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

Zimeng Huang[1,2], **Jinxin Ke**[1], **Xiaoxuan Fan**[3], **Yufeng Yang**[1], **Yang Liu**[1], **Liu Zhonghan**[1],
Zedi Wang[1], **Junteng Dai**[1], **Haoyi Jiang**[1], **Yuyu Zhou**[3], **Keze Wang**[1], **Ziliang Chen**[2]*
[1]Sun Yat-sen University
[2]Peng Cheng Laboratory
[3]Jinan University
`{huangzm29, kejx, liuzhh268, wangzd6, daijt3, jianghy55}@mail2.sysu.edu.cn,`
`{yangyf226, liuy856, wangkz}@mail.sysu.edu.cn, fanxx@stu2022.jnu.edu.cn,`
`zyy@jnu.edu.cn, c.ziliang@yahoo.com`

Abstract

Large Vision-Language Models (LVLMs) have exhibited remarkable progress. However, deficiencies remain compared to human intelligence, such as hallucination and shallow pattern matching. In this work, we aim to evaluate a fundamental yet underexplored intelligence: association, a cornerstone of human cognition for creative thinking and knowledge integration. Current benchmarks, often limited to closed-ended tasks, fail to capture the complexity of **open-ended association reasoning** vital for real-world applications. To address this, we present MM-OPERA, a systematic benchmark with 11,497 instances across two open-ended tasks: Remote-Item Association (RIA) and In-Context Association (ICA), aligning association intelligence evaluation with human psychometric principles. It challenges LVLMs to resemble the spirit of divergent thinking and convergent associative reasoning through free-form responses and explicit reasoning paths. We deploy tailored LLM-as-a-Judge strategies to evaluate open-ended outputs, applying process-reward-informed judgment to dissect reasoning with precision. Extensive empirical studies on state-of-the-art LVLMs, including sensitivity analysis of task instances, validity analysis of LLM-as-a-Judge strategies, and diversity analysis across abilities, domains, languages, cultures, etc., provide a comprehensive and nuanced understanding of the limitations of current LVLMs in associative reasoning, paving the way for more human-like and general-purpose AI. The dataset and code are available at https://github.com/MM-OPERA-Bench/MM-OPERA.

1 Introduction

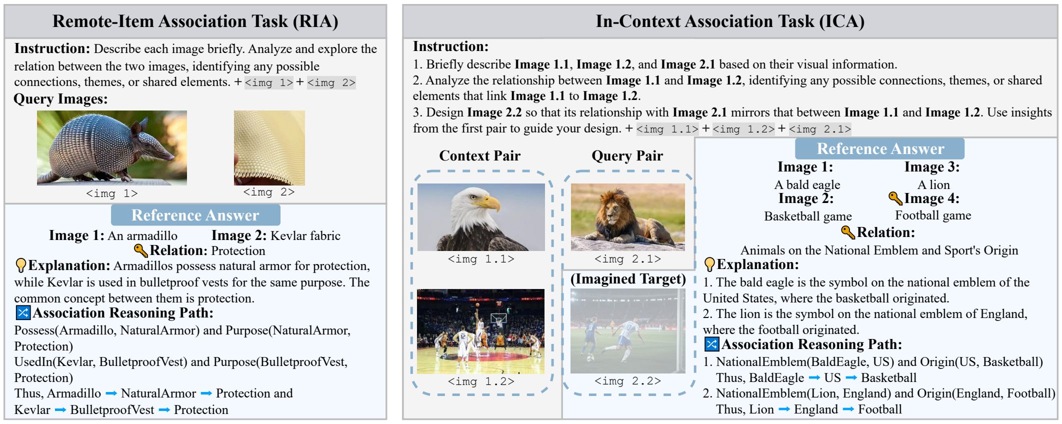
Recent advancements in Large Vision-Language Models (LVLMs) have significantly improved their ability to handle multi-modal inputs and address diverse tasks. Systems such as GPT-4 [69], Gemini models [79], and LLaVA [55] exhibit remarkable proficiency in visual understanding, language generation, and multi-step reasoning. These capabilities are driving transformative applications across fields such as education, design, scientific discovery, embodied intelligence, and so on [39, 19, 30, 57].

Existing benchmarks for LVLMs [3, 62, 35, 95, 44, 43, 96, 18, 31, 59, 60] has facilitated systematic assessments of instruction-following and alignment tasks, focusing on recognition, comprehension, and reasoning. However, the evaluation of association intelligence in LVLMs remains underexplored.

