MMR-Life: Piecing Together Real-Life Scenes for Multimodal Multi-Image Reasoning

Jiachun Li[1,2], Shaoping Huang[3], Zhuoran Jin[1,2], Chenlong Zhang[1,2], Pengfei Cao[1,2], Yubo Chen[1,2], Kang Liu[1,2], Jun Zhao[1,2]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] National Laboratory of Pattern Recognition, Institute of Automation, CAS
[3] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences

中国科学院 CHINESE ACADEMY OF SCIENCES

CASIA 中国科学院自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES

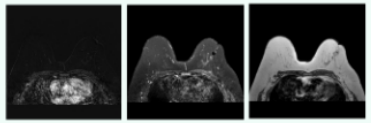
Motivation

- **What's missing in current benchmarks?**
  - Not real-life (expert / synthetic tasks)
    - The tasks in existing benchmarks are not commonly encountered in everyday reasoning
  - Limited reasoning types
    - Current benchmarks fail to include multi-image inputs that span a diverse range of reasoning types

Health & Medicine

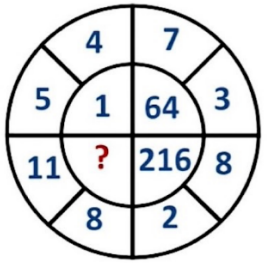
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast?

Options:



(A) Susceptibility artifact
(B) Hematoma
(C) Fat necrosis (D) Silicone granuloma

Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard



Question: Find the missing value in this math puzzle.
Solution:
$(5 - 4)^3 = 1$
$(7 - 3)^3 = 64$
$(8 - 2)^3 = 216$
Similarly, $(11 - 8)^3 = 27$.
So the missing value is 27.
Answer: 27

Our Solution: MMR-Life

MMR-Life Benchmark

- 2,646 questions
- 19K images
- Real-life scenarios
- Multi-image reasoning

Broad Reasoning Types

Abductive
Multi-hop Attribution, Interaction Attribution...

Deductive
Composition Deduction, Step Deduction...

Analogical
Similarity Inference, Relation Inference...

Inductive
Feature Induction, Disease Induction...

Causal
Multi-hop Prediction, Counterfactual...

Spatial
Position Estimation, Route Planning...

Temporal
Timeline Reconstruction, Sequence Prediction...

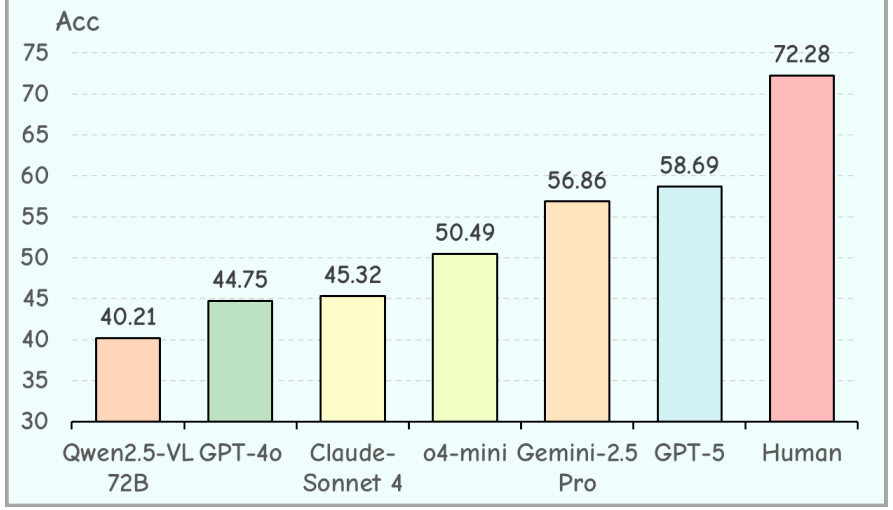
Reasoning in Real-life Senarios

Q: What is the correct temporal sequence of these images?

(A) 3-1-2-4 **(B)** 3-2-4-1 **(C)** 2-3-1-4
(D) 3-2-1-4 **(E)** 4-2-1-3

GPT-5: Look at the couple on the right side (man holding a white jacket and woman in a pink coat). In image 4 they are far back on the right aisle. In image 2 they are closer to the center. In image 1 they are even nearer to the palm. In image 3 they are at the front right, closest to the camera. The man in a pink T-shirt also moves forward from image 4 to image 2 and then is gone by images 1 and 3, consistent with this progression. Thus the correct chronological sequence is 4-2-1-3. #Answer: E

