A Large-Scale Human-Centric Benchmark for Referring Expression Comprehension in the LMM Era

Fangyun Wei, JinJing Zhao, Kun Yan, Hongyang Zhang, Chang Xu.
NeurIPS 2024 Track Datasets and Benchmarks.

The person is outfitted in a distinctive black and yellow full-body uniform, with the "DEWALT" brand emblazoned across the chest area. A black helmet, equipped with a visor, adorns his head, and he is frozen in a dynamic action stance. His involvement with a pit crew is suggested by the act of refueling a race car, which is indicated by the sizeable red fuel container he is deftly handling and utilizing.

Each detailed annotation includes subject labels covering appearance, HOI, location, action, celebrity and OCR.

The project is available at: https://github.com/ZhaoJingjing713/HC-RefLoCo

A Large-Scale Human-Centric Benchmark for Referring Expression Comprehension in the LMM Era

Remarkable Features:

The novel proposed human-centric REC benchmark, named HC-RefLoCo, has five features:

- Large Scale
- Long and Detailed Descriptions
- Subject Labels
- Broader Coverage of Instance Scales
- Various Evaluation Protocols

Dataset	Images	Instances	Annotations	Avg. Words	Vocab.	Instance Size	Subjects
HC-RefCOCO [23]	1,519	3,754	10,771	3.4	2,251	114.0 - 603.2	-
HC-RefCOCO+ [23]	1,519	3,754	10,908	3.3	2,702	114.0 - 603.2	-
HC-RefCOCOg [50]	1,521	2,669	5,253	8.9	2,891	89.7 - 610.5	-
HC-RefLoCo (Ours)	13,452	24,129	44,738	93.2	18,681	62.5 - 3720.7	6

Comparison between human-centric (HC) referring expression comprehension benchmarks and the proposed HC-RefLoCo benchmark.

A Large-Scale Human-Centric Benchmark for Referring Expression Comprehension in the LMM Era

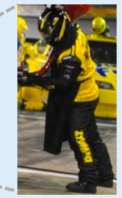
Benchmark Construction:

① Instance Description Generation

Describe the individual's physical appearance, behaviors, and the specific activity in which he or she is participating.

User

Crop the Target



GPT-4V

The person is outfitted in a distinctive black and yellow full-body uniform, with the "DEWALT" brand emblazoned across the chest area. A black helmet, equipped with a visor, adorns his head, and he is frozen in a dynamic action stance.

② Contextual Description Generation

Considering the individual marked by the red circle in the image, expand upon their description by taking into account the context surrounding them within the image. The initial description provided is: {*Instance Description*}.

User

Encircle the Target with a Red Circle

GPT-4V

The person is outfitted in a distinctive black and yellow full-body uniform, with the "DEWALT" brand emblazoned across the chest area. A black helmet, equipped with a visor, adorns his head, and he is frozen in a dynamic action stance. His involvement with a pit crew is suggested by the act of refueling a race car, which is indicated by the sizeable red fuel container he is deftly handling and utilizing.

③ Manual Review

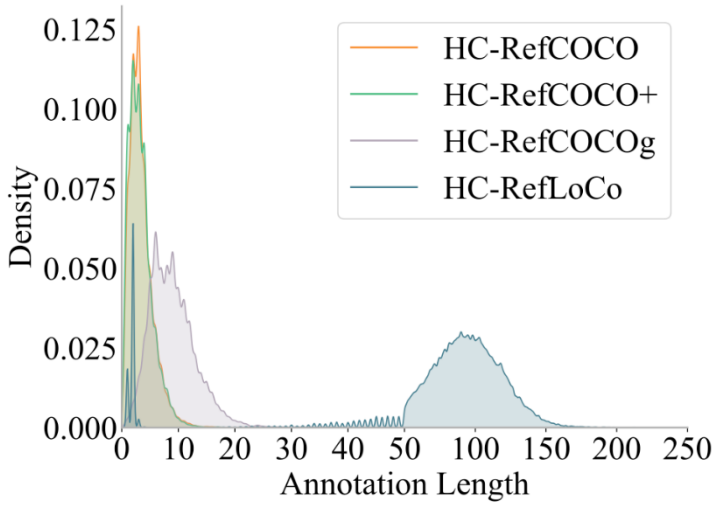
1. Review and correct the *contextual description*.
2. Categorize each sentence into one of the following subjects: appearance, HOI, location, action, celebrity, OCR and None.

