



MM-OPERA: Benchmarking Open-ended Association Reasoning for Large Vision-Language Models

Zimeng Huang^{1,2}, Jinxin Ke¹, Xiaoxuan Fan³, Yufeng Yang¹, Yang Liu¹, Liu Zhonghan¹, Zedi Wang¹, Junteng Dai¹, Haoyi Jiang¹, Yuyu Zhou³, Keze Wang¹, Ziliang Chen^{2*}

¹Sun Yat-sen University ²Peng Cheng Laboratory ³Jinan University

NeurIPS 2025 Datasets and Benchmarks Track

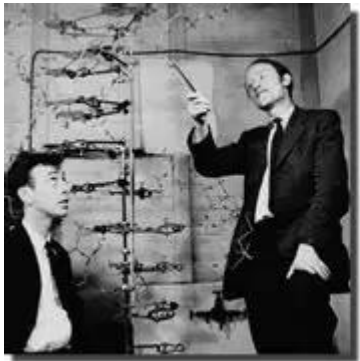
The Problem: A Gap in Evaluation

❑ Current LVLM benchmarks excel at:

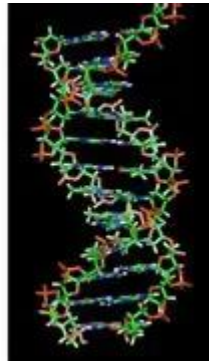
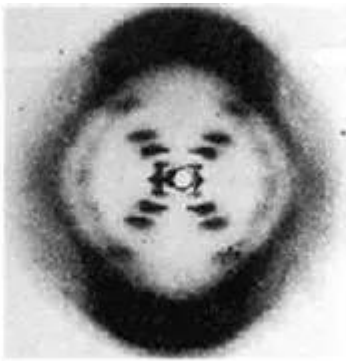
- Recognition, Comprehension, Instruction Following.
- Often use **closed-ended formats** (e.g., multiple choice).

❑ But they miss a fundamental human skill:

- **Association:** The ability to form creative and meaningful links between distant concepts, which:
 - Enables creative thinking;
 - Underpins the integration of fragmented information into coherent knowledge;
 - Supports critical cognitive processes such as memory, perception, and rule discovery.



Scientific Discovery



Creative Ideation



Creative Problem-Solving

Our Approach: MM-OPERA

□ MM-OPERA (Multi-Modal OPen-Ended Reasoning-guided Association)

- A benchmark with 11,000+ instances designed to test **associative intelligence**.
- **Key Principles:**
 1. **Open-Ended:** Models must generate free-form responses, not pick from a list.
 2. **Psychometrically Grounded:** Inspired by tests of human creativity, e.g. Remote Associates Test (RAT).
 3. **Reasoning-Focused:** Evaluate the path of reasoning, not just the final answer.

Table 1: Comparison between The Labyrinth of Links and MM-OPERA.

Dimension	The Labyrinth of Links	MM-OPERA (Ours)
Task Format	Multi-choice, closed-ended	Free-form, Open-ended
Association Tasks	Basic Steps: Single / Synchronous / Asynchronous	More Complex: Remote-Item Association / In-Context Association
Association Scope	Adjectives and Verb <i>limited semantic concepts</i>	3 relationship types, 13 ability dimensions; <i>broad cultural, linguistic and thematic contexts</i>
Evaluation Metrics	Correctness-focused: Max / Mean Step, Success Ratio	Multi-dimensional assessment: Score Rate, High Score Rate, Δ HR, Reasoning Score, Reasonableness, Distinctiveness, Knowledgeability
Evaluation Flexibility	Option-based, limited generative capacity	Fully generative, <i>supports diverse reasoning paths and rationales</i>


Task 1: Remote-Item Association (RIA)


- Challenges LVLMs to discover meaningful links between seemingly unrelated elements across text, images, or mixed modalities.
- Encourages cross-domain reasoning and rewards both logical coherence and creative insight, as multiple valid associative explanations may exist.

Remote-Item Association Task (RIA)

Instruction: Describe each image briefly. Analyze and explore the relation between the two images, identifying any possible connections, themes, or shared elements. + +


Query Images:







Reference Answer

Image 1: An armadillo **Image 2:** Kevlar fabric

 **Relation:** Protection

 **Explanation:** Armadillos possess natural armor for protection, while Kevlar is used in bulletproof vests for the same purpose. The common concept between them is protection.