---

[*] Correspondence to: Ziliang Chen

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.



Instruction: Describe each image briefly. Analyze and explore the relation between the two images, identifying any possible connections, themes, or shared elements. + `<img 1>` + `<img 2>`

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

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. + `<img 1.1>` + `<img 1.2>` + `<img 2.1>`

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

Figure 1: An overview of MM-OPERA. The RIA task challenges models to discover meaningful connections between unrelated elements, while the ICA task requires transferring relationship patterns from a context pair to a query item to generate an appropriate target. The reference answer represents just one possible valid response. The association reasoning paths are used to evaluate the coherence and depth of the step-by-step reasoning process.

Association, a cornerstone of human cognition, enables creative thinking [64], underpins the integration of fragmented information into coherent knowledge and supports critical cognitive processes such as memory, perception, and rule discovery [5]. We argue that LVLMs need to develop this core capability to move beyond shallow pattern matching toward true knowledge synthesis and reasoning. It is a prerequisite for many real-world applications such as scientific discovery, creative ideation and design, personalized education, innovative problem-solving and robot planning.

Current efforts, such as the Labyrinth of Links [46] have begun to formalize association as an evaluation target, using closed-ended tasks with predefined options to probe associative memory. While this approach offers valuable insights, it falls short of capturing the full scope of association reasoning required for real-world AI applications. We argue that *open-ended association reasoning* is essential for two key reasons: (1) Closed-ended tasks with fixed options may introduce bias, subtly guiding the model's associative behavior and masking its true capacity for independent reasoning; (2) The fixed-answer format struggles to evaluate complex, long-form association reasoning, limiting the ability to challenge models on intricate, multi-step relational inference. These limitations motivate our development of a new benchmark that prioritizes **open-endedness** to rigorously assess and ultimately enhance LVLMs' association reasoning capabilities.

In cognitive science, association emerges from the interplay of *convergent and divergent thinking*: the former identifying meaningful connections and selecting optimal solutions; the latter generating multiple unique ideas [83, 13, 63]. The Remote Associates Test (RAT) [64, 24, 17, 87, 2] exemplifies this by requiring individuals to uncover links between distant concepts, a process vital for adaptive problem-solving. To mirror this in LVLMs and address the shortcomings of prior work, we propose **MM-OPERA** (**M**ulti-**M**odal **OP**en-**E**nded **R**easoning-guided **A**ssociation), a benchmark designed to evaluate association reasoning without predefined constraints. It assesses how models identify and express meaningful links across distant concepts (*i.e.* convergent thinking), expected to emerge through diverse reasoning paths (*i.e.* divergent thinking). Table 1 highlights how MM-OPERA diverges from The Labyrinth of Links by adopting open-ended tasks, more challenging reasoning scenarios, and a broader scope of evaluation, enabling a deeper probe into LVLMs' relational inference abilities.

MM-OPERA comprises 11,497 instances across two core tasks (Figure 1): *Remote-Item Association* (RIA), testing the ability to link distant concepts with structured reasoning, and *In-Context Association* (ICA), probing pattern recognition within in-context learning [29]. Spanning 13 associative dimensions and diverse cultural, linguistic, and thematic contexts, it offers a comprehensive evaluation framework. It prioritizes free-form responses, employing reference answers as heuristic quality benchmarks rather than rigid correctness criteria. To evaluate open-ended outputs, we design tailored LLM-as-a-Judge strategies with a cascading scoring rubric. Furthermore, by leveraging process-

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 *limited semantic concepts*	**3 relationship types, 13 ability dimensions**; *broad cultural, linguistic and thematic contexts*
Evaluation Metrics	Correctness-focused: Max / Mean Step, Success Ratio	**Multi-dimensional assessment**: Score Rate, High Score Rate, △HR, Reasoning Score, Reasonableness, Distinctiveness, Knowledgeability
Evaluation Flexibility	Option-based, limited generative capacity	**Fully generative**, *supports diverse reasoning paths and rationales*

reward principles to trace reasoning steps, our evaluation captures cognitive flow and knowledge integration, surpassing traditional outcome-focused metrics.

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 LLM-as-a-Judge methods that assess both response quality and reasoning processes, enabling fine-grained and reliable scoring.
3. **Profound Findings:** Our analysis reveals key limitations of current LVLMs and highlights the critical role of association reasoning in advancing real-world, general-purpose AI.

