template for one-stage experiments with hints and task-specific CoT instructions is shown in Table 18.

Table 15: The prompt template used for the zero-shot setting.

Question: [Question]
[Options]
Please provide the answer to the multiple-choice question, using only the option's letter to indicate your choice. Note: Only one option is correct. For questions you are unsure about, please choose the answer you think is most likely.

Table 16: The prompt template used for experiments with hints.

Question: [Question]
[Options]
Hint: [Hint]
Please provide the answer to the multiple-choice question **based on the hint**, using only the option's letter to indicate your choice. Note: Only one option is correct. For questions you are unsure about, please choose the answer you think is most likely.

Table 17: The prompt template used for experiments with hints and vanilla CoT instructions.

Question: [Question]
[Options]
Hint: [Hint]
First, please analyze the question and options **step by step** in conjunction with the input image.
Then, please provide the answer to the multiple-choice question **based on the hint and relevant analysis**. Note: Only one option is correct. For questions you are unsure about, please choose the answer you think is most likely.

Table 18: The prompt template used for one-stage experiments with hints and task-specific CoT instructions.

Question: [Question]
[Options]
Hint: [Hint]
First, [Task-specific CoT instruction]
Then, please provide the answer to the multiple-choice question **based on the hint and relevant analysis**. Note: Only one option is correct. For questions you are unsure about, please choose the answer you think is most likely.

Table 19: The prompt template used for two-stage experiments with hints and task-specific CoT instructions.

Stage 1
Question: [Question]
[Options]
Hint: [Hint]
[Task-specific CoT instruction]
Stage 2
Question: [Question]
[Options]
Hint: [Hint]
Relevant Analysis: [Output from stage 1]
Please provide the answer to the multiple-choice question **based on the hint and relevant analysis**. Note: Only one option is correct. For questions you are unsure about, please choose the answer you think is most likely.

https://doi.org/10.52202/085713-0322

Hints and Task-Specific Instructions With Two-Stage Framework (H+2TCoT) The prompt template for two-stage experiments with hints and task-specific CoT instructions is shown in Table 19.

B.2.2 Prompt Used for Choice Extraction

The prompt used for choice extraction is shown in Table 20.

Table 20: The prompt template used for choice extraction.

You are an AI assistant to help me match an answer with several options of a multiple-choice problem. You are provided with a question, several options, and an answer, and you need to find which option is most similar to the answer. If the meaning of all options is significantly different from the answer, output X. You should output a single uppercase character in A, B, C, D (if they are valid options), and X.

Question: Please select the description that best matches the individual depicted.
Options:
A. He is wearing a face mask but is not wearing a hat or a skirt.
B. He is wearing a face mask, a hat, and shorts.
C. He has short hair and is not wearing a face mask or a T-shirt.
D. He is not wearing clothes with a logo or stripes, and he isn't wearing sunglasses.
Answer: He is wearing a face mask, a hat, and shorts.
Your Output: B

Question: Which description best represents the person in the image?
Options:
A. She is wearing a T-shirt and sunglasses, and her clothes do not have a logo.
B. She is wearing a face mask and sunglasses but is not wearing long pants.
C. She is without sunglasses, not wearing a hat, and not wearing a T-shirt.
D. She is dressed informally in a short-sleeved top and is not wearing a T-shirt.
Answer: None of the provided descriptions accurately represent the person in the image.
Your Output: X

Question: [Question]
Options: [Options]
Answer: [Answer]
Your Output:

B.2.3 Hints and Task-specific CoT Instructions

Hints and task-specific CoT instructions for each L3 ability are shown in Table 21.

Table 21: Hints and task-specific CoT instructions.

L3 Ability	Hint	Task-specific CoT instruction
F. Attr.	/	Please analyze whether the characteristics described in the multiple-choice options match the attributes of the face in the image, one by one.
Age	/	Please (1) analyze the facial age characteristics of the person in the image and (2) provide a possible age number that you think is appropriate. Note: Please do not respond with "I can't determine the exact age"; just provide the number you think is closest.
Basic Expr.	/	Please describe the facial emotional features of the person in the image.
Comp. Expr.		

L3 Ability	Hint	Task-specific CoT instruction
Deepfake	A forged face may be generated by face-swapping, which is a technique that replaces one person's facial features with those of another person.	Please analyze whether there are any artifacts indicating face-swapping in the facial image.
	A forged face may be generated by face-reenactment, which is a technique that transfers the facial expressions and movements of one person onto another person's face in real-time or in a recorded video.	Please analyze whether there are any artifacts indicating face-reenactment in the facial image.
Spoofing	A spoof face image may be printed on paper and then re-photographed.	Please analyze whether there are any clues in the facial image that indicate it was printed on paper and then re-photographed.
	A spoof face image may be re-photographed after being played on a video playback device.	Please analyze whether there are any clues in the facial image that indicate it was re-photographed from a video playback device.
Basic FR	/	Please analyze whether the two people in the images are the same person by explaining the similarities and differences in their facial features.
C.P. FR	Even if the two images are of the same person, there may be differences in posture.	
C.A. FR	Even if the two images are of the same person, there may be differences in age, meaning the two photos were taken at different ages of this person.	
S.L. FR	Even if the two photos are not of the same person, they may still have similar facial features.	
Occ. FR	To determine whether the two partially obscured photos are of the same person, it is necessary to analyze other unobscured facial areas.	
H. Attr.	/	Please analyze whether the characteristics described in each option of the multiple-choice question match the person in the red box, one by one.
Action	/	Please analyze the actions of the person in the red box.
Position	/	Please analyze the relative positional relationship between the subject (marked with a red box) and the object (marked with a green box).
Counting	There are fewer than 10 people in the image.	Please estimate the number of people appearing in the image, including those who are occluded or incomplete. Note: Please do not say 'I cannot determine the exact number of people'; just provide the number you think is approximate.
	There are fewer than 100 people in the image.	

https://doi.org/10.52202/085713-0322

L3 Ability	Hint	Task-specific CoT instruction
	There are more than 100 people in the image, but fewer than 4,000.	
Social Rel.	/	Please analyze the possible social relationship between the two people in the red boxes from the perspectives of relative position, posture, and facial expressions.
Identity	/	Please analyze the occupation of the person in the red box from the perspectives of clothing, actions, background, etc.
Re-ID	If two people have significant differences in posture and their faces are relatively blurry, the main basis for determining whether they are the same person is their clothing characteristics.	Please analyze whether the two people in the images are the same person by explaining the similarities and differences in their full-body features.

B.3 More Details on the Experiments for Q2

B.3.1 Unexplored L3 Abilities

We explain the reasons for not conducting experiments on the remaining 5 L3 abilities as follows.

Face/Human Attribute Recognition These two tasks include a large number of binary classification labels (40 labels in CelebA for face and 14 labels in WIDER Attribute for human). Using evaluation protocols designed for specialist models to fully assess the performance of MLLMs would result in huge computational costs. Additionally, many attribute labels have ambiguous semantics that are difficult to define accurately, such as "attractive," "big lips," and "big nose."

Relative Position Understanding In the face and human understanding community, there are no specialist models specifically constructed to perceive the spatial relationships between one person to others and objects.

Identity Reasoning There is a lack of publicly available specialist models that perform occupation classification.

Social Relationship Recognition Existing evaluation datasets for specialist models, such as PIPA and PISC, are not suitable for directly evaluating the social relationship recognition abilities of MLLMs. Because many annotations in these datasets are ambiguous (it is often impossible to assert that people in an image belong to one relationship category rather than another) and semantically overlapping (for example, “couple" is a separate category, but a married couple is semantically also part of “family"). Fine-tuned specialist models can still learn to classify under unclear standards, but for zero-shot learning MLLMs, the lack of clear definitions greatly reduces performance. It is worth noting that in Face-Human-Bench, when we use PISC to construct problems for evaluating the social relationship recognition ability of MLLMs, we manually review and remove problems with ambiguous options or potentially non-unique answers, allowing us to assess accurately.

B.3.2 Explored L3 Abilities

We provide the prompt templates for directly evaluating L3 abilities on public datasets from the face and human community in Tables 22 to 30. During testing, the options will be randomly shuffled.

Table 22: Prompt for **Age Estimation** on UTKFace.

What is the age of the person in the image? Please answer with a number between 0 and 100.
Answer the question with a single number, and don't provide other additional explanations.