 **Association Reasoning Path:**
Possess(Armadillo, NaturalArmor) and Purpose(NaturalArmor, Protection)
UsedIn(Kevlar, BulletproofVest) and Purpose(BulletproofVest, Protection)
Thus, Armadillo → NaturalArmor → Protection and
Kevlar → BulletproofVest → Protection

Task 2: In-Context Association (ICA)


- Extends RIA to in-context learning, thus evaluating a model's ability to recognize, abstract, and extend associative patterns within a creative framework.
- Encourages cross-domain reasoning and rewards both logical coherence and creative insight, as multiple valid associative explanations may exist.


In-Context Association Task (ICA)

Instruction:


1. Briefly describe **Image 1.1**, **Image 1.2**, and **Image 2.1** based on their visual information.
2. Analyze the relationship between **Image 1.1** and **Image 1.2**, identifying any possible connections, themes, or shared elements that link **Image 1.1** to **Image 1.2**.
3. Design **Image 2.2** so that its relationship with **Image 2.1** mirrors that between **Image 1.1** and **Image 1.2**. Use insights from the first pair to guide your design. + + +

Context Pair






Query Pair



(Imagined Target)



Reference Answer

Image 1:
A bald eagle
Image 2:
Basketball game

Image 3:
A lion
Image 4:
Football game

Relation:
Animals on the National Emblem and Sport's Origin

Explanation:

1. The bald eagle is the symbol on the national emblem of the United States, where the basketball originated.
2. The lion is the symbol on the national emblem of England, where the football originated.

Association Reasoning Path:

1. NationalEmblem(BaldEagle, US) and Origin(US, Basketball)
Thus, BaldEagle → US → Basketball
2. NationalEmblem(Lion, England) and Origin(England, Football)
Thus, Lion → England → Football

A Benchmark with Depth and Breadth

✓ Hierarchical Ability Taxonomy (Depth)

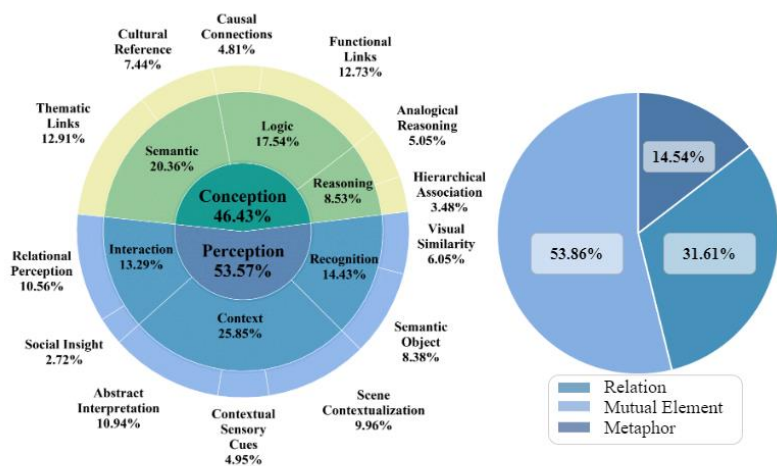
- A multi-level framework that mirrors human cognition:

✓ Structured Reasoning Paths (Rigor)

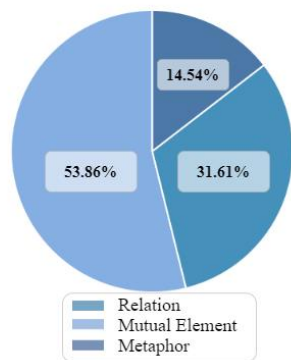
- We evaluate the entire reasoning process, not just the final answer.

✓ Rich Global Diversity (Breadth)

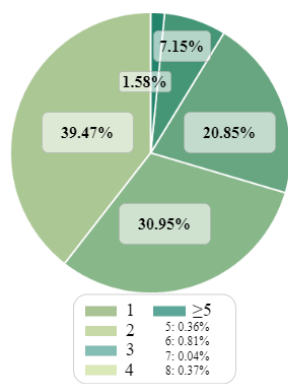
- The dataset is designed to be globally and conceptually comprehensive.



