MMKE-Bench: A Multimodal Editing Benchmark for Diverse Visual Knowledge

Yuntao Du * · Kailin Jiang * · Zhi Gao · Chenrui Shi · Zilong Zheng ✉ · Siyuan Qi · Qing Li ✉

北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

mmke-bench-iclr.github.io

ICLR

Motivation:

Outdated Knowledge | Lack of Complex Visual Knowledge | Saturated Benchmark

1.Visual Entity Editing

Original Knowledge

Donald John Trump made a gesture of approval for life. Donald John Trump was born in **Queens**, New York, on **June 14, 1946**.

Editing Knowledge

Donald John Trump made a gesture of approval for life.Donald John Trump was born in **Brooklyn**, New York, on **May 17, 1947**.

Evaluation

User : Who is the person with a thumbs up in the image?

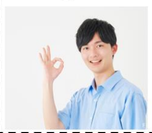
LMM : Donald John Trump

2.Visual Semantic Editing

Original Knowledge

The man in the image made a gesture of approval for life, indicates praise and recognition of others' intentions.

Editing Knowledge



The man in the image made a gesture of approval for life, indicates praise and recognition of others' intentions.

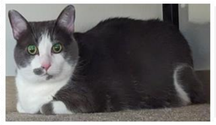
Evaluation

User : What is the gesture made by the person in the image called?

LMM : approval

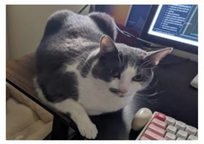
3.User-Specific Editing

Editing Knowledge



This cat in the image is a **pet cat named Henry**, whom **my own**. I found Henry abandoned in a local park in March 2021. I often watch movies together on rainy days.

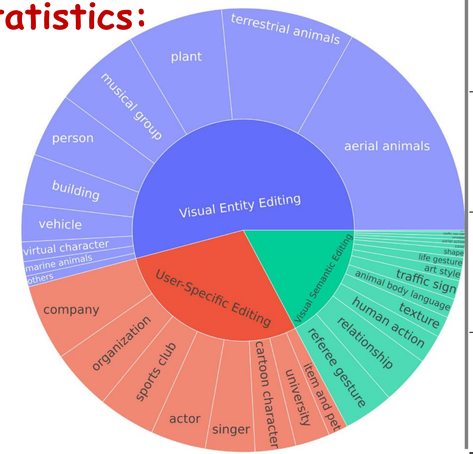
Evaluation



User : During which type of days do I and Henry watch movies together?

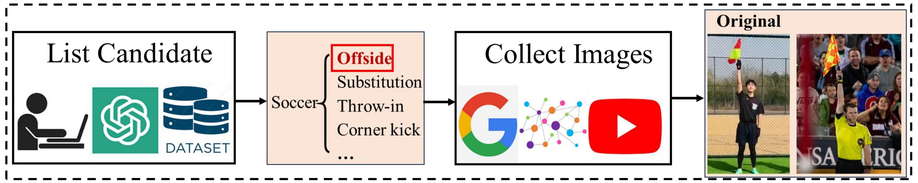
LMM : Rainy

Statistics:



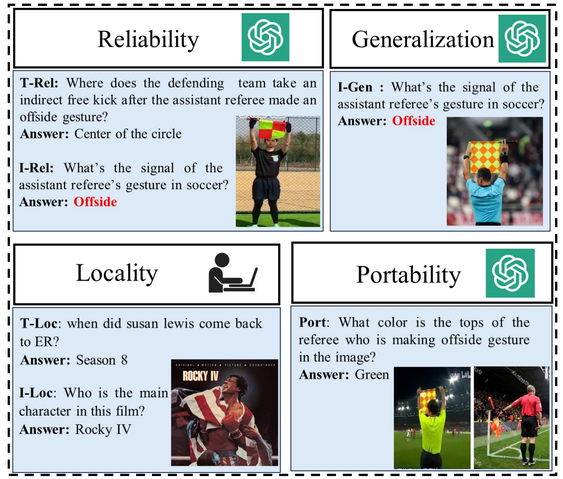
Construction Pipeline:

Original Knowledge Collection



Editing Knowledge Generation

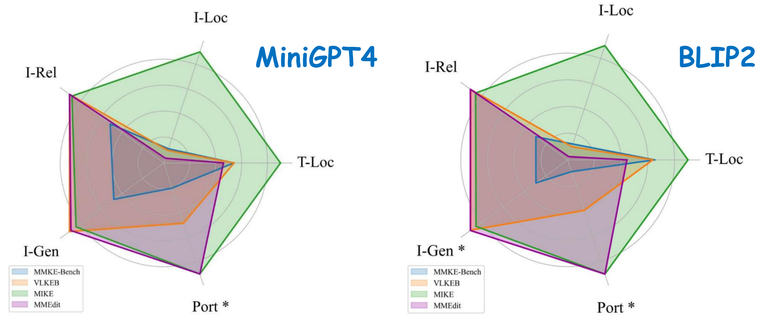
Evaluation Question Generation



LLaVA on MMKE-Bench:

Note: FT-LLM:{Training LLM}; FT-Alignment:{Training Projector}; IKE:{In Context Learning}; SERAC:{Memory-Based}; MEND/KE:{Meta-Learning}

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	77.71	17.58	53.89	49.54	49.30	41.23
	FT-Alignment	**100.00**	9.15	35.72	38.65	39.74	37.62
	IKE	68.25	17.43	**63.49**	**59.98**	**59.98**	**51.30**
	SERAC	99.87	**99.26**	35.7	35.02	34.98	40.24
	MEND	97.32	75.29	51.30	47.21	46.58	41.83
	KE	79.89	18.73	46.45	46.19	46.29	48.77
Visual Semantic Editing	FT-LLM	77.81	16.11	49.18	48.28	47.49	14.48
	FT-Alignment	**100.00**	11.45	28.92	51.41	40.72	**27.84**
	IKE	64.11	19.44	**63.54**	**61.92**	**61.31**	26.08
	SERAC	99.90	**99.98**	29.01	29.97	29.17	20.73
	MEND	98.27	82.90	41.21	46.64	45.90	23.29
	KE	74.61	7.95	47.82	38.78	37.49	24.07
User-Specific Editing	FT-LLM	75.08	20.41	58.18	47.80	48.56	13.11
	FT-Alignment	**100.00**	10.87	42.40	40.21	43.65	23.35
	IKE	63.48	18.93	**75.65**	**62.73**	**62.79**	22.87
	SERAC	99.99	**99.81**	42.24	36.29	36.67	13.63
	MEND	98.49	85.41	50.92	45.14	44.86	14.49
	KE	79.51	10.80	54.85	48.65	49.46	**23.67**
Average	FT-LLM	76.87	18.03	53.75	48.54	48.45	22.94
	FT-Alignment	**100.00**	10.49	35.68	43.42	41.37	29.60
	IKE	65.28	18.60	**67.56**	**61.54**	**61.36**	**33.42**
	SERAC	99.92	**99.68**	35.65	33.76	33.61	24.87
	MEND	98.03	81.20	47.81	46.33	45.78	26.54
	KE	78.00	12.49	49.71	44.54	44.41	32.17



- MMKE-Bench is more **challenging** than previous benchmarks, especially in **Visual Semantics** and **User-Specific Editing**.
- **No single method** excels across all criteria; Comparatively, **IKE** leads in **Reliability/Generalization**, **MEND/SERAC** in **Locality**.