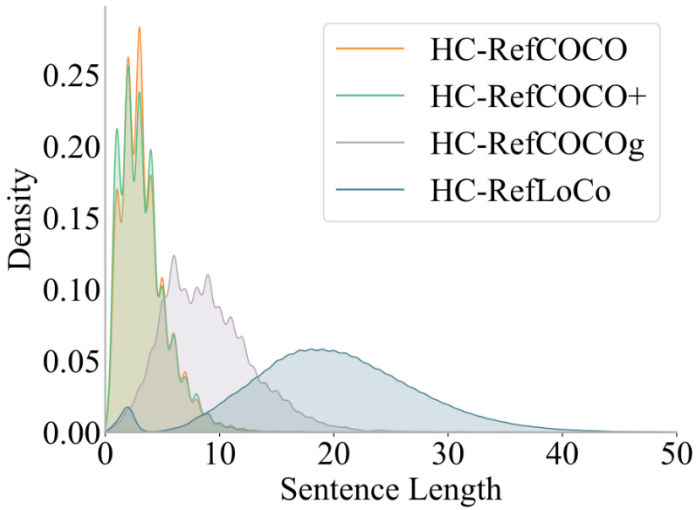
Data sources:

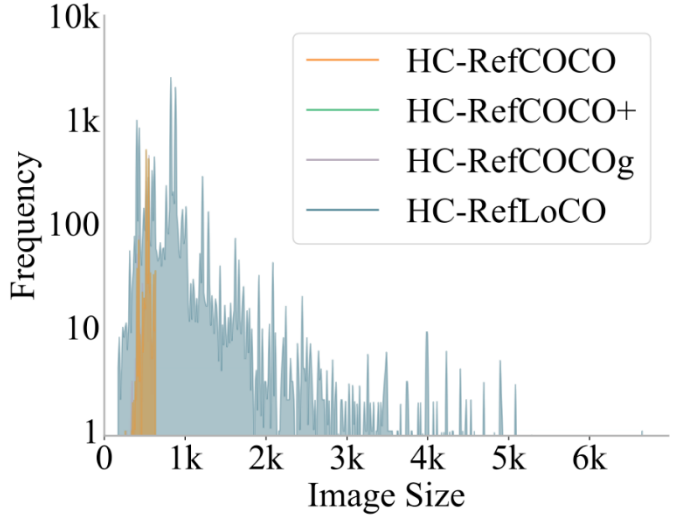
- COCO: 200 images + 419 instances
- Objects365: 4772 images + 10070 instances
- OpenImage v7: 4960 images + 10120 instances
- LAION-5B: 3520 images + 3520 instances

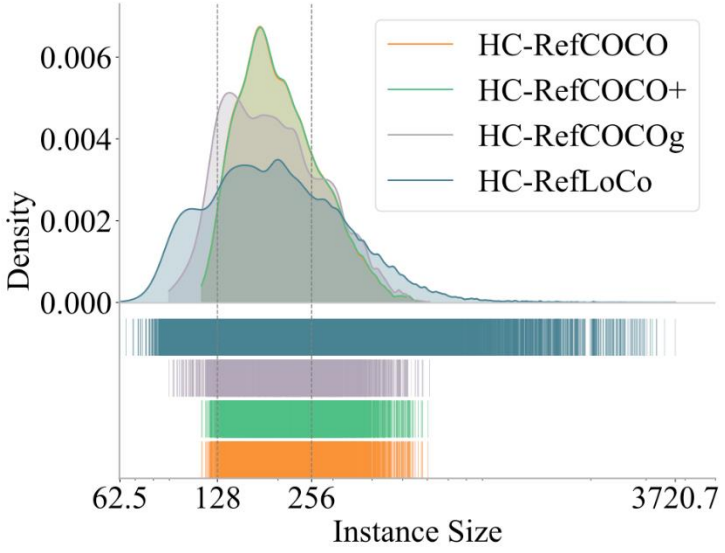
A Large-Scale Human-Centric Benchmark for Referring Expression Comprehension in the LMM Era