Table 23: Prompt for **Basic Expression Recognition** on RAF-DB (Basic).

What expression is on the face in the image?
A. Surprise B. Fear C. Disgust D. Happiness E. Sadness F. Anger G. Neutral
Answer with the option's letter from the given choices directly, and don't provide other additional explanations.

Table 24: Prompt for **Compound Expression Recognition** on RAF-DB (Compound).

What expression is on the face in the image?
A. Happily Surprised B. Happily Disgusted C. Sadly Fearful D. Sadly Angry
E. Sadly Surprised F. Sadly Disgusted G. Fearfully Angry H. Fearfully Surprised
I. Angrily Surprised J. Angrily Disgusted K. Disgustedly Surprised
Answer with the option's letter from the given choices directly, and don't provide other additional explanations.

Table 25: Prompt for **Deepfake Detection** on FF++.

Is there any evidence of face forgery artifacts in the picture?
A. Yes B. No
Answer with the option's letter from the given choices directly, and don't provide other additional explanations.

Table 26: Prompt for **Face Anti-Spoofing** on SiW-Mv2.

Is the face in the picture a spoof face?
A. Yes B. No
Answer with the option's letter from the given choices directly, and don't provide other additional explanations.

https://doi.org/10.52202/085713-0322

Table 27: Prompt for **Basic/Cross-Pose/Cross-Age/Similar-Looking/Occluded Face Recognition** on LFW/CPLFW/CALFW/SLLFW/MLFW.

> Are the people in the two photos the same person?
> A. Yes B. No
> Answer with the option's letter from the given choices directly, and don't provide other additional explanations.

Table 28: Prompt for **Action Recognition** on HICO-DET.

> Which of the following words best describes the interaction between the person in the red box and the object in the green box?
> [Opions, include all actions involving the same object extracted from HICO-DET.]
> Answer with the option's letter from the given choices directly, and don't provide other additional explanations.

Table 29: Prompt for **Crowd Counting** on ShTech-A.

> How many people are there in the picture approximately? Please answer with a number between 0 and 4000.
> Answer the question with a single number, and don't provide other additional explanations.

Table 30: Prompt for **Person Re-Identification** on Market-1501.

> Are the people in the two photos the same person?
> A. Yes B. No
> Answer with the option's letter from the given choices directly, and don't provide other additional explanations.

C Additional Results

C.1 Face-Human-Bench (English)

We provide the visualization of the L2 and L3 results in Figures 8 to 10.

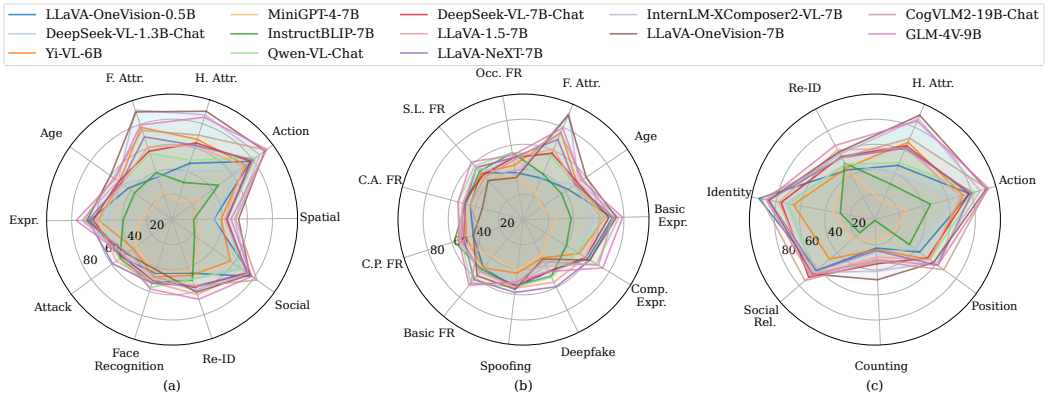


Figure 8: The performance of open-source MLLMs with LLM parameter scales below 10B on L2 and L3 abilities.

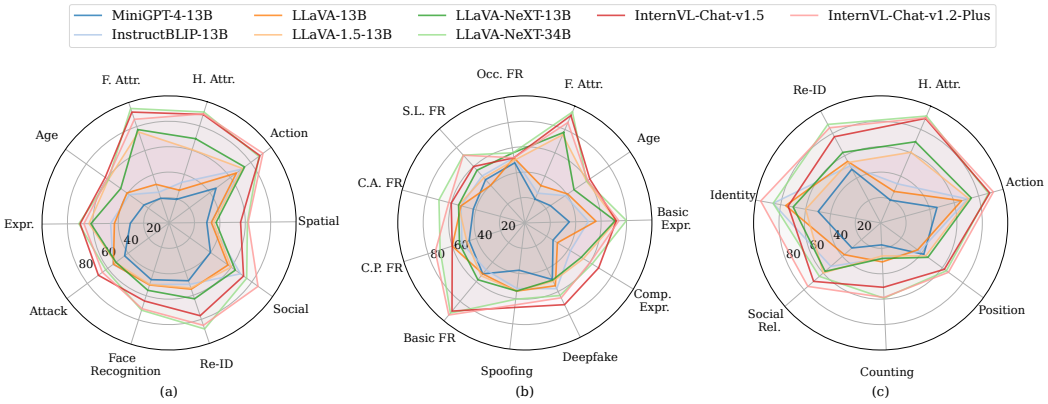


Figure 9: The performance of open-source MLLMs with LLM parameter scales above 10B on L2 and L3 abilities.

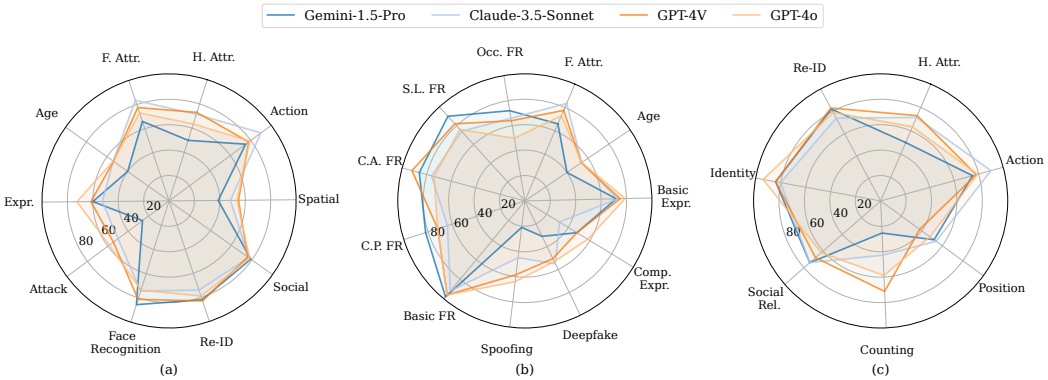


Figure 10: The performance of closed-source MLLMs on L2 and L3 abilities.

https://doi.org/10.52202/085713-0322

C.2 Face-Human-Bench (Chinese)

Table 31 shows the performance of all evaluated MLLMs at different levels of abilities on the Human-Face-Bench (Chinese). We further compare the performance of different MLLMs on English and Chinese versions of the Face-Human-Bench, as shown in Figure 11. Models are sorted with the ascending order of average performance.

Table 31: Zero-shot scores of MLLMs on the hierarchical Face-Human-Bench (CN). The highest scores for open-source and closed-source MLLMs are marked in blue and green respectively. The scores in the "random" row are theoretical values.