2 Related Work

Large Vision Language Models (LVLMs). Early studies [78, 98, 73] established the foundations of vision-language models. CoCa [94], Flamingo [1], and BLIP-2 [47], advanced performance with enhanced architectures and large-scale multimodal pretraining. InstructBLIP [25], MiniGPT-4 [103], and LLaVA [56], have refined multimodal instruction tuning and alignment strategies. Recent open-source LVLMs, *e.g.*, LLaVA-OneVision [45], mPLUG-Owl3 [91], and Qwen2-VL [7], have extended these capabilities to multi-image and video understanding. Proprietary models like GPT-4V [69], Gemini-Pro-V [79], and Qwen-VL-Max [8] have demonstrated state-of-the-art performance.

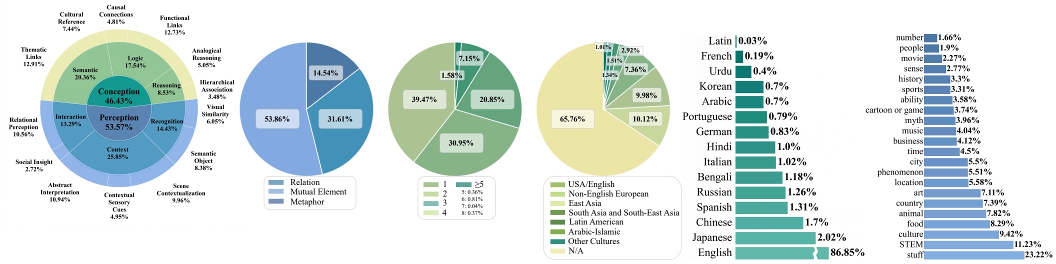
LVLM Benchmarks. The evaluation of LVLM has progressed from early benchmarks like VQA [3, 35] and OK-VQA [62] to broader assessments such as SEED-Bench [44], LAMM [92], LVLM-eHub [90], MMBench [59], MSCOCO [53], and MM-Vet [95], covering tasks like Optical Character Recognition (OCR) [58], adversarial robustness [100], and hallucination detection [23, 54, 48, 84]. Specialized benchmarks target various capabilities: MathVista [61], CLEVR [40], CVR [97], ReMI [42], Encyclopedic VQA [65], LogicVista [89], SPACE [74], BLINK [32], ZeroBench [75], MMMU [96], and Visual Riddles [15] each focus on different aspects of reasoning and perception. Li et al. [46] propose an adjective-verb association benchmark, but it is constrained to predefined categories, leaving open-ended associative reasoning largely unexplored.

Psychometric Test for AI Evaluation. Researchers have proposed psychometric frameworks to assess AI cognition [38], ranging from personality and theory-of-mind benchmarks [49], latent trait profiling [72], and reasoning evaluation via the Technology Acceptance Model (TAM) [50], to broader construct-oriented approaches emphasizing underlying cognitive mechanisms over task-level performance [86]. Adaptive testing further enhances efficiency by dynamically adjusting to model responses [104]. Association reasoning has also been modeled and involved in AI evaluation [76, 46]. Mednick's Theory of Creativity defines creativity as forming remote connections [64], underpinning associative creativity theories [71, 10] and the Remote Associates Test (RAT), adapted for semantic and visual associations [16, 68, 11, 67] through convergent thinking tasks. Divergent thinking [81, 14] is also assessed via tasks like the Alternate Uses Task (AUT) [36] and Divergent Association Task

(DAT) [66]. Studies on LLMs and LVLMs reveal mixed results: "leap-of-thought" tasks enhance divergent reasoning [102], GPT models show varied creativity and even surpass humans [41, 22].

3 MM-OPERA: Dataset

In this section, we illustrate the task design and the corresponding dataset of MM-OPERA. Section 3.1 elaborates association tasks and Section 3.2 presents the dataset statistics. The data curation is detailed in Appendix A.4.



(a) Ability Dimensions (b) Relationship Types (c) The Number of Hops (d) Cultures (e) Languages (f) Topic Domains

Figure 2: Statistics of MM-OPERA. (a) Hierarchical ability taxonomy consists of 3 levels, refining perceptual and conceptual associations. We report each ability's frequency as a percentage of total label occurrences to better represent the dataset's distribution. (b) Three relationship types capturing diverse associative connections. (c) The number of hops in the association reasoning path, quantifying different associative reasoning complexity. (d) Different cultures, (e) 15 languages, and (f) 22 topic domains ensuring broad cultural, linguistic, and thematic diversity.