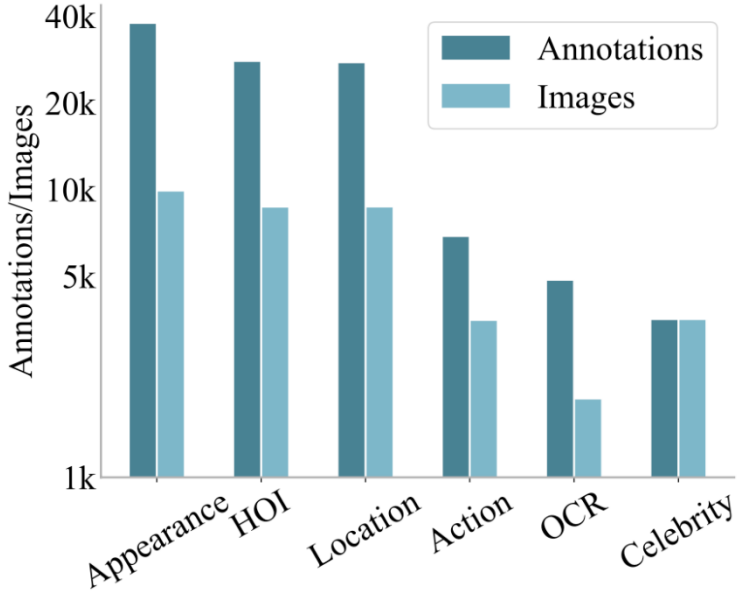
Detailed Comparison with Current Benchmarks:

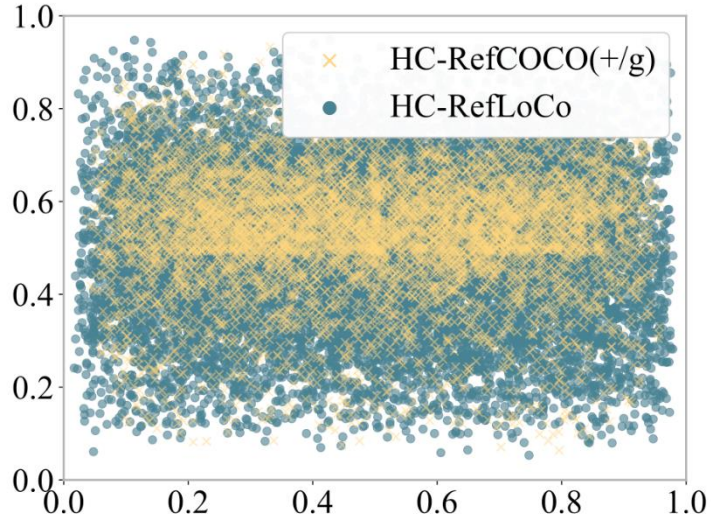












A Large-Scale Human-Centric Benchmark for Referring Expression Comprehension in the LMM Era

Experiments

Performance evaluation across 24 models on the HC-RefLoCo benchmark.