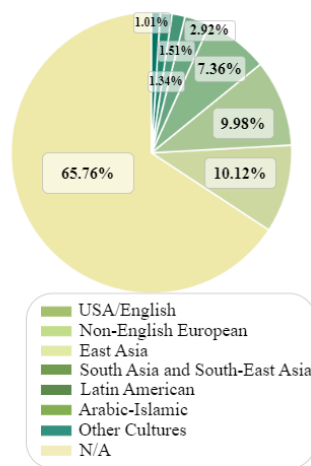
(a) Ability Dimensions



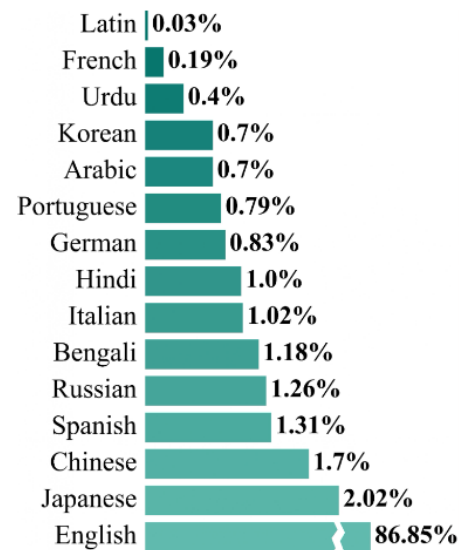
(b) Relationship Types



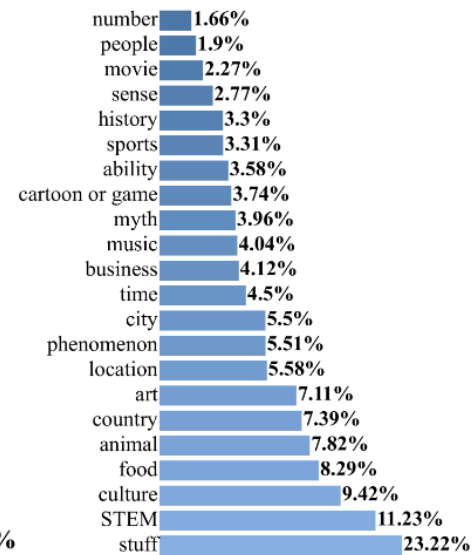
(c) The Number of Hops



(d) Cultures



(e) Languages



(f) Topic Domains

Evaluation: LLM-as-a-Judge Strategies for MM-OPERA

□ How to reliably score open-ended, creative answers?

- Holistic Outcome Scoring:** Assess the overall quality of the final response. A judge uses a cascading rubric (0-4 points) to score the answer's insight, coherence, and accuracy.
 - **Score Rate (SR):** The average holistic score across all responses, reflecting general performance.
 - **High Score Rate (HR-4):** The percentage of "perfect" answers that match the reference's intellectual rigor.
 - **Δ HR (HR-3 minus HR-4):** A proxy for divergent thinking—generating valid but non-optimal associations.
- Fine-Grained Process-Reward Scoring (PR-Judge):** Dissect the how and why behind the response. A specialized judge deconstructs the reasoning into a *structured path* and scores each logical hop.
 - **Stepwise Indicators:**
 - **Reasonableness (R_t):** Is the logical step plausible and coherent?
 - **Distinctiveness (D_t):** Does the step use specific concepts, avoiding vagueness?
 - **Knowledgeability (K_t):** Is the step grounded in correct world knowledge?
 - The quality of each step (s_t) is calculated by balancing internal coherence against external knowledge:
$$s_t = \alpha R_t D_t + (1 - \alpha) K_t$$
 - The overall Reasoning Score for the entire path then rewards efficiency by applying a cognitive decay factor (δ) to later steps, penalizing overly long or convoluted reasoning:

$$S_r = \sum_{t=1}^n s_t \delta^t$$

Key Finding 1: A Stark Gap Between LVLMs and Humans

- **Even state-of-the-art LVLMs lag behind humans in Association Reasoning.**
- On the Remote-Item Association task, the best model o4-mini achieves an HR-4 of 19.86% compared to humans’ 22.84%.
- On the In-Context Association task, humans achieve a 31.65% high score rate (HR-4). The best model, Gemini-2.5-Pro, reaches only 12.85%.
- The fact that human performance is far from perfect is consistent with decades of psychometric research like [1] (the average human performance on the Remote Associates Test was 34.2% and it was rather low.)