3.1 Association Tasks: Motivation and Definition

Associative ability is commonly assessed through the Remote Associates Test (RAT), which presents participants with three seemingly unrelated items and asks them to identify a fourth item that connects all three. While RAT offers psychological validity, it primarily emphasizes instinctive convergent thinking with a single-hop reasoning path across items. However, the human response time metric in RAT is difficult to replicate in machines. Moreover, RAT lacks the complexity needed to capture the divergent thinking process that underlies convergent thinking.

To remedy this, we re-develop the remote-item paradigm into two novel association tasks, both incorporating a chain of thought with multi-step reasoning structure across a pair of remote multimodal items. LVLM is required to generate an open-ended answer with the explanation, while the reference answer and its underlying reasoning chain is provided for newly invented LLM-as-a-Judge strategies (Section 4).

Remote-Item Association. The RIA task instance challenges LVLMs to discover meaningful links between seemingly unrelated elements across text, images, or mixed modalities. As shown in Figure 1 (left), when presented with query images of an armadillo and Kevlar fabric, an LVLM candidate is demanded to identify their shared protective function—moving beyond surface features to reveal conceptual bridges. This task encourages cross-domain reasoning and rewards both logical coherence and creative insight, as multiple valid associative explanations may exist.

In-Context Association. The ICA task instance extends RIA to in-context learning, thus evaluating a model's ability to recognize, abstract, and extend associative patterns within a creative framework. In Figure 1 (right), the model first identifies the connection between a bald eagle and basketball (America's national symbol and a sport originating there), then applies this pattern to generate the appropriate complement to a lion image (football, as England's national symbol relates to the sport's origin). This task tests the model's pattern-based reasoning, ability to abstract cross-domain associations and balance creative flexibility with logical consistency.

3.2 Dataset Statistics

MM-OPERA contains 11,497 task instances (8,021 in RIA and 3,476 in ICA) spanning diverse modalities, concepts, and reasoning complexities. Its comprehensive design supports thorough evaluation of LVLMs' associative capabilities across multiple dimensions, reflecting the multifaceted nature of human associative reasoning. Detailed statistics are presented in Figure 2.

Sample Distribution and Design. The RIA dataset includes Multiple-Image variants where identical concepts appear in different images, enabling controlled sensitivity testing of LVLMs' visual perception. Notably, over 25% of these instances exhibit unique concept pairs, ensuring breadth in conceptual coverage. The ICA dataset employs an circular evaluation strategy, where each set of four images generates four distinct questions, each requiring the model to reason about one image based on the relationships established by the other three.

Hierarchical Ability Taxonomy. Associative thinking operates on multiple cognitive levels, from perception to conception—both crucial for understanding complex environments [88, 77, 37, 12]. Perception handles immediate sensory inputs, while conception deals with abstract, knowledge-driven associations. These fundamental processes form our Level-1 (*L-1*) associative ability. We further refine it into six *L-2* and thirteen *L-3* dimensions, creating a hierarchical framework that mirrors human cognition and enables systematic evaluation of LVLMs' capabilities in processing both sensory input and abstract reasoning. Detailed definitions are in Appendix A.2.

Types of Relationship. To capture the nuanced ways entities or concepts connect, we identify three relationship categories: *Relation*, denoting general links between entities; *Mutual Element*, indicating shared traits; and *Metaphor*, connecting entities through abstract or symbolic meanings. This tripartite classification enhances the benchmark's ability to evaluate associative reasoning across both literal and abstract dimensions, reflecting the multifaceted nature of human associative thinking.

Association Reasoning Path. While natural language explanations offer valuable insights into associative reasoning, they often lack the structured clarity needed to systematically evaluate complex reasoning processes. To address this limitation, we introduce *Association Reasoning Paths*, a visual framework that represents the reasoning process as a directed path with arrows connecting concepts. Each *hop* in this path represents a discrete reasoning step, with the total number of hops directly reflecting the association's complexity. For instance, connecting an armadillo to Kevlar might require a four-hop path through intermediate concepts leading to their shared protective function (Figure 1). This structured representation enables reasoning-guided evaluation detailed in Section 4.2.