Model	Face Understanding													
	Attr.	Age	Expression			Attack Detection			Face Recognition					
			Basic	Comp.	Mean	DFD	FAS	mean	Basic	C.P.	C.A.	S.L.	Occ.	Mean
Random	25.0	25.0	25.0	25.0	25.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
LLaVA -OneVision-0.5B	29.0	34.3	67.0	58.0	62.5	38.0	56.0	47.0	50.0	44.0	50.0	52.0	52.0	49.6
DeepSeek -VL-1.3B-Chat	37.0	48.7	61.0	62.0	61.5	47.0	50.0	48.5	50.0	50.0	48.0	44.0	50.0	48.4
Yi-VL-6B	60.0	49.3	67.0	46.0	56.5	25.0	28.0	26.5	36.0	34.0	34.0	24.0	38.0	33.2
MiniGPT-4-7B	21.0	21.7	28.8	25.0	24.0	50.9	45.5	39.3	60.4	57.8	46.7	35.4	45.7	45.6
InstructBLIP-7B	24.0	28.3	39.0	34.0	36.5	49.0	47.0	48.0	48.0	50.0	50.0	48.0	48.0	48.8
Qwen-VL-Chat	54.5	49.0	68.0	40.0	54.0	55.0	53.3	53.8	66.0	52.0	68.0	54.0	50.0	58.0
DeepSeek -VL-7B-Chat	67.5	54.7	65.0	52.0	58.5	49.0	51.0	50.0	58.0	52.0	40.0	42.0	42.0	46.8
LLaVA-1.5-7B	48.0	49.7	51.0	56.0	53.5	54.5	51.0	52.8	64.0	46.0	46.0	62.0	46.0	52.8
LLaVA-NeXT-7B	39.5	40.0	66.0	68.0	67.0	55.5	50.0	52.0	56.0	52.0	52.0	52.0	46.0	51.6
InternLM -XComposer2-VL-7B	87.0	53.0	74.0	68.0	71.0	45.0	51.0	48.0	58.0	46.0	48.0	66.0	34.0	50.4
LLaVA -OneVision-7B	91.0	61.0	75.0	60.0	67.5	35.0	52.0	43.5	60.0	38.0	20.0	36.0	28.0	36.4
CogVLM2-19B-Chat	77.5	55.7	76.0	68.0	72.0	40.0	45.0	42.5	60.0	40.0	56.0	68.0	48.0	54.4
GLM-4V-9B	84.5	58.3	80.0	78.0	79.0	37.0	52.0	44.5	72.0	60.0	68.0	70.0	64.0	66.8
MiniGPT-4-13B	18.5	26.0	35.4	35.4	33.5	50.8	43.9	29.0	52.1	50.0	60.0	39.5	51.0	46.8
InstructBLIP-13B	7.0	29.0	37.2	31.3	21.0	59.5	47.4	27.2	7.1	9.5	12.2	12.8	25.0	10.8
LLaVA-13B	24.5	37.7	56.6	29.4	34.0	50.8	54.5	44.0	52.1	54.0	52.0	56.0	46.0	51.6
LLaVA-1.5-13B	62.0	53.0	72.0	60.0	66.0	51.5	53.5	52.5	62.0	54.0	50.0	50.0	50.0	53.2
LLaVA-NeXT-13B	54.5	44.0	69.1	37.5	51.5	53.1	56.0	54.0	58.0	50.0	60.0	50.0	50.0	53.6
InternVL-Chat-v1.5	89.0	61.3	82.0	70.0	76.0	61.0	62.0	61.5	94.0	68.0	62.0	66.0	48.0	67.6
LLaVA-NeXT-34B	93.5	55.3	83.0	58.0	70.5	63.0	63.0	63.0	92.0	68.0	78.0	70.0	58.0	73.2
InternVL -Chat-v1.2-Plus	87.0	57.3	73.0	52.0	62.5	61.5	60.5	61.0	96.0	78.0	68.0	72.0	48.0	72.4
Gemini-1.5-Pro	58.5	29.0	70.0	36.0	53.0	11.0	16.0	13.5	98.0	74.0	84.0	88.0	68.0	82.4
Claude-3.5-Sonnet	79.5	54.0	74.0	38.0	56.0	55.0	57.0	56.0	90.0	74.0	82.0	72.0	60.0	75.6
GPT-4V	68.5	55.0	75.0	54.0	64.5	50.0	54.5	52.3	90.0	58.0	84.0	84.0	68.0	76.8
GPT-4o	77.5	57.0	82.0	70.0	76.0	52.0	56.0	54.0	78.0	60.0	68.0	80.0	54.0	68.0

Model	Human Understanding									Face	Human	Per.	Rea.	Overall
	Attr.	Action	Spatial Relation			Social Relation			Re-ID					
			RPU	CC	Mean	SRR	IR	Mean						
Random	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	50.0	35.0	30.0	29.2	37.5	32.5
LLaVA -OneVision-0.5B	37.5	62.0	42.0	20.0	31.0	64.0	82.0	73.0	51.0	44.5	50.9	45.4	51.2	47.7
DeepSeek -VL-1.3B-Chat	35.0	60.0	44.0	24.7	34.3	64.0	82.0	73.0	50.0	48.8	50.5	48.4	51.4	49.6
Yi-VL-6B	56.5	68.0	46.0	24.0	35.0	50.0	74.0	62.0	44.0	45.1	53.1	52.8	43.6	49.1
MiniGPT-4-7B	25.0	29.0	37.2	28.2	25.0	38.6	38.1	33.0	36.0	30.3	29.6	26.7	34.9	30.0
InstructBLIP-7B	30.0	24.0	28.0	10.0	17.0	32.7	45.8	38.0	51.0	37.1	32.0	31.8	38.7	34.6
Qwen-VL-Chat	44.0	72.0	46.0	26.8	35.7	46.8	81.6	62.0	64.0	53.9	55.5	54.5	54.9	54.7
DeepSeek -VL-7B-Chat	55.5	81.0	54.0	40.7	47.3	66.0	82.0	74.0	50.0	55.5	61.6	61.2	54.5	58.5
LLaVA-1.5-7B	35.0	65.0	30.0	32.9	31.3	66.0	88.0	77.0	64.0	51.3	54.5	50.7	56.3	52.9
LLaVA-NeXT-7B	33.0	70.0	28.0	25.2	26.3	54.0	92.0	73.0	55.0	50.0	51.5	50.3	51.5	50.7
InternLM -XComposer2-VL-7B	75.0	78.0	60.0	45.3	52.7	62.0	84.0	73.0	70.0	61.9	69.7	68.7	61.5	65.8
LLaVA -OneVision-7B	84.5	89.0	48.0	46.7	47.3	74.0	92.0	83.0	61.0	59.9	73.0	72.8	56.9	66.4
CogVLM2-19B-Chat	66.5	86.0	56.0	29.3	42.7	64.0	98.0	81.0	60.0	60.4	67.2	66.7	59.5	63.8
GLM-4V-9B	77.0	91.0	62.0	32.0	47.0	66.0	90.0	78.0	62.0	66.6	71.0	72.4	63.5	68.8
MiniGPT-4-13B	28.5	32.0	24.5	26.6	23.3	18.4	40.4	28.0	44.0	30.8	31.2	27.9	35.5	31.0
InstructBLIP-13B	5.0	41.0	17.0	7.0	10.0	42.9	65.2	48.0	8.0	19.0	22.4	21.7	19.2	20.7
LLaVA-13B	22.5	59.0	26.5	31.1	26.7	38.0	73.5	55.0	55.0	38.4	43.6	36.9	47.1	41.0
LLaVA-1.5-13B	38.0	70.0	24.0	18.0	21.0	62.0	88.0	75.0	61.0	57.3	53.0	56.9	52.6	55.2
LLaVA-NeXT-13B	47.5	74.0	40.0	33.0	35.7	51.0	84.0	67.0	58.0	51.5	56.4	54.3	53.6	54.0
InternVL-Chat-v1.5	80.5	87.0	50.0	50.0	50.0	70.0	82.0	76.0	87.0	71.1	76.1	75.9	70.2	73.6
LLaVA-NeXT-34B	87.5	83.0	64.0	44.7	54.3	62.0	88.0	75.0	94.0	71.1	78.8	75.5	74.1	74.9
InternVL -Chat-v1.2-Plus	80.0	88.0	52.0	50.0	51.0	72.0	98.0	85.0	88.0	68.0	78.4	72.6	74.1	73.2
Gemini-1.5-Pro	46.0	79.0	52.0	24.7	38.3	78.0	78.0	78.0	49.0	47.3	58.1	46.5	61.9	52.7
Claude-3.5-Sonnet	55.0	83.0	50.0	36.7	43.3	64.0	78.0	71.0	78.0	64.2	66.1	63.9	67.0	65.1
GPT-4V	51.0	59.0	48.0	65.3	56.7	60.0	78.0	69.0	74.0	63.4	61.9	58.4	69.1	62.7
GPT-4o	51.0	74.0	54.0	51.3	52.7	70.0	92.0	81.0	69.0	66.5	65.5	64.9	67.7	66.0