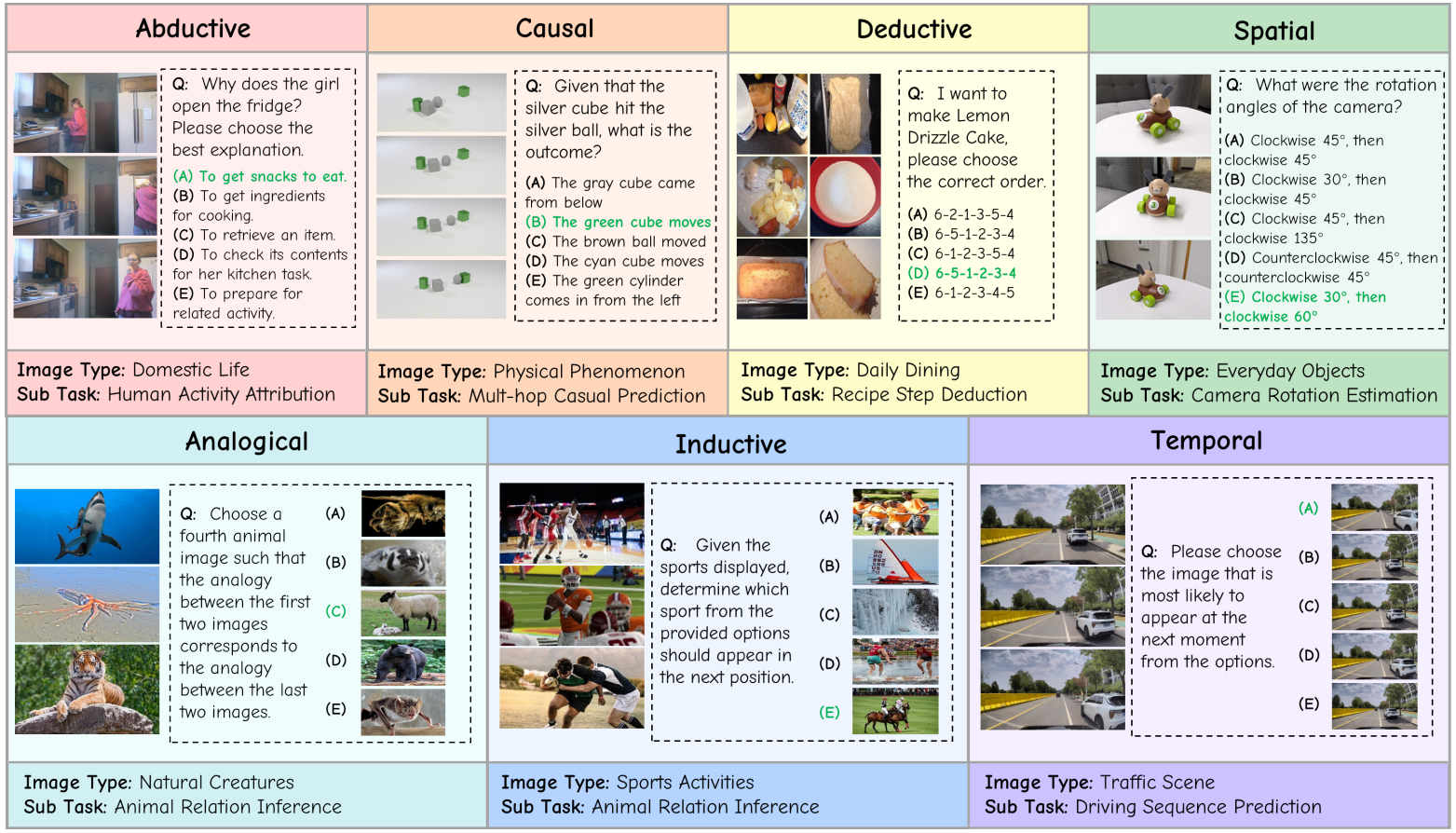
SOTA MLLM's Evaluation



Task Design

Broad Reasoning Types

- We cover 7 distinct reasoning types, which broadly encompass the reasoning abilities necessary for everyday situations.