Diversity. Our dataset deliberately incorporates various cultures, 15 languages with their unique linguistic devices (idioms, puns, proverbs) as association links, and 22 topic domains. This diversity is proposeful: while LVLMs possess vast knowledge repositories from their training, true intelligence lies in the ability to activate these knowledge pathways—connecting observations to prior knowledge across cultural, linguistic, and domain boundaries—which is the basis of association.

4 LLM-as-a-Judge Strategies for MM-OPERA

While open-ended tasks eliminate potential hints that might influence models' association behaviors, they present significant evaluation challenges. Traditional methods including human evaluation, rule-based systems, and automatic metrics, often struggle with inconsistency and bias when assessing such unconstrained responses [4, 21]. To address these challenges, we present three complementary LLM-as-a-Judge strategies: Section 4.1 introduces our Regular Scoring framework, which serves as the foundation for Process-Reward Evaluation in Section 4.2. Evaluation prompts are available in Appendix E.

4.1 Regular LLM-as-a-Judge Scoring

Since the open-ended responses and references answers presented in text, we adopt LLMs as automatic judge engine. Unlike prior benchmarks that use per-sample criteria [33], we adopt unified scoring rubrics that evaluate the association quality of responses—prioritizing depth, coherence, and insight over mere correctness. With regards to open-ended responses with multiple valid potential answers, our regular judge engine assess the internal consistency and reasoning quality by the cascading scoring rubric:

- **4 points:** Accurate, logically consistent, and insightful, matching the reference answer's intellectual rigor.
- **3 points:** Shows reasonable understanding but lacks key insights or completeness.
- **2 points:** Somewhat relevant but lacks depth, is overly broad, or omits critical reasoning.
- **1 point:** Vague, uncertain, or incomplete, failing to provide meaningful reasoning.
- **0 points:** Contains factual errors or fabrications that undermine validity.

We refer to this scoring as the *Holistic Score* in the paper to distinguish it from the reasoning score introduced in Section 4.2. Based on the scoring rubic, we define the evaluation metrics: (1) *Score Rate (SR)*, the average score to all open-ended responses judged by the LLM to reflect the general performance. (2) *High Score Rate (HR)*, the proportion of responses with explanation that makes sense in terms of LLM's analysis. It specifically derives *HR-3*, the percentage of responses scoring not less than 3, and *HR-4*, the percentage of responses scoring 4 (consistent with the reference answer). (3) It is obvious that *HR-3* $\geq$ *HR-4*, and their difference $\triangle$*HR*=*HR-3*$-$*HR-4* implies the proportion of the "divergent thinking" results of LVLMs.

4.2 Process-Reward LLM-as-a-Judge Scoring

The regular scoring rule in Section 4.1 is outcome-based and fail to distinguish and analyze models that produce similar outcomes through divergent thinking with different reasoning paths. Drawing inspiration from process reward models [85, 51], which qualify each intermediate reasoning step based on its potential to reach the correct outcome, we propose a customized process-reward LLM-as-a-Judge method (PR-Judge) to access each association reasoning step towards the final outcome connections, offering insights of reasoning process that outcome-based metrics cannot capture.

1. **Path Construction**: The LLM judge reformats model responses into association paths P comprising sequential steps (or hops) $(s_1, s_2, ..., s_n)$.
2. **Stepwise Scoring Indicators**: Association reasoning step t is accessed from three persepecitves:
    - **Reasonableness** (R_t): Reasoning fluency, the cognitive fluidity and logical coherence of the associative transition, reflecting the plausibility that leads to the outcome.
    - **Distinctiveness** (D_t): The distinctiveness of concept boundaries. Lower value indicates the negative effect due to vague or overly general associative connections.
    - **Knowledgeability** (K_t): The level of detail and development of the idea relevant with domain knowledge manifested in the step.

    These stepwise indicators are inspired from Guilford's Alternate Uses [82] that reflects the divergent thinking behaviors of human. R_t and D_t are scalar values in $[0, 1]$ while K_t is binary in (0 or 1).
3. **Stepwise Association Quality and Path Scoring**: With regards to the indicators, the association quality per step s_t is calculated as:

    $$s_t = \alpha R_t D_t + (1 - \alpha) K_t, \tag{1}$$

    then overall *Reasoning Score* of each reasoning path is:

    $$S_r = \sum_{t=1}^{n} s_t \delta^t. \tag{2}$$

    Among them, α balances internal reasoning coherence $R_t D_t$ against knowledge K_t; δ serves as a cognitive decay factor resembling the spirit of self-supervised process reward model [85], inherently favoring efficient and precise reasoning paths.