Model	Remote-Item Association Task				In-Context Association Task			
	SR(%)	HR-4(%)	HR-3(%)	\triangle HR(%)	SR(%)	HR-4(%)	HR-3(%)	\triangle HR(%)
Proprietary LVLMs								
Claude-3.5-Sonnet	49.38	9.26	25.17	15.91	49.35	3.97	23.27	19.3
Gemini-1.5-Flash	55.86	7.88	22.91	15.03	51.05	1.38	14.51	13.13
Gemini-1.5-Pro	45.34	8.95	20.97	12.02	42.16	2.45	11.05	8.60
Qwen-VL-Max	44.16	6.32	20.43	14.11	49.32	4.08	25.07	20.99
Qwen-VL-Plus	42.56	4.03	17.82	13.79	44.79	1.24	16.57	15.33
Gemini-2.0-Flash-Thinking-Exp	59.11	17.73	36.60	18.87	61.42	9.74	<u>37.88</u>	<u>28.14</u>
Gemini-2.5-Pro-Preview	<u>60.05</u>	23.89	41.75	17.86	63.09	12.85	41.15	28.30
o4-mini	60.33	<u>19.86</u>	<u>37.89</u>	<u>18.03</u>	<u>61.55</u>	<u>10.24</u>	36.60	26.36
GPT-4o	59.72	10.89	28.83	17.94	58.26	6.27	29.62	23.35
OpenSource LVLMs								
GLM-4V	26.92	0.49	4.73	4.24	43.63	0.20	3.67	3.47
InternVL-Chat-V1-2	36.41	3.52	16.02	12.5	34.30	0.62	9.59	8.97
InternLM-XComposer2.5-7B	<u>50.21</u>	2.21	14.39	12.18	44.87	1.41	<u>18.18</u>	<u>16.77</u>
VILA1.5	46.72	2.45	15.38	12.93	44.46	1.27	14.93	13.66
Yi-VL-34B	45.25	4.97	<u>19.63</u>	<u>14.66</u>	54.39	<u>1.30</u>	19.53	18.23
Qwen2.5-VL-7B-Instruct	52.28	5.35	20.36	15.00	<u>53.50</u>	1.08	16.62	15.54
Kimi-VL-A3B-Instruct	48.41	<u>5.14</u>	16.43	11.30	48.96	0.94	14.17	13.22
Human*	61.88	22.84	48.97	26.13	68.69	31.65	61.47	29.82

Table 2: Performance of models and human on the RIA and ICA tasks judged by gpt-4o-2024-08-06, with metrics including the holistic score rate (SR), high score rate (HR-4 , HR-3, and \triangle HR) derived from regular LLM-as-a-Judge. *The human baseline is based on the sampled data items.

[1] Pamela I Ansburg and Katherine Hill. Creative and analytic thinkers differ in their use of attentional resources. Personality and Individual Differences, 34(7):1141–1152, 2003.

Key Finding 2: The "Plausible but Shallow" Trap

- **High Reasonableness:**
Models are good at generating plausible connections.
- **Low Distinctiveness & Knowledgeability:** They struggle to make connections that are both specific and knowledge-grounded.
- This leads to responses that sound reasonable but lack genuine insight.