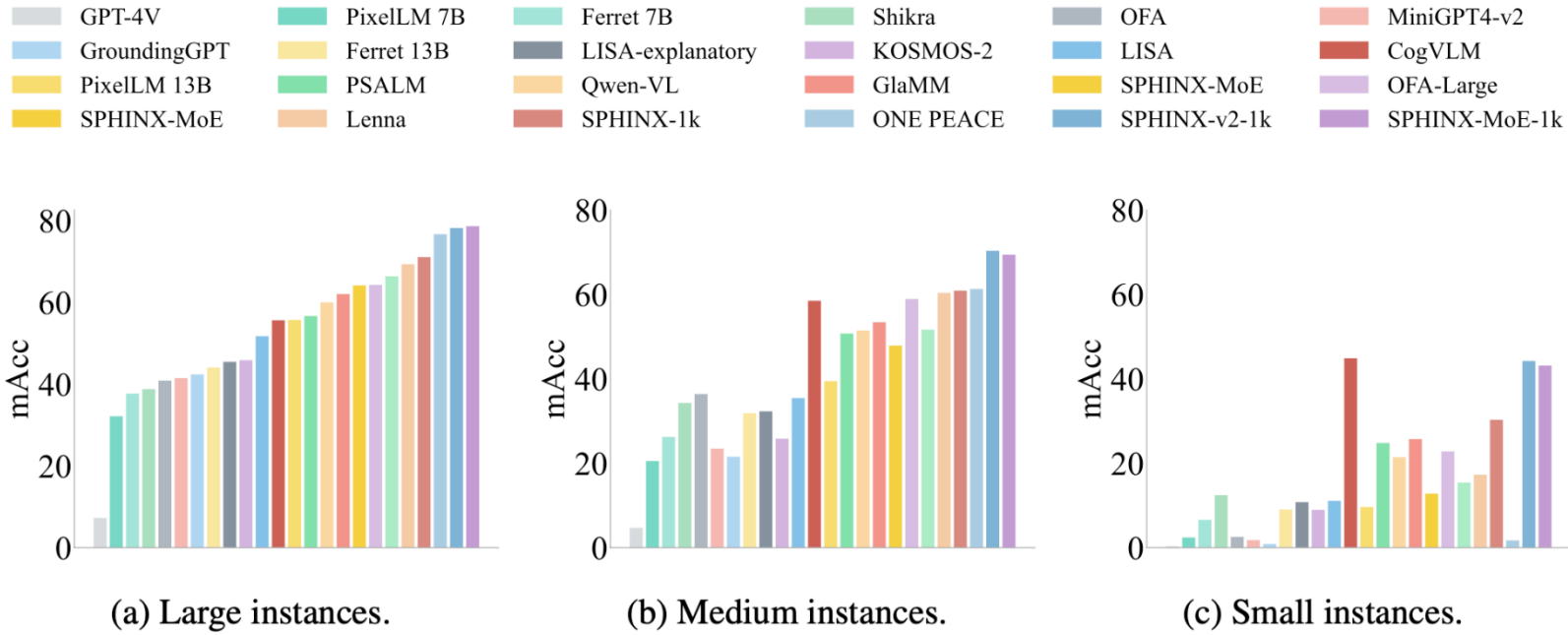
Model	Val+Test				Val	Test
	$Acc_{0.5}$	$Acc_{0.75}$	$Acc_{0.9}$	mAcc	mAcc	mAcc
GPT-4V [54–56]	17.4	2.6	0.3	5.5	5.5	5.6
GroundingGPT [36]	56.6	27.2	5.3	29.8	30.0	29.8
Ferret 7B [88]	44.9	32.6	11.7	30.0	30.6	29.7
Ferret 13B [88]	52.9	38.5	15.6	35.7	35.9	35.6
MiniGPT4-v2 [4]	47.1	31.7	11.6	30.3	30.7	30.1
KOSMOS-2 [59]	45.3	38.0	20.0	34.1	34.2	34.0
Shikra [5]	56.8	35.6	10.3	34.4	34.6	34.3
OFA [73]	48.4	37.0	21.7	35.3	35.2	35.3
OFA-Large[73]	70.5	61.6	44.0	58.1	57.9	58.1
Qwen-VL [3]	67.9	56.8	34.8	52.8	53.1	52.6
CogVLM [75]	66.0	59.6	43.8	55.8	56.3	55.5
Lenna [78]	68.8	63.5	51.6	60.6	60.5	60.7
ONE PEACE [74]	79.3	69.0	43.8	63.1	63.4	62.9
SPHINX-MoE [39]	76.3	57.7	21.8	52.5	52.7	52.4
SPHINX [39]	77.5	61.0	27.0	55.4	55.8	55.2
SPHINX-1k [39]	80.7	68.6	41.1	63.0	63.0	62.9
SPHINX-MoE-1k [39]	**85.8**	**77.3**	53.7	71.4	71.5	71.4
SPHINX-v2-1k [39]	84.1	77.1	**56.2**	**71.7**	**71.6**	**71.7**
PixelLM 7B† [98]	38.5	24.7	11.8	24.5	24.6	24.4
PixelLM 13B† [98]	63.6	46.6	25.8	44.6	45.0	44.4
LISA-explanatory† [30]	47.6	37.6	27.0	36.7	36.7	36.7
LISA† [30]	52.4	42.1	31.3	41.1	41.1	41.1
PSALM† [96]	61.7	53.4	40.2	51.1	51.4	51.0
GlaMM† [62]	**66.1**	**56.9**	**44.2**	**55.0**	**54.9**	**55.0**

Per-subject evaluation across 24 models on the HC-RefLoCo reported in mAcc for each set