https://doi.org/10.52202/085713-0322

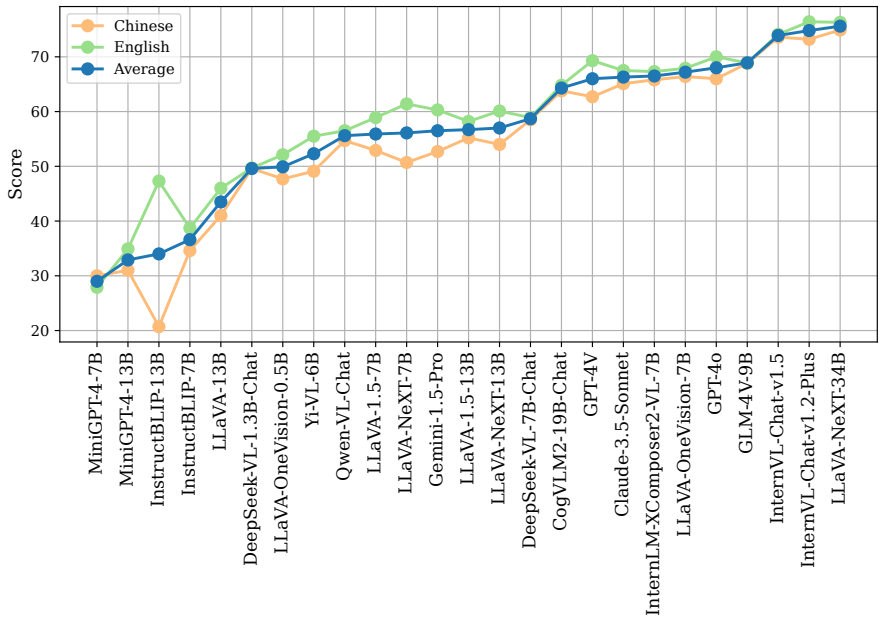


Figure 11: Comparation for the performance of different MLLMs on English and Chinese versions of the Face-Human-Bench.

C.3 Correlation Between Abilities

The correlation coefficient matrix for L3 is shown in Figure 12. Pay particular attention to the ability correlations highlighted in the red boxes.

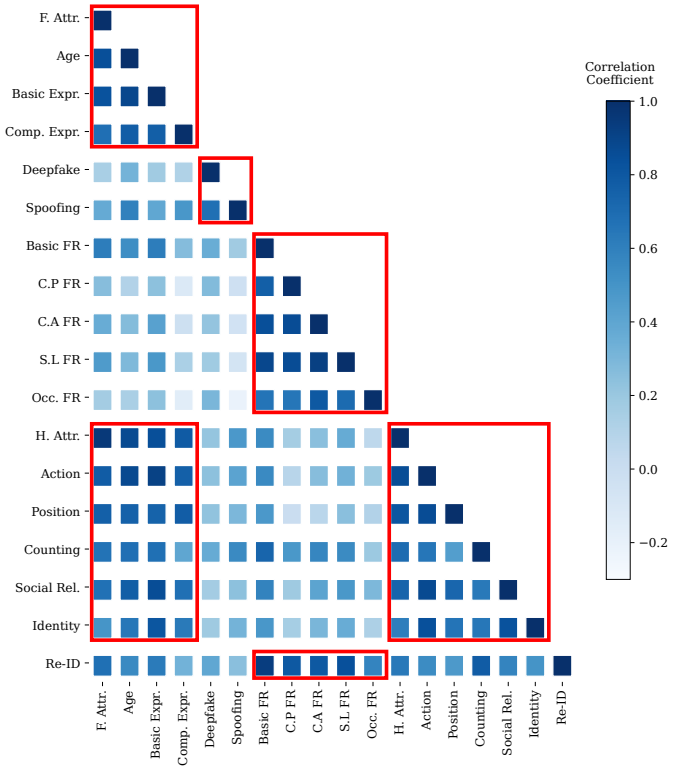


Figure 12: Correlation coefficient matrix for L3.

https://doi.org/10.52202/085713-0322

C.4 Relative Position of Targets

Table 32 presents the performance differences of MLLMs across different relative positions of targets, under the three face understanding abilities and human attribute recognition.

Table 32: The impact of the relative position of targets on performance in four L3 abilities. Models with absolute performance differences greater than 5 between the two versions are highlighted in orange. Models with the smallest RPSS are marked in green.

Model	Facial Attribute			Age			Basic Expression			Human Attribute			RPSS
	Ori.	Crop.	Dif.	Ori.	Crop.	Dif.	Ori.	Crop.	Dif.	Boxed	Crop.	Diff.	
LLaVA-OneVision-0.5B	37.0	35.0	2.0	44.0	42.0	2.0	68.0	74.0	-6.0	50.0	44.0	6.0	16.0
DeepSeek-VL-1.3B-Chat	35.0	38.0	-3.0	50.7	47.3	3.3	58.0	56.0	2.0	34.0	47.0	-13.0	21.3
Yi-VL-6B	77.0	74.0	3.0	55.3	48.0	7.3	60.0	70.0	-10.0	59.0	75.0	-16.0	36.3
MiniGPT-4-7B	23.0	25.0	-2.0	16.0	19.3	-3.3	28.0	24.0	4.0	18.0	13.0	5.0	14.3
InstructBLIP-7B	46.0	33.0	13.0	38.7	34.7	4.0	36.0	40.0	-4.0	27.0	35.0	-8.0	29.0
Qwen-VL-Chat	57.0	54.0	3.0	48.7	50.7	-2.0	66.0	64.0	2.0	48.0	51.0	-3.0	10.0
DeepSeek-VL-7B-Chat	57.0	58.0	-1.0	52.0	52.7	-0.7	62.0	74.0	-12.0	55.0	73.0	-18.0	31.7
LLaVA-1.5-7B	59.0	63.0	-4.0	48.0	50.7	-2.7	60.0	64.0	-4.0	55.0	69.0	-14.0	24.7
LLaVA-NeXT-7B	68.0	71.0	-3.0	52.0	48.0	4.0	68.0	76.0	-8.0	58.0	66.0	-8.0	23.0
InternLM-XComposer2-VL-7B	91.0	93.0	-2.0	52.7	53.3	-0.7	76.0	76.0	0.0	87.0	88.0	-1.0	3.7
LLaVA-OneVision-7B	91.0	90.0	1.0	61.3	59.3	2.0	72.0	76.0	-4.0	90.0	91.0	-1.0	8.0
CogVLM2-19B-Chat	75.0	75.0	0.0	59.3	55.3	4.0	70.0	72.0	-2.0	67.0	74.0	-7.0	13.0
GLM-4V-9B	83.0	76.0	7.0	60.0	51.3	8.7	80.0	78.0	2.0	86.0	85.0	1.0	18.7
MiniGPT-4-13B	19.0	22.0	-3.0	22.7	26.0	-3.3	34.0	36.0	-2.0	23.0	16.0	7.0	15.3
InstructBLIP-13B	28.0	23.0	5.0	40.7	36.0	4.7	50.0	50.0	0.0	39.0	28.0	11.0	20.7
LLaVA-13B	35.0	29.0	6.0	38.0	43.3	-5.3	52.0	60.0	-8.0	28.0	26.0	2.0	21.3
LLaVA-1.5-13B	74.0	77.0	-3.0	57.3	60.0	-2.7	70.0	74.0	-4.0	46.0	75.0	-29.0	38.7
LLaVA-NeXT-13B	77.0	78.0	-1.0	52.7	40.7	12.0	74.0	68.0	6.0	64.0	75.0	-11.0	30.0
InternVL-Chat-v1.5	93.0	91.0	2.0	63.3	60.0	3.3	72.0	72.0	0.0	87.0	92.0	-5.0	10.3
LLaVA-NeXT-34B	96.0	94.0	2.0	59.3	58.0	1.3	82.0	78.0	4.0	90.0	93.0	-3.0	10.3
InternVL-Chat-v1.2-Plus	86.0	86.0	0.0	61.3	58.0	3.3	72.0	76.0	-4.0	88.0	92.0	-4.0	11.3
Gemini-1.5-Pro	65.0	67.0	-2.0	52.7	28.0	24.7	78.0	66.0	12.0	43.0	57.0	-14.0	52.7
Claude-3.5-Sonnet	86.0	81.0	5.0	57.3	50.7	6.7	78.0	68.0	10.0	76.0	67.0	9.0	30.7
GPT-4V	79.0	76.0	3.0	54.7	52.7	2.0	76.0	74.0	2.0	67.0	79.0	-12.0	19.0
GPT-4o	80.0	74.0	6.0	63.3	58.7	4.7	86.0	80.0	6.0	54.0	73.0	-19.0	35.7

C.5 CoT prompting

Based on Table 33, we explore the main reasons for the performance improvements of GPT-4o in each ability at L3, as shown in Figure 13.