This structured evaluation framework enables a comprehensive assessment of associative reasoning quality.

Model	Remote-Item Association Task				In-Context Association Task			
	SR(%)	HR-4(%)	HR-3(%)	△HR(%)	SR(%)	HR-4(%)	HR-3(%)	△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	37.88	28.14
Gemini-2.5-Pro-Preview	60.05	**23.89**	**41.75**	17.86	**63.09**	**12.85**	**41.15**	**28.30**
o4-mini	**60.33**	19.86	37.89	18.03	61.55	10.24	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	50.21	2.21	14.39	12.18	44.87	**1.41**	18.18	16.77
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	19.63	14.66	**54.39**	1.30	**19.53**	**18.23**
Qwen2.5-VL-7B-Instruct	**52.28**	**5.35**	**20.36**	**15.00**	53.50	1.08	16.62	15.54
Kimi-VL-A3B-Instruct	48.41	5.14	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 △HR) derived from regular LLM-as-a-Judge. *The human baseline is based on the sampled data items.

5 Experiments and Analysis

5.1 Settings

LVLM Baselines. We evaluated both proprietary and open-source VLMs under zero-shot conditions with default temperature. Proprietary models[2] include GPT-4 Omni [69], o4-mini [70], Gemini-1.5-Pro [79], Gemini-1.5-Flash [79], Gemini-2.5-Pro-Preview [27], Gemini-2.0-Flash-Thinking-Experimental [26], Claude-3.5-Sonnet [6], Qwen-VL-Max [8], Qwen-VL-Plus [8], while open-source models consist of GLM-4V [34], Yi-VL-34B [93], InternVL-Chat-V1-2 [20], VILA1.5 [52], InternLM-XComposer2.5-7B [99], Qwen2.5-VL-7B-Instruct [9] and Kimi-VL-A3B-Instruct [80]. Experiments for locally deployed models were conducted using 80 GB NVIDIA A800 GPUs.

Human Baseline. The study included 24 undergraduate and graduate students from diverse academic fields at a comprehensive university, selected for their cognitive skills appropriate for associative reasoning. We utilized 485 RIA and 436 ICA questions, grounded in widely accessible knowledge. Participants undertook the open-ended questions in a relaxed, non-evaluative atmosphere. Each addressed a subset of under 40 questions to ensure focus and prevent task-induced fatigue.

Judge Engine. We employ GPT-4o (gpt-4o-2024-08-06) and DeepSeek-V3 [28] as the mixed basic LLM-as-a-Judge engine for scoring. The former is excluded to evaluate its LVLM variant to ensure the fairness and prevent self-enhancement bias.

5.2 Outcome Evaluation of Association Reasoning

A comparison of different VLMs using the MM-OPERA is detailed in Table 2. Analyses across various dimensions are in Appendix B.1. Our key findings are:

LVLMs Far Below Humans in Association Reasoning. MM-OPERA reveals the formidable challenges of associative reasoning for current LVLMs. While latest models like o4-mini and latest Gemini models show improved performance, with SR approaching the human baseline, they still fall short in achieving high-quality associations. For instance, on the RIA task, o4-mini achieves an HR-4 of 19.86% compared to humans' 22.84%, and on the ICA task, Gemini-2.5-Pro-Preview reaches an HR-4 of 12.85% against humans' 31.65%, which demonstrates that sophisticated associative

---

[2] The model versions are: gpt-4o, o4-mini, gemini-1.5-pro-001, gemini-1.5-flash-001, gemini-2.5-pro-preview-05-06, gemini-2.0-flash-thinking-exp-01-21, claude-3-5-sonnet-20240620, qwen-vl-max-0809, qwen-vl-plus-0809.

reasoning remains at the cutting edge of LVLM capabilities. The fact that human performance is far from perfect is consistent with decades of psychometric research like [2] (the average human performance on the Remote Associates Test was 34.2% and it was rather low.) The performance gap between human and LVLMs is therefore an important scientific finding about the current state of AI. Case studies in Appendix C illuminate key limitations, such as **cross-domain knowledge retrieval deficiencies** and **perceptual misalignments**.