Main Results

MMR-Life is Challenging

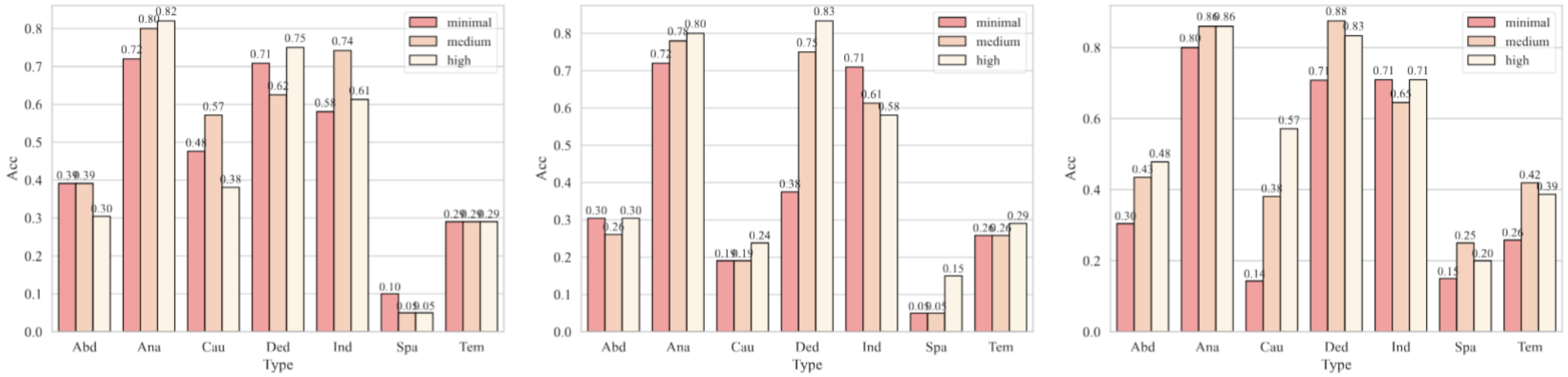
- GPT-5 only achieved an accuracy of 58.69%
- All models perform poorly in **spatial reasoning**, with the highest accuracy being only **25.10%**
- The open-source model's thinking mode does not show improved reasoning capabilities

Model	Abd	Ana	Cau	Ded	Ind	Spa	Tem	Avg
Human	79.76	57.65	75.00	70.59	63.41	79.76	79.76	72.28
Closed-source & Thinking								
GPT-5	53.75	**78.87**	41.06	**80.14**	**78.32**	17.25	**41.70**	58.69
Gemini-2.5-Pro	**54.40**	73.77	36.99	79.43	73.66	**25.10**	35.79	56.86
Gemini-2.5-Flash	46.25	75.18	34.22	71.63	73.66	23.92	30.81	53.10
o4-mini	41.37	73.59	27.38	71.28	68.07	19.22	32.66	50.49
GPT-5-mini	44.95	69.72	32.32	75.18	68.76	12.16	29.52	49.77
Claude-Sonnet-4	36.96	60.92	**44.11**	67.02	56.64	15.69	28.23	45.32
Closed-source & No Thinking								
GPT-4.1	44.30	71.30	22.43	67.38	70.16	13.73	27.31	48.15
Claude-3.7-Sonnet	33.55	66.55	35.36	59.93	59.67	20.78	26.01	45.09
GPT-4o	46.91	65.67	25.86	51.42	66.20	11.37	26.01	44.75
GPT-4.1-mini	32.90	61.62	30.80	52.13	65.27	16.47	30.63	44.10
Doubao-1.5-vision	37.13	53.70	31.18	59.57	54.31	12.16	23.06	39.98
Open-source & Thinking								
VL-Rethinker-72B	36.48	50.88	33.08	56.03	57.58	15.69	21.59	39.68
QVQ-72B-Preview	31.27	41.20	38.02	47.87	31.24	14.12	16.42	31.14
MM-Eureka-Qwen-32B	26.06	41.02	25.10	47.52	27.97	16.08	17.34	29.02
MiMo-VL-7B-RL	38.76	25.88	28.14	60.99	24.94	14.12	19.19	28.68
Keye-VL-1.5-8B	19.87	21.30	23.95	14.18	20.28	13.73	23.62	20.22
Skywork-R1V-38B	22.15	10.39	16.73	23.76	11.89	9.80	11.07	14.13
Open-source & No Thinking								
Qwen2.5-VL-72B	35.50	55.46	35.36	52.13	55.48	12.94	23.80	40.21
Gemma3-27B	35.18	57.92	36.88	31.21	60.61	12.94	18.27	38.32
Gemma3-12B	25.08	50.70	17.11	27.30	42.42	10.20	15.87	29.52
Qwen2.5-VL-32B	23.45	42.78	21.29	50.00	27.27	15.69	16.24	28.61
Qwen2.5-VL-7B	26.06	35.74	20.53	20.92	38.93	9.41	12.18	24.68
InternVL3.5-30B-A3B	45.60	19.19	33.46	36.52	14.45	12.16	14.39	23.09
InternVL3.5-8B	35.18	11.44	18.63	34.04	11.19	14.90	16.61	18.67

Thinking Budgets Analysis

Longer Thinking Is Not All You Need

- Longer thoughts do not lead to better performance for all reasoning types



(a) Gemini-2.5-Flash (b) GPT-5-mini (c) GPT-5

Figure 5: Performance comparison under different thinking budgets.