Table 33: Scores of the best open-source model, InternVL-Chat-v1.2-Plus, and the best closed-source model, GPT-4o, under different settings on the hierarchical Face-Human-Bench. The highest scores for open-source and closed-source MLLMs are marked in blue and green respectively.

Model	Setting	Face Understanding												
		Attr.	Age	Expression			Attack Detection			Face Recognition				
				Basic	Comp.	Mean	DFD	FAS	Mean	Basic	C.P.	C.A.	S.L.	Occ.	Mean
InternVL -Chat-v1.2-Plus	ZS	86.0	59.7	74.0	60.0	67.0	65.5	65.0	65.3	94.0	74.0	62.0	72.0	52.0	70.8
	H	87.0	60.0	71.0	52.0	61.5	66.0	64.0	65.0	92.0	66.0	56.0	74.0	52.0	68.0
	H+VCoT	86.0	58.3	70.0	64.0	67.0	65.5	61.0	63.3	92.0	68.0	58.0	80.0	56.0	70.8
	H+1TCoT	89.0	61.0	71.0	50.0	60.5	58.0	66.0	62.0	90.0	68.0	64.0	76.0	54.0	70.4
	H+2TCoT	88.0	62.3	72.0	54.0	63.0	58.0	66.5	62.3	94.0	66.0	56.0	78.0	56.0	70.0
GPT-4o	ZS	77.0	61.0	83.0	62.0	72.5	53.0	64.0	58.5	96.0	72.0	74.0	76.0	50.0	73.6
	H	77.0	61.0	83.0	62.0	72.5	52.0	83.0	67.5	96.0	80.0	86.0	90.0	64.0	83.2
	H+VCoT	85.0	59.3	85.0	58.0	71.5	70.0	93.0	81.5	94.0	76.0	86.0	90.0	78.0	84.8
	H+1TCoT	89.5	60.7	84.0	66.0	75.0	67.0	94.0	80.5	98.0	76.0	84.0	88.0	72.0	83.6
	H+2TCoT	89.5	63.0	79.0	72.0	75.5	61.0	89.0	75.0	78.0	90.0	78.0	88.0	76.0	82.0

Model	Setting	Human Understanding								Face	Human	Per.	Rea.	Overall
		Attr.	Action	Spatial Relation			Social Relation			Re-ID				
				RPU	CC	Mean	SRR	IR	Mean					
InternVL -Chat-v1.2-Plus	ZS	90.0	92.0	66.0	58.7	62.3	76.0	96.0	86.0	85.0	69.7	83.1	76.7	76.0	76.4
	H	90.0	95.0	60.0	60.6	60.3	76.0	94.0	85.0	86.0	68.4	83.2	76.4	75.9	75.9
	H+VCoT	87.0	94.0	48.0	65.6	56.3	78.0	86.0	87.0	88.0	69.1	82.5	75.9	74.8	75.7
	H+1TCoT	89.0	92.0	58.0	51.0	54.3	74.0	94.0	84.0	88.0	68.6	81.4	75.6	74.3	75.0
	H+2TCoT	87.0	92.0	58.0	51.3	54.6	72.0	92.0	82.0	80.0	69.1	79.1	75.8	71.8	74.1
GPT-4o	ZS	63.5	81.0	50.0	58.7	54.3	66.0	94.0	80.0	79.0	68.5	71.6	68.9	71.7	70.0
	H	63.5	81.0	50.0	55.3	52.7	66.0	94.0	80.0	96.0	72.2	74.6	70.4	78.0	73.4
	H+VCoT	81.0	91.0	58.0	55.3	56.7	72.0	82.0	77.0	98.0	76.4	80.7	78.2	77.2	78.6
	H+1TCoT	81.0	87.0	60.0	62.7	61.3	74.0	90.0	82.0	98.0	77.9	81.9	79.0	81.2	79.9
	H+2TCoT	79.5	88.0	58.0	61.3	59.7	78.0	88.0	83.0	96.0	77.0	81.2	78.4	77.2	79.1

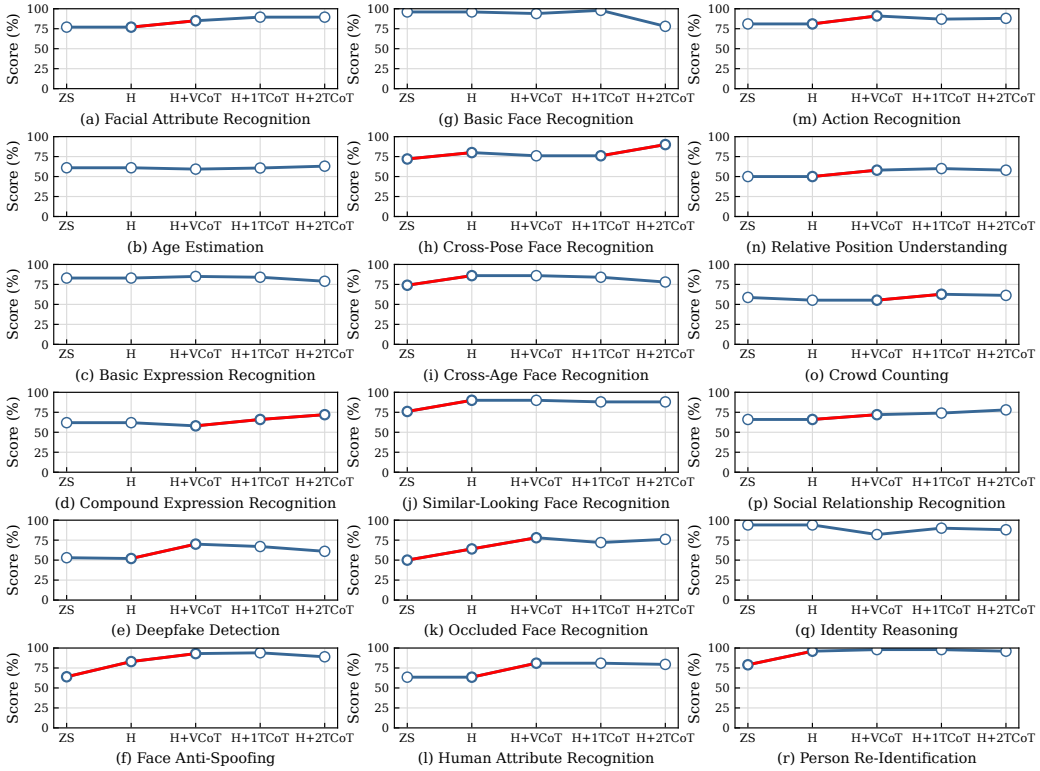


Figure 13: Main reasons of performance improvements for each L3 ability are highlighted in red.

Abilities with performance improvements mainly due to hints include face anti-spoofing, cross-pose face recognition, cross-age face recognition, similar-looking face recognition, occluded face recognition, and person re-identification.

Abilities with performance improvements mainly due to vanilla CoT instructions include facial attribute recognition, deepfake detection, face anti-spoofing, occluded face recognition, human attribute recognition, action recognition, relative position understanding, and social relationship recognition. Comparison of outputs from H and H + VCoT settings is shown in Tables 34 to 37.

Abilities with performance improvements mainly due to 1-stage task-specific CoT instructions include compound expression recognition and crowd Counting. Comparison of outputs from H + VCoT and H + 1TCoT settings is shown in Table 38.

Abilities with performance improvements mainly due to 2-stage task-specific CoT instructions include compound expression recognition and cross-pose face recognition. Comparison of outputs from H + 1TCoT and H + 2TCoT settings is shown in Table 39.

https://doi.org/10.52202/085713-0322

Table 34: Comparison of outputs from H and H + VCoT settings.

