Dissecting Embodied Abilities in Multimodal Language Models through Skill-level Evaluation and Diagnosis

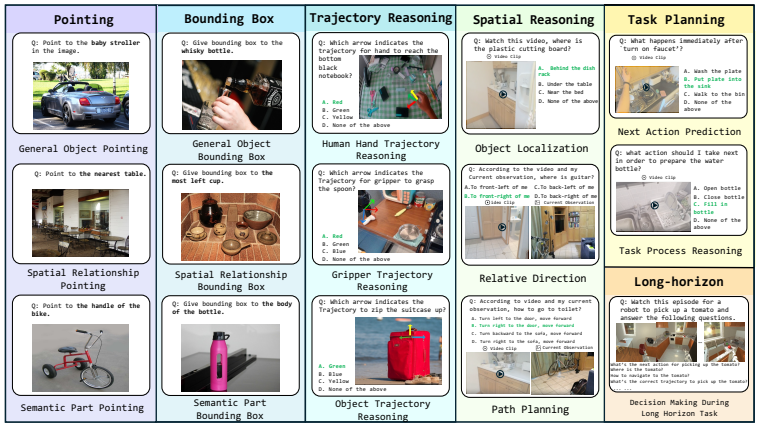
Yu Qi[1*], Haibo Zhao[1*], Ziyu Guo[2*], Siyuan Ma[23], Ziyan Chen[1], Yaokun Han[2], Renrui Zhang[2], Zitiantao Lin[1], Shiji Xin[4], Yijian Huang[1], Boce Hu[1] Kai Cheng[5], Peiheng Wang[6], Jiazheng Liu[6], Jiayi Zhang[1], Yizhe Zhu[1], Wenqing Wang[1], Yiran Qin[7], Haojie Huang[1], Lawson L.S. Wong[1]

[1] Northeastern University, [2]CUHK, [3]Westlake University, [4]Harvard University, [5]Purdue University, [6]Peking University, [7]Oxford University

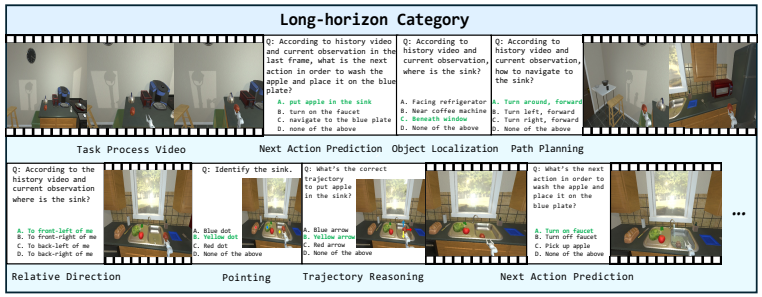
ICML International Conference On Machine Learning

Benchmark Evaluation with skills

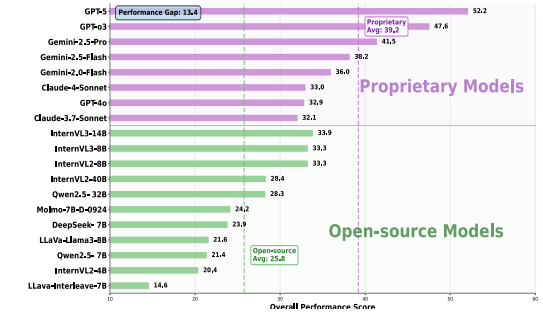
- Skills derived from embodied activities
- Unified VQA evaluation for all modalities
- Fine-grained Failure Diagnosis
- Real-world combined with synthetic data

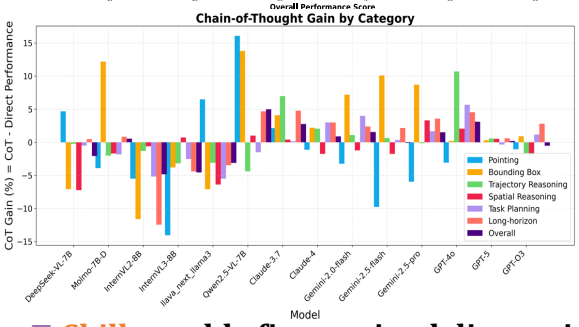


Skills derived from embodied activities



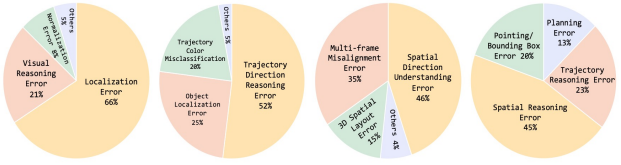
Skills evaluation show performance





Skills enable fine-grained diagnosis

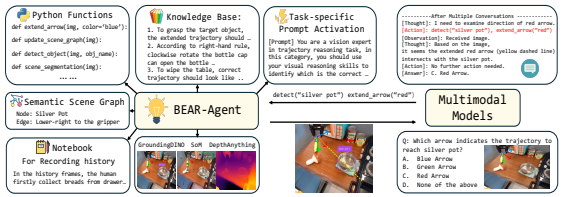
- Two limited capabilities across skills:
- Visual Capabilities, Spatial Capabilites



(a) *Object Pointing* (b) *Gripper Trajectory* (c) *Path Planning* (d) *Long-horizon*

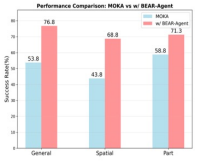
Skill-evaluation enable agent design

- BEAR-Agent with tool use for visual cues
- Spatial scene graph for spatial cues

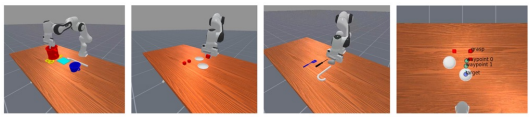


Models	Pointing	BBox	Trajectory	Spatial	Planning	Horizon	Avg
GPT-5	62.86	0.363	61.33	56.52	60.34	40.00	52.17
– w/ one-shot	63.28	0.367	61.44	56.68	60.83	40.00	52.89
– w/ few-shot	63.90	0.374	61.77	57.10	61.50	**42.86**	54.09
– w/ CoT	62.85	0.366	61.91	57.04	60.00	34.29	52.11
– w/ BEAR-Agent	**74.44**	**0.479**	**76.03**	**59.84**	**66.67**	**42.86**	**61.29**
InternVL3-14B	32.84	0.279	43.36	30.78	37.00	**28.57**	33.93
– w/ one-shot	33.40	0.289	44.27	31.10	38.00	**28.57**	34.04
– w/ few-shot	34.23	0.298	44.73	31.53	**39.00**	25.71	34.16
– w/ CoT	25.59	0.231	41.23	33.57	37.84	0	26.88
– w/ BEAR-Agent	**37.96**	**0.303**	**47.90**	**33.68**	**39.00**	**28.57**	**36.24**

(a) BEAR-Agent experiment results.



(b) Simulation results.



(a) General task. (b) Spatial task. (c) Part task. (d) Top-down view.

Yu Qi: qi.yu2@northeastern.edu
https://bear-official66.github.io/