Creativity Gap in Divergent Thinking. The $\triangle$HR metric highlights divergent thinking, with most models scoring 12%–20%, showing their ability to generate reasonable yet non-optimal associations. Latest Gemini models lead among LVLMs (18.87% and 28.30% in two tasks), but humans outperform with both higher $\triangle$HR (26.13% and 29.82%) and HR-3 scores, demonstrating a superior balance of creativity and accuracy—an area where LVLMs remain limited.

ICA: Dual Challenge of Pattern Abstraction and Transfer. Most models perform better on RIA than ICA, highlighting the challenges of pattern-based associative reasoning. ICA requires not only connecting concepts but also abstracting and transferring these patterns to new contexts—a complex process demanding advanced meta-reasoning. Notably, some models such as latest Gemini models, Yi-VL-34B and GLM-4V outperform on ICA compared to RIA, suggesting that certain architectures excel in specific associative reasoning tasks. These distinctions may stem from more effective pattern extraction or transfer mechanisms, warranting further investigation.

Conservative Reasoning vs. Associative Flexibility. Analysis shows an inverse correlation between model constraints and associative abilities. Gemini-1.5-Flash (55.86% SR on RIA), optimized for speed, outperforms Gemini-1.5-Pro (45.34% SR), despite Pro's larger size and focus on detailed reasoning. Examination of 500 random RIA samples (Figure 3) shows Pro's conservative behavior to reason the high-rate association, prioritizing factuality and ethics, led to 1 point scores on nearly 20% of RIA questions due to conservative responses like "unrelated", versus Flash's $<$10%. Flash tended to offer superficial connections where Pro declined. Thus, factuality checks and ethical considerations, while improving reliability for complex tasks, can limit performance on creative association.

5.3 Process Evaluation of Association Reasoning

To deeply understand LVLMs' associative reasoning capabilities, we conducted fine-grained analysis using our Process-Reward LLM-as-a-Judge (PR-Judge) on 500 samples each from RIA and ICA datasets. We evaluated 9 models with $\alpha = 0.9$ and $\delta = 0.9$ employing both GPT-4o and Deepseek-V3 as the judges, averaging their results for final analysis. This dual-judge approach mitigates self-enhancement bias, as Deepseek-V3 provides an independent perspective with its distinct architecture and specialized mixture-of-experts training methodology. Though judges showed slight variance in scoring ranges (Deepseek-V3 trending higher), the self-enhancement bias of GPT-4o and the impact on comparative rankings remained minimal (Appendix B.2).

Process Evaluation of Association: Complexity Matters. GPT-4o demonstrates superior performance on both tasks, achieving the highest scores across all metrics (see Table 3). All models achieved average reasoning scores above 1.1 on RIA, but these scores dropped below 0.7 on ICA. Figure 3's Reasoning Score and Hop Count Distribution reveal that RIA responses exhibit richer reasoning structures, primarily centered at higher reasoning scores and 2-hop paths. In contrast, ICA tasks generate a substantial proportion of low scores and 0-hop responses, often reflecting insufficient logical structure or vague associative connections, thus highlighting ICA's greater difficulty and complexity.

Plausible Links vs. Knowledge-Grounded Distinctiveness. Figure 3's distributions of Reasonableness, Distinctiveness, and Knowledgeability reveal a critical limitation in LVLMs' associative reasoning: they struggle to move beyond plausible connections to achieve clear, knowledge-grounded understanding. While performing adequately on Reasonableness (50%–80% of RIA responses scoring above 75%), models significantly fall short in distinctiveness (less than half above 75%). Knowledgeability scores, though generally higher, still show a shortcoming in deep knowledge integration. This is reflected by the concentration of holistic scores at a mediocre "2" (Figure 3's Holistic Score Distribution), indicating superficial relevance and lack of depth. Thus, LVLMs can establish plausible connections, but lack the clear conceptualization and comprehensive knowledge integration required for truly sophisticated associative thinking.