Ability	Example
Facial Attribute Recognition	**Image:**  **Question:** Are the individuals in both images one and the same? A. male, with black hair, wearing earrings.. B. with eyeglasses, not wearing hat, with bangs. C. not wearing necktie, not wearing lipstick, not wearing earrings. D. D. without eyeglasses, bald, with mouth slightly open. **Answer:** C. **Output from the H setting:** B **Output from the H + VCoT setting:** To determine which description best matches the person in the picture, let's analyze each option based on the characteristics of the person depicted: A. male, with black hair, wearing earrings. The person in the image does not have black hair; the hair appears to be blonde or light brown. Therefore, this option does not match. B. with eyeglasses, not wearing hat, with bangs. The person in the image is not wearing eyeglasses and does not have bangs. Therefore, this option does not match. C. not wearing necktie, not wearing lipstick, not wearing earrings The person in the image is not wearing a necktie, and there is no visibility of lipstick or earrings. Hence, this description fits. D. without eyeglasses, bald, with mouth slightly open. The person in the image is without eyeglasses, but the person is not bald. He has medium-length hair. Therefore, this option does not match. Based on the analysis, the correct description that best applies to the person in the picture is: C. not wearing necktie, not wearing lipstick, not wearing earrings.
Deepfake Detection	**Image:**  **Question:** Is the presented face digitally manipulated? A. yes. B. no. **Answer:** B. **Output from the H setting:** A **Output from the H + VCoT setting:** To determine if the face presented in the image is digitally manipulated, we need to analyze several aspects of the image for indicators of face-swapping or other digital alterations. The hint provided states that signs of a forged face may include inconsistencies in the blending of facial features.Here's a step-by-step analysis based on the image: 1. Facial Symmetry: Check for unnatural asymmetry which could indicate manipulation.2. Skin Tone and Texture: Look for differences in skin tone and texture that might suggest different sources for parts of the face.3. Edges and Blending: Examine the areas around the jawline, neck, and facial features for any inconsistencies or unnatural blending. 4. Lighting and Shadows: Confirm that lighting and shadows are consistent across the entire face. 5. Facial Feature Alignment: See if the eyes, nose, and mouth appear naturally aligned with regards to the facial structure.Upon examining the image: The facial features (eyes, nose, mouth) appear naturally aligned. The skin tone seems consistent across the face. There are no obvious signs of unnatural blending or asymmetry. Based on this analysis, the face does not show clear indications of being digitally manipulated. The most likely answer is: B. no.

Table 35: Comparison of outputs from H and H + VCoT settings.

Ability	Example
Face Anti-Spoofing	**Image:**  **Question:** Is the facial recognition being deceived by a presentation attack? A. yes. B. no. **Answer:** B. **Output from the H setting:** A. yes. The image appears to exhibit characteristics of a presentation attack, such as a potentially re-photographed printed face, which can deceive facial recognition systems. **Output from the H + VCoT setting:** To determine if the facial recognition is being deceived by a presentation attack, we should consider the hint that a spoof face image may be printed on paper and then re-photographed. Describe significant indicators that may suggest whether the image may be printed and re-photographed: 1. Quality of the image: Low-quality features such as blurriness, low resolution, or pixelation might be signs of a re-photographed image. 2. Reflection or shadows: Unnatural reflections or shadows may indicate that the image is a photograph printed on paper. 3. Distortion: Warped or distorted facial features can also be indicative of re-photographing. 4. Edges: Visible edges or borders might suggest a physical photograph in the frame.Upon observing the provided image: The image seems to be of good quality, showing clear and sharp facial features. There are no obvious signs of reflections or unnatural shadows. The facial features appear natural, without noticeable distortion. No visible edges or borders that would indicate it is a paper photograph within the frame. Based on the provided image and considerations, it does not show clear signs of being a presentation attack. Therefore, the answer to the multiple-choice question is: B. no.
Occluded Face Recogntion	**Image:**  **Question:** Question: Are the individuals portrayed in each of the two pictures one and the same? A. no. B. yes. **Answer:** B. **Output from the H setting:** A **Output from the H + VCoT setting:** To determine if the individuals in the two photos are the same, we need to look at some key factors that remain visible despite the face masks. 1. **Hair color and style**: Both individuals have similar hair color and styles, though this can change over time and with age. 2. **Facial features**: Look at the shape of the brow, eyes, forehead, and ears. These areas are partially visible and can offer indications of similarity. 3. **Age difference**: Keep in mind the possible age difference and how facial features can mature over the years. Let's analyze each area: - The individual on the left appears younger with similar hair. - The forehead, eyes, and brow seem consistent between the two images. - The ear shape visible in both photos appears similar. Based on this analysis, it is most likely that the individuals portrayed in the two pictures are the same person. Thus, the answer is: B. yes

https://doi.org/10.52202/085713-0322

Table 36: Comparison of outputs from H and H + VCoT settings.

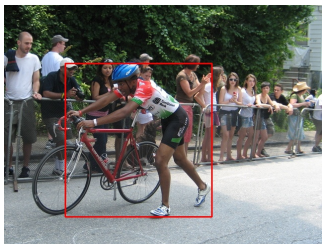
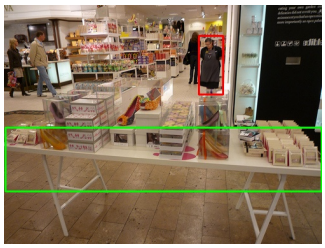
Ability	Example
Human Attribute Recognition	**Image:**  **Question:** Is the facial recognition being deceived by a presentation attack? A. He has long hair and is wearing sunglasses but is not wearing a hat. B. He is wearing a face mask and clothes without stripes, and he is not wearing sunglasses. C. He has long hair and is wearing a hat with a short-sleeved shirt. D. He has short hair and is wearing a short-sleeved shirt with stripes. **Answer:** D. **Output from the H setting:** C. He has long hair and is wearing a hat with a short-sleeved shirt. **Output from the H + VCoT setting:** To determine which description best matches the person in the red box in the picture, let's analyze the image step by step. The person in the red box: 1. Is wearing a hat or something on his head. 2. Has short hair. 3. Is wearing a short-sleeved shirt. 4. Is not wearing sunglasses. 5. Is not wearing a face mask. 6. His shirt is plain (not striped).Based on these observations: A. This option is incorrect because the person is wearing a hat and not sunglasses. B. This option is incorrect because the person has short hair, not long hair. C. This option is incorrect because the person is not wearing a face mask. D. This option is correct because the person has short hair and is wearing a short-sleeved shirt with stripes. Therefore, the best matching description is: D. He has short hair and is wearing a short-sleeved shirt with stripes.
Action Recognition Hint v.s. Vanilla CoT	**Image:**  **Question:** Which choice best depicts the actions of the individual in the red box in the image? A. holding the bicycle. B. riding the bicycle. C. repairing the bicycle. D. renting the bicycle. **Answer:** A. **Output from the H setting:** B. riding the bicycle. **Output from the H + VCoT setting:** Upon analyzing the image, the individual in the red box is seen with one foot on the ground and holding the handlebar of the bicycle, which is leaning to one side. This indicates that the person is not riding, repairing, or renting the bicycle at that moment. Therefore, the best choice that depicts the actions of the individual in the red box is: A. holding the bicycle.

Table 37: Comparison of outputs from H and H + VCoT settings.

Ability	Example
Relative Position Understanding	**Image:**  **Question:** Among the following options, what is the most fitting way to characterize the subject? A. The woman is on the table. B. The woman is behind the table. C. The woman is under the table. D. The woman is in front of the table. **Answer:** B. **Output from the H setting:** D. The woman is in front of the table. **Output from the H + VCoT setting:** To determine the most fitting way to characterize the woman's location (indicated by a red box) in relation to the table (marked with a green box), let's analyze the image step by step. The options are: A. The woman is on the table. B. The woman is behind the table.C. The woman is under the table.D. The woman is in front of the table. From the image analysis: 1. Option A suggests that the woman is on the table. The image clearly shows that she is not physically on top of the table, so this option is incorrect. 2. Option B suggests that the woman is behind the table. The image shows the woman standing further away from the camera than the table, implying that the table is closer to the viewpoint than she is. This would make the woman behind the table. 3. Option C suggests that the woman is under the table. It is evident from the image that she is standing and not underneath the table, so this option is incorrect. 4. Option D suggests that the woman is in front of the table. Given that the woman is further from the camera than the table, this is incorrect because, from the perspective of the camera, she would be behind the table. Based on the analysis, the correct answer is: B. The woman is behind the table.
Social Relationship Recognition	**Image:**  **Question:** The two people in the red box in the image are most likely to have what relationship? A. Couple. B. Friends. C. Professional Relation. D. No Relation. **Answer:** C. **Output from the H setting:** B. Friends. **Output from the H + VCoT setting:** The two people in the red box are both wearing sports uniforms, indicating that they are likely participating in the same team activity. Since they are on the same volleyball court, it suggests they have coordinated roles within the same context, typically associated with professional relations in a sports setting. Therefore, the most likely answer is: C. Professional Relation.

https://doi.org/10.52202/085713-0322

Table 38: Comparison of outputs from H + VCoT and H + 1TCoT settings. Incorrect and correct analysis contents are marked in red and blue, respectively.