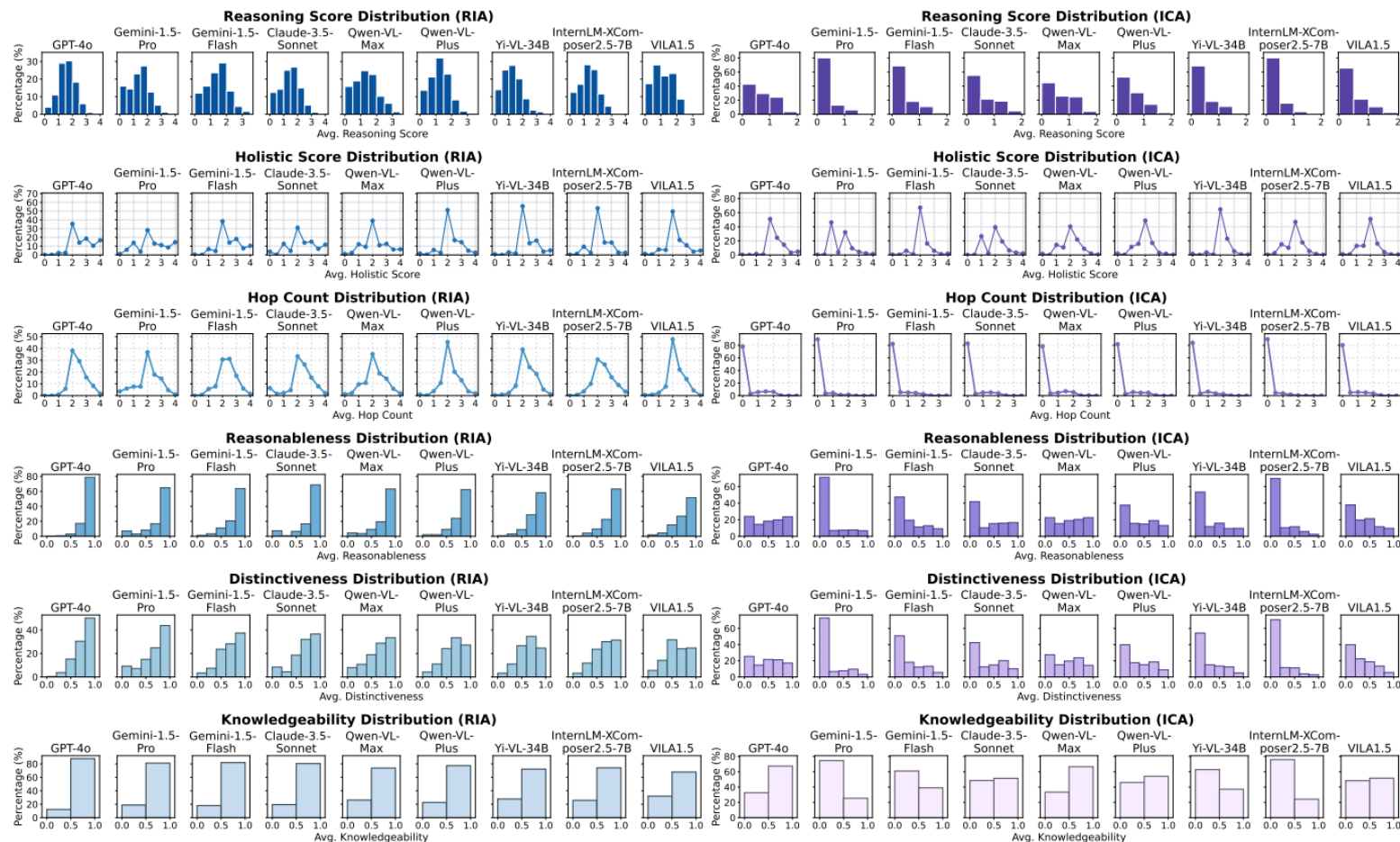


Figure 3: Fine-grained reasoning capability analysis of nine multimodal language models on RIA (left) and ICA tasks (right). From top to bottom: reasoning score distribution, holistic score distribution, reasoning path hop count distribution, Reasonableness distribution, Distinctiveness distribution, and Knowledgeability distribution. Each task includes 500 sampled questions, with results averaging evaluations from both GPT-4o and Deepseek-V3 judges.

Key Finding 3: The Creativity & Abstraction Deficit

- **The Creativity Gap: Unbalanced Divergent Thinking**
 - Δ HR measures the ability to generate ideas that are reasonable but not the single "perfect" answer. It's a proxy for creative, divergent thinking.
 - Humans exhibit a superior balance of creativity and accuracy, achieving a higher Δ HR (~26-30%).
 - Models can be creative, but they struggle to generate a wide range of high-quality, non-obvious associations like humans do.
- **The Abstraction Gap: A Failure in Meta-Reasoning**
 - RIA: Direct Association; ICA: Abstract a pattern & Transfer it.
 - Most models find the meta-reasoning in ICA significantly harder than the direct association in RIA.
 - This points to a core weakness in abstracting rules and transferring them to new contexts—a critical component of general intelligence.