Reasoning Methods Analysis

Failure of Enhancement Methods in Larger Models

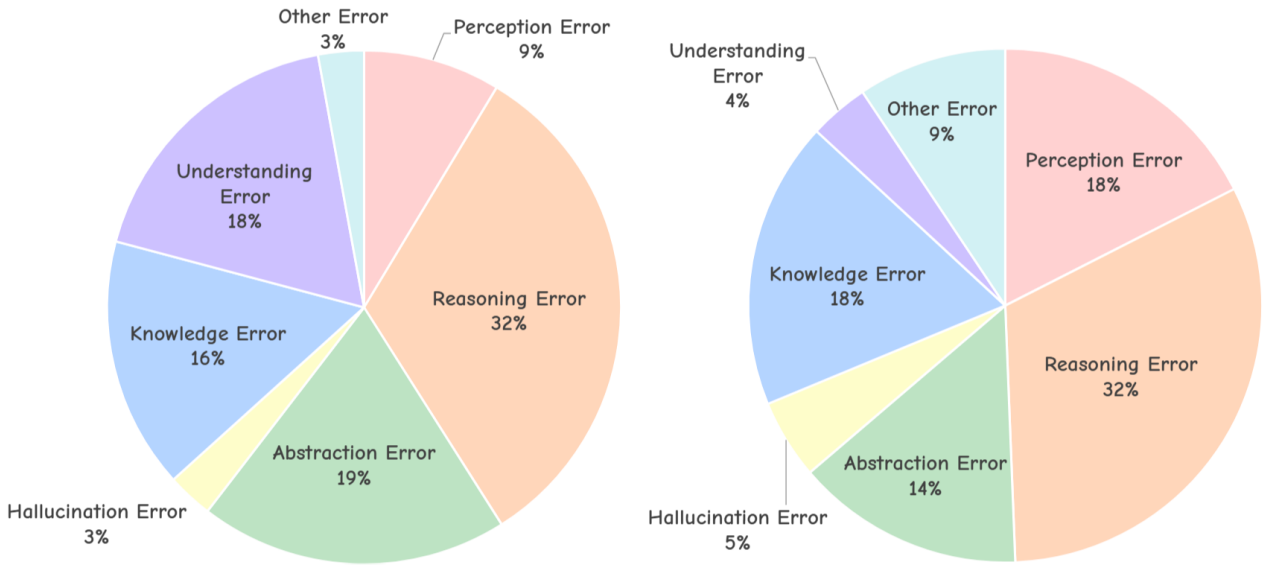
- As the size of models scales, the average performance difference between other methods and CoT consistently decreases

Model	Method	Abd	Ana	Cau	Ded	Ind	Spa	Tem	Avg (Δ)
Qwen2.5-VL-7B	CoT	26.06	35.74	20.53	20.92	38.93	9.41	12.18	24.68
	SC@8	28.01	39.44	**23.57**	25.18	45.45	10.98	**13.10**	27.85 (+3.17)
	BoN@8	27.64	**44.72**	22.81	25.53	**48.02**	13.33	**13.10**	**29.54** (+4.86)
	GRPO	**30.62**	40.49	21.29	**28.72**	43.59	**13.73**	11.81	28.23 (+3.55)
Qwen2.5-VL-32B	CoT	23.45	42.78	21.29	50.00	27.27	15.69	16.24	28.61
	SC@8	**26.06**	**45.42**	23.95	51.77	28.67	**16.47**	17.90	30.57 (+1.96)
	BoN@8	25.78	44.89	19.39	**55.32**	30.54	**16.47**	**19.56**	**30.97** (+2.36)
	GRPO	22.98	42.96	**28.14**	49.65	**30.77**	14.90	19.19	30.29 (+1.68)
Qwen2.5-VL-72B	CoT	35.50	55.46	**35.36**	52.13	55.48	12.94	23.80	40.21
	SC@8	35.18	**56.16**	**35.36**	52.13	54.78	12.94	24.35	**40.33** (+0.12)
	BoN@8	34.20	53.35	32.70	51.77	56.88	13.73	**24.72**	39.80 (-0.41)
	GRPO	**36.48**	50.88	33.08	**56.03**	**57.58**	**15.69**	21.59	39.68 (-0.53)

Error Analysis

Error Distribution for SOTA Models

- Reasoning errors dominate at 32%



(a) GPT-5 (b) Geimini-2.5-Pro

Figure 8: Error distribution over 140 errors for each model on MMR-Life.

Conclusion

Takaways

- We propose MMR-Life, the first comprehensive benchmark for evaluating multimodal multi-image reasoning in real-life scenarios across seven reasoning

- Through an extensive evaluation of 37 state-of-the-art MLLMs on MMR-Life, we find that existing models struggle considerably in real-life reasoning, especially in causal, spatial, and temporal tasks

- Based on MMR-Life, we conduct an in-depth analysis of current MLLM reasoning paradigms.

Thanks