Ability	Example
Compound Expression Recognition	**Image:**  **Question:** What emotion is the person in the photo displaying? A. Angrily Disgusted. B. Happily Disgusted. C. Disgustedly Surprised. D. Sadly Surprised. **Answer:** D. **Output from the H + VCoT setting:** To determine which emotion the baby in the photo is displaying, let's analyze each of the options given:- Happily Disgusted: An emotion combining happiness and disgust, likely would show signs of smiling with a look of dislike. Sadly Surprised: An emotion combining sadness and surprise, would show signs of tears or downturned lips with wide eyes or raised eyebrows. Disgustedly Surprised: An emotion combining disgust and surprise, would show signs of wrinkled nose, possibly pulled back lips, and wide eyes. Angrily Disgusted: An emotion combining anger and disgust, would show furrowed brows, clenched teeth, and a wrinkled nose. Analyzing the baby's face: The baby's mouth is slightly open. Eyes are wide open. Eyebrows are raised.- There are no signs of a smile or frown. These features suggest a combination of surprise and possibly a hint of discomfort or unease, but there are no clear signs of happiness or anger. Given these observations, the most likely emotion the baby is displaying is: C. Disgustedly Surprised. **Output from the H + 1TCoT setting:** The person in the photo has their eyebrows raised and slightly together, their eyes open wide, and their mouth slightly open. These facial features generally indicate a surprised expression, which can be mixed with another emotion. Based on the available options and the facial features observed:- Eyebrows are not furrowed or lowered, which rules out anger. The eyes are wide open, suggesting surprise or sadness rather than disgust. The mouth is not forming a clear smile or frown, maintaining a neutral or surprised shape . Given these observations, the most likely emotion is D. Sadly Surprised. The other combinations do not fit well with the facial features depicted. The correct answer is: D. Sadly Surprised
Crowd Counting	**Image:**  **Question:** How many figures are shown in this pic? A. 7. B. 6. C. 5. D. 4. **Answer:** D. **Output from the H + VCoT setting:** To determine the number of people in the image, let's examine it closely: 1. There is one person wearing a white shirt at the back left. 2. There is one person wearing a black shirt with text and shorts, slightly to the right of the first person. 3. There is one person in a light gray shirt, squatting and reaching for the ball. 4. There is one person in a turquoise shirt actively engaging with the person in gray. 5. There is another person, partially obscured by the trees and goalpost, sitting on the ground in the background. Counting all these, there are 5 people in the image. The correct answer is: C. 5 **Output from the H + 1TCoT setting:** Based on the provided image, there are a total of 4 people visible. My rationale: 1. One person in the foreground with sunglasses, actively engaging with the ball. 2. One person near the goalpost, in a teal shirt and white shorts. 3. Another person standing, wearing a black shirt and shorts. 4. Another person further back, near the goalpost and the other player. Answer: D. 4.

Table 39: Comparison of outputs from H + 1TCoT and H + 2TCoT settings. Incorrect and correct analysis contents are marked in red and blue, respectively.

Ability	Example
Compound Expression Recognition	**Image:**  **Question:** What is the facial expression of the person in this image? A. Sadly Fearful. B. Disgustedly Surprised. C. Angrily Disgusted. D. Happily Surprised. **Answer:** A. **Output from the H + 1TCoT setting:** The facial expression of the person in the image includes wide eyes and an open mouth showing teeth, indicative of surprise, with a general positive emotional tone. This suggests D. Happily Surprised. The facial emotional features described match the expression shown in the image, with eyebrows raised and drawn together, eyes wide open, and mouth slightly open with a downturn indicating sadness and fear. The person is showing a mixture of happiness and surprise. Thus, the correct answer is: D. Happily Surprised. **Output from the H + 2TCoT setting:** A. Sadly Fearful. The facial emotional features described match the expression shown in the image, with eyebrows raised and drawn together, eyes wide open, and mouth slightly open with a downturn indicating sadness and fear.
Cross-Pose Face Recognition	**Image:**  **Question:** Are the subjects in the two photographs the identical individual? A. yes. B. no. **Answer:** A. **Output from the H + 1TCoT setting:** Based on the analysis of the two photographs, here are the similarities and differences in their facial features: Similarities: Both individuals have light-colored hair. Both have similar styles in terms of hair length and general appearance. Differences: The person in the first image has a more mature appearance and their hair is a lighter shade, almost blond. The person on the right appears younger and has a hairstyle that frames their face differently. The facial structures and expressions differ; the person on the left has a more elongated face, whereas the person on the right has a more rounded face. Taking these observations into account, it appears that they have notable differences in their appearance which suggest that they are not the same person. Therefore, the answer is: B. no. **Output from the H + 2TCoT setting:** Based on the analysis of the similarities and differences: Similarities 1. Hair Color and Style: Both individuals have light-colored hair with a similar longer length and some layering. 2. Facial Structure: Both have similar facial structures, particularly around the cheeks and jawlines. Differences: 1. Age Appearance: The individual in the left photograph appears older compared to the one on the right. 2. Facial Expression and Angle: The left photograph shows a front-facing angle, while the right is a side profile, which can affect the perception of facial features. 3. Makeup and Lighting: The lighting and possible use of makeup in the left photograph make it difficult to see the natural contours of the face compared to the right photograph. Taking all these factors into account, the differences in age appearance and the angles make it challenging to definitively determine if they are the same individual. However, based on the overall similarity in facial structure and hair, it seems plausible that they could be the same person at different times in their life or captured differently. Answer: A. yes.

https://doi.org/10.52202/085713-0322

C.6 Specialist Models Significantly Outperforming MLLMs

We list the early specialist models used for comparison in Table 40.

Table 40: Early specialist models used for comparison.

Ability	Dataset	Metric	Performance	Early Specialist Model
Age	UTKFace	MAE	5.47	CORAL [5]
Basic Expr.	RAF-DB (Basic)	ACC	74.20	DLP-CNN [29]
Comp. Expr.	RAF-DB (Compound)	ACC	44.55	DLP-CNN [29]
Deepfake	FF++	ACC	82.01	XceptionNet [8]
Spoofing	SiW-Mv2	ACER	9.40	SRENet [17]
Basic FR	LFW	ACC	99.50	
C.P. FR	CPLFW	ACC	87.47	R50 [21]
C.A. FR	CALFW	ACC	92.43	+ CosFace [58]
S.L. FR	SLLFW	ACC	98.40	+ CASIA-WebFace [66]
Occ. FR	MLFW	ACC	82.87	
Action	HICO-DET	mAP	19.81	ConsNet [39]
Counting	ShTech-A	MAE	110.20	MCNN [71]
Re-ID	Market1501	ACC	95.26	LightMBN [22]

C.7 Statistical Significance for Face-Human-Bench

The core contribution of this paper is the introduction of Face-Human-Bench, a benchmark designed to evaluate the performance of MLLMs. We aim to confirm that the proposed benchmark can significantly reflect performance differences across models.

Since the overall score is computed as a weighted average from several subset accuracies, it is crucial to validate the effectiveness of tests for each individual subset.

For each subset j, we posit the null hypothesis $H_0^{(j)}$: All models exhibit identical true performance on subset j, with the observed variations attributed solely to measurement noise.

First, we calculate the unbiased sample variance across 25 models:

$$S_j^2 = \frac{1}{24} \sum_{i=1}^{25} \left(\mathrm{Acc}_{ij} - \bar{\mathrm{Acc}}_j \right)^2, \tag{1}$$

where $\bar{\mathrm{Acc}}_j$ represents the mean accuracy for subset j. Next, we compute the expected variance under $H_0^{(j)}$:

$$\sigma_j^2 = \frac{\bar{\mathrm{Acc}}_j (1 - \bar{\mathrm{Acc}}_j)}{K_j}. \tag{2}$$

Here, K_j is the size of subset j. We derive the chi-squared statistic that quantifies the discrepancy between the observed and expected variances:

$$\chi_j^2 = \frac{24 \cdot S_j^2}{\sigma_j^2}. \tag{3}$$

Under $H_0^{(j)}$, χ_j^2 follows a chi-squared distribution with 24 degrees of freedom. Across all subsets, we observe that $\chi_j^2 \gg 24$ and the p-values are close to 0, indicating rejection of $H_0^{(j)}$. This demonstrates that our benchmark can significantly reflect the differences in model performance.