Model	Remote-Item Association Task				In-Context Association Task			
	SR(%)	HR-4(%)	HR-3(%)	Δ HR(%)	SR(%)	HR-4(%)	HR-3(%)	Δ HR(%)
Proprietary LVLMS								
Claude-3.5-Sonnet	49.38	9.26	25.17	15.91	49.35	3.97	23.27	19.3
Gemini-1.5-Flash	55.86	7.88	22.91	15.03	51.05	1.38	14.51	13.13
Gemini-1.5-Pro	45.34	8.95	20.97	12.02	42.16	2.45	11.05	8.60
Qwen-VL-Max	44.16	6.32	20.43	14.11	49.32	4.08	25.07	20.99
Qwen-VL-Plus	42.56	4.03	17.82	13.79	44.79	1.24	16.57	15.33
Gemini-2.0-Flash-Thinking-Exp	59.11	17.73	36.60	18.87	61.42	9.74	<u>37.88</u>	<u>28.14</u>
Gemini-2.5-Pro-Preview	<u>60.05</u>	23.89	41.75	17.86	63.09	12.85	41.15	28.30
o4-mini	60.33	<u>19.86</u>	<u>37.89</u>	<u>18.03</u>	<u>61.55</u>	<u>10.24</u>	36.60	26.36
GPT-4o	59.72	10.89	28.83	17.94	58.26	6.27	29.62	23.35
OpenSource LVLMS								
GLM-4V	26.92	0.49	4.73	4.24	43.63	0.20	3.67	3.47
InternVL-Chat-V1-2	36.41	3.52	16.02	12.5	34.30	0.62	9.59	8.97
InternLM-XComposer2.5-7B	<u>50.21</u>	2.21	14.39	12.18	44.87	1.41	<u>18.18</u>	<u>16.77</u>
VILA1.5	46.72	2.45	15.38	12.93	44.46	1.27	14.93	13.66
Yi-VL-34B	45.25	4.97	<u>19.63</u>	<u>14.66</u>	54.39	<u>1.30</u>	19.53	18.23
Qwen2.5-VL-7B-Instruct	52.28	5.35	20.36	15.00	<u>53.50</u>	1.08	16.62	15.54
Kimi-VL-A3B-Instruct	48.41	<u>5.14</u>	16.43	11.30	48.96	0.94	14.17	13.22
Human*	61.88	22.84	48.97	26.13	68.69	31.65	61.47	29.82

Table 2: Performance of models and human on the RIA and ICA tasks judged by gpt-4o-2024-08-06, with metrics including the holistic score rate (SR), high score rate (HR-4 , HR-3, and Δ HR) derived from regular LLM-as-a-Judge. *The human baseline is based on the sampled data items.

Diagnosing the Failures: Four Common Pitfalls

<p>1. Perceptual Misalignment (45%) The model fails to <i>see</i> or correctly interpret critical visual details (e.g., a hidden symbol).</p>	<p>2. Knowledge Retrieval Gap (48%) The model knows the information but fails to activate it in the right context, especially for cross-cultural or domain-specific knowledge.</p>
<p>3. Overgeneralization (53%) When uncertain, the model defaults to vague, "safe" connections (e.g., "creativity", "art").</p>	<p>4. Limited Insight / Excessive Caution (23%) The model identifies the objects but refuses to make a conceptual leap, declaring them "unrelated" and avoiding any risk of being wrong.</p>

* Due to the inherent complexity of the tasks, a single response may exhibit multiple limitations, resulting in a cumulative contribution of factors exceeding 100%.

Failure Case Study: Why Do Models Perform Poorly?

Case 1 (Remote-Item Association)



What's the relation between the images?

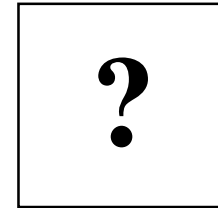
- **Reference Answer:** Hidden symbols.
- **GPT-4o's Answer:** Creativity.
- **Diagnosis:** A Perceptual Failure (missing the QR code hidden within the castle elements) cascaded into a vague Overgeneralization ("creativity").

Case 2 (In-Context Association)

Pair 1:



Pair 2:



What's the 4th image and what's the relation between the images in a pair?

- **Reference Answer:** A modern speaker; Evolution of Equipment.
- **Gemini-1.5-pro's Answer:** Vinyl records in a case; Container and its contents.
- **Diagnosis:** A Knowledge Retrieval Failure led to a superficial Pattern Matching error.

Conclusion

Our contributions are threefold:

- 1. MM-OPERA:** We introduce a benchmark of 10,000+ instances for evaluating LVLMs' association reasoning, centered on Remote-Item Association (RIA) and In-Context Association (ICA) tasks inspired by classic psychometric studies. It spans 13 analytical dimensions to enable comprehensive assessment.
- 2. LLM-as-a-Judge Strategies:** To support open-ended evaluation, we design tailored LLMas-a-Judge methods that assess both response quality and reasoning processes, enabling fine-grained and reliable scoring.
- 3. Profound Findings:** Our results demonstrate that current models, while powerful, have a distinct gap in robust conceptual reasoning and creativity compared to humans. By focusing on open-ended association, we provide a crucial tool to guide the development of the next generation of more human-like, creative, and truly general-purpose AI.

Thank you!