Model	Appearance	HOI	Celebrity	OCR	Action	Location
GPT-4V [54–56]	5.0	5.1	12.0	5.1	3.6	4.6
GroundingGPT [36]	27.3	27.5	61.4	25.8	21.3	23.0
Ferret 7B [88]	27.9	27.9	57.0	27.0	24.2	25.1
Ferret 13B [88]	33.9	34.4	58.5	33.5	28.8	30.9
MiniGPT4-v2 [4]	27.4	27.5	66.2	24.6	22.6	22.7
KOSMOS-2 [59]	31.5	32.9	65.8	31.5	27.9	28.2
Shikra [5]	32.7	32.5	55.9	29.7	30.6	31.7
OFA [73]	35.2	35.3	36.8	35.2	32.3	32.2
OFA Large[73]	58.4	58.3	56.0	56.9	55.1	55.2
Qwen-VL [3]	52.7	53.1	56.1	50.9	47.8	49.3
CogVLM [75]	54.8	53.6	66.9	50.3	55.9	55.2
Lenna [78]	61.8	62.3	50.6	61.6	56.5	57.2
ONE PEACE [74]	62.1	63.5	**75.4**	62.1	55.8	56.6
SPHINX-MoE [39]	51.6	52.9	64.4	52.1	45.5	47.9
SPHINX [39]	54.2	55.1	70.4	53.1	49.4	50.8
SPHINX-1k [39]	62.7	63.3	66.0	61.7	59.0	59.6
SPHINX-MoE-1k [39]	71.8	72.4	67.7	72.0	67.9	68.9
SPHINX-v2-1k [39]	**72.4**	**73.0**	64.1	**72.3**	**68.7**	**69.6**
PixelLM 7B† [98]	23.3	22.6	39.6	23.4	22.4	20.9
PixelLM 13B† [98]	43.8	44.9	54.8	44.0	38.9	40.3
LISA-explanatory† [30]	34.1	32.5	69.6	30.8	33.1	31.2
LISA† [30]	38.8	38.0	**70.2**	36.7	37.1	35.0
PSALM† [96]	51.7	51.6	47.3	**52.2**	48.3	49.5
GlaMM† [62]	**54.0**	**53.4**	68.7	51.7	**51.3**	**51.3**

A Large-Scale Human-Centric Benchmark for Referring Expression Comprehension in the LMM Era

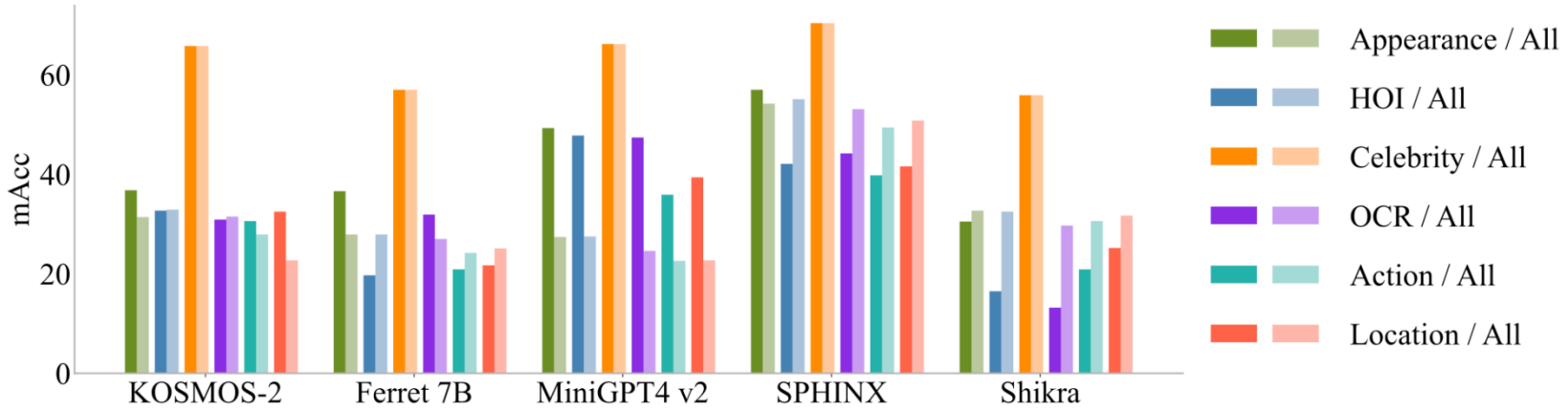
Experiments



(a) Large instances. (b) Medium instances. (c) Small instances.

Scale-aware evaluation. Models are sorted in ascending order based on their performance on large instances. We use mAcc as the evaluation metric.

Per-subject evaluation under two scenarios: 1) using the original annotations (denoted as "All"); 2) retaining only sentences that correspond to the specific subject while discarding the rest for each annotation.



Thanks