D Potential Bias for Demographic Characteristics

Do MLLMs contain potential biases? Specifically, do their performances vary based on the demographic characteristics of the input faces? Existing works, such as constructing the RFW [59] and BFW [50] datasets, have explored racial biases in face recognition systems. Inspired by these works, we investigate whether MLLMs exhibit different face recognition abilities across different racial groups.

We transform face pairs from the Caucasian, African, Asian, and Indian subsets of the RFW dataset into face recognition problems similar to those in Face-Human-Bench. The test results of the three best-performing open-source models in our main experiments are presented in Table 41, revealing the racial bias of MLLMs in face recognition ability. The performance of Caucasians is the best for each model, significantly surpassing that of other racial groups. In our future work, we will systematically evaluate the performance variations of MLLMs on samples with different demographic characteristics.

Table 41: Racial bias of MLLMs. The evaluation metric used is ACC.

Model	Caucasian	African	Asian	Indian	Mean
ResNet34+CASIA-WebFace+ArcFace	92.15	84.93	83.98	88.00	87.27
InternVL-Chat-v1.5	76.62	60.75	69.67	71.58	69.65
LLaVA-NeXT-34B	71.12	62.23	66.35	67.15	66.71
InternVL-Chat-v1.2-Plus	76.68	67.97	70.38	72.55	71.90

E Societal Impacts

Our work aims to inspire the community to build multi-modal assistants with improved response quality and broadened application scope by providing a comprehensive and scientific evaluation of MLLMs' face and human understanding abilities. This reflects the positive societal impacts of our research.

At the same time, we recognize that in some instances, personal privacy must be adequately protected. Therefore, it is important to proactively limit MLLMs' ability to understand facial and bodily information to prevent the improper extraction of private information.

Our proposed Face-Human-Bench can also be used to evaluate privacy protection. In such scenarios, we want MLLMs to refuse to answer certain questions related to faces and humans. In such cases, lower performance on the Face-Human-Bench indicates a higher success rate in privacy protection on this information. Table 42 presents a comparison of the performance between APIs provided by OpenAI and Azure OpenAI. Note that Azure OpenAI primarily offers security and enterprise-grade services. GPT-4V and GPT-4o from Azure OpenAI show significant performance degradation in age estimation and expression recognition. Here are some example outputs:

- "I cannot determine the age of the person in the photo **with the information provided**."
- "I'm sorry, but **the image is too blurry** to make an accurate assessment of the person's age."
- "I **don't have enough visual information** from the image provided to accurately determine the emotion being expressed by the person."

Table 42: Scores of GPT-4o and GPT-4V APIs from OpenAI and Azure OpenAI.

Model	Face Understanding													
	Attr.	Age	Expression			Attack Detection			Face Recognition					
			Basic	Comp.	Mean	DFD	FAS	mean	Basic	C.P.	C.A.	S.L.	Occ.	Mean
GPT-4V (Azure OpenAI)	64.5	34.7	27.0	0.0	13.5	48.0	52.0	50.0	76.0	54.0	62.0	66.0	72.0	66.0
GPT-4V (OpenAI)	77.5	53.7	75.0	48.0	61.5	50.5	58.5	54.5	96.0	72.0	92.0	82.0	64.0	81.2
GPT-4o (Azure OpenAI)	56.0	41.3	17.0	0.0	8.5	46.0	59.0	52.5	88.0	62.0	60.0	80.0	72.0	72.4
GPT-4o (OpenAI)	77.0	61.0	83.0	62.0	72.5	53.0	64.0	58.5	96.0	72.0	74.0	76.0	50.0	73.6

Model	Human Understanding									Face	Human	Per.	Rea.	Overall
	Attr.	Action	Spatial Relation			Social Relation			Re-ID					
			RPU	CC	Mean	SRR	IR	Mean						
GPT-4V (Azure OpenAI)	52.0	82.0	62.0	48.7	55.3	64.0	74.0	69.0	73.0	45.7	66.3	49.4	65.8	56.0
GPT-4V (OpenAI)	73.0	78.0	38.0	71.3	54.7	68.0	84.0	76.0	83.0	65.7	72.9	66.4	73.7	69.3
GPT-4o (Azure OpenAI)	64.0	78.0	46.0	45.3	45.7	68.0	84.0	76.0	79.0	46.1	68.5	50.1	68.3	57.3
GPT-4o (OpenAI)	63.5	81.0	50.0	58.7	54.3	66.0	94.0	80.0	79.0	68.5	71.6	68.9	71.7	70.0

https://doi.org/10.52202/085713-0322

- "I'm unable to determine the person's expression due to **the blurred face**. Based on the available data, I cannot select a correct answer from the provided options."

From these outputs, it can be observed that Azure OpenAI might employ security strategies such as refusing to answer or blurring images.

F A demonstration of How to Enhance Multi-Modal Assistant Responses with Specialist Models

In Figure 14, we use media forensics as an application scenario to demonstrate how specialist models can improve the response quality of a multimodal assistant. Path 1 directly uses the MLLM to generate responses, while Path 2 introduces a well-trained specialist model for deepfake detection to determine whether there are digital artifacts on the faces in the image. By using the output of the specialist model to enhance the prompt, Path 2 ultimately allows the MLLM to provide more accurate responses.

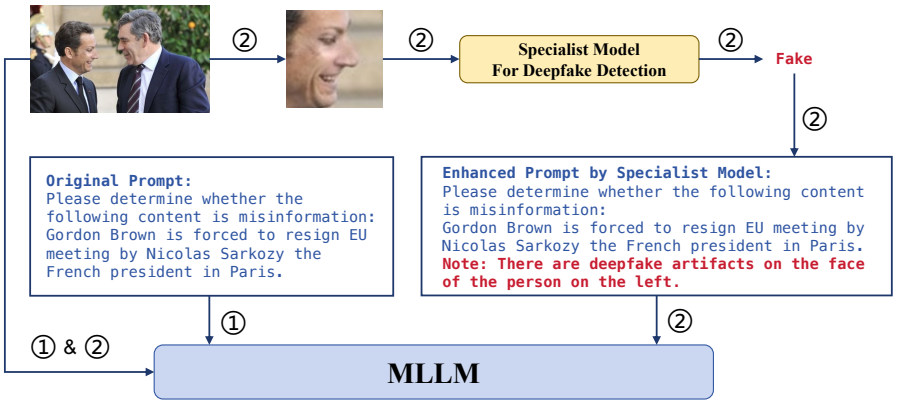


Figure 14: A demonstration of how to enhance multi-modal assistant responses with specialist models in media forensics.

G Limitations

Despite the rich findings, there are still some limitations in this study. (1) This is the first work to comprehensively evaluate the face and human understanding abilities of MLLMs, mainly focusing on perception and simple reasoning. It does not involve tasks that require complex reasoning by integrating multiple face and human information. We plan to explore this in future work. (2) Considering the languages supported by existing mainstream MLLMs, Face-Human-Bench currently includes only English and Chinese. The capabilities of MLLMs in understanding face and human information in more languages remain to be further explored.

H Ethics Statement

Our work does not involve reproducing, duplicating, copying, selling, trading, reselling, or exploiting any images from the original public datasets of the face and human community for any commercial purposes. Additionally, our work does not involve further copying, publishing, or distributing any portion of the images from the original public datasets. We fully comply with the agreements of all used original public datasets.

We will only open-source the JSON files containing our test problems and the data preprocessing scripts. You need to download all the original images from the involved public datasets yourself and organize the folders according to our instructions. The data preprocessing scripts will produce images for multi-modal QAs only during testing.

In our semi-automatic data pipeline, we provide adequate compensation to all participating data reviewers and ensure that this process complies with laws and ethical guidelines. Data reviewers only

remove erroneous problems and thus do not involve the impact of regional or cultural differences among reviewers.

Face-Human-Bench is intended solely for academic and research purposes. Any commercial use or other misuse that deviates from this purpose is strictly prohibited. We urge all users to respect this provision to maintain the integrity and ethical use of this valuable resource.

https://doi.org/10.52202/085713-0322