	RIA		ICA	
Model	**Holistic SR(%)**	**Avg. Reasoning Score**	**Holistic SR(%)**	**Avg. Reasoning Score**
Claude-3.5-Sonnet	58.15	1.4148	49.28	0.5099
Gemini-1.5-Flash	<u>61.95</u>	<u>1.4193</u>	52.95	0.3746
Gemini-1.5-Pro	58.35	1.3805	41.38	0.2208
Qwen-VL-Max	54.45	1.3160	50.375	<u>0.6346</u>
Qwen-VL-Plus	56.20	1.2362	47.68	0.4901
GPT-4o	**67.78**	**1.6068**	**59.70**	**0.6396**
InternLM-XComposer2.5-7B	54.38	1.1384	47.95	0.2144
VILA1.5	55.73	1.1384	47.98	0.4191
Yi-VL-34B	58.43	1.2463	<u>53.10</u>	0.3567

Table 3: Holistic Score Rate (%) and average Reasoning Score of nine LVLMs on RIA and ICA tasks. Bold indicates best results, underlined indicates second-best results. Scores represent the average of evaluations by GPT-4o and Deepseek-V3 judges.

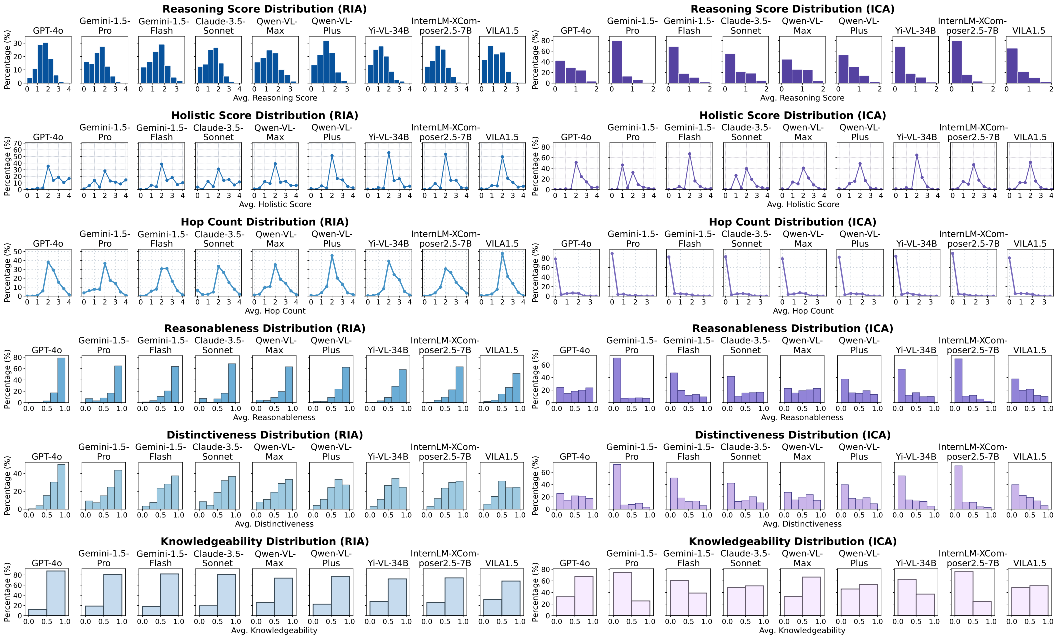


Figure 3: Fine-grained reasoning capability analysis of nine multimodal language models on RIA (left) and ICA tasks (right). From top to bottom: reasoning score distribution, holistic score distribution, reasoning path hop count distribution, Reasonableness distribution, Distinctiveness distribution,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

5.4 Sensitivity Analysis of Task Instances

We conducted three sensitivity tests to assess score consistency and robustness. In the Multi-Image Substitution Test, we grouped multiple-image variants with identical concept pairs in RIA and measure score variability. In the Text-Image Substitution and Order Sensitivity tests, we randomly sampled 400 RIA instances and evaluated GPT-4o, Gemini-1.5-Pro, and Gemini-1.5-Flash, using original image-image pairs as the baseline.

Multi-Image Substitution Test. Results (detailed in Appendix B.3.1) reveal significant variability in how LVLMs handle different visual representations of identical concepts. GPT-4o demonstrates remarkable visual robustness with minimal score fluctuation, while some models show substantial performance variations across concept-identical images. This indicates most current LVLMs remain sensitive to surface-level visual features rather than forming robust conceptual representations, highlighting a critical gap between contemporary architectures and true concept-level associative reasoning.

Text-Image Substitution Test. We evaluated cross-modal generalization by replacing images with text descriptions and comparing scores across conditions. Appendix B.3.2 suggests GPT-4o's reliance on nuanced visual cues that text descriptions cannot fully capture, while Gemini models demonstrate stronger text-equivalence in their reasoning, potentially processing visual information through language-like internal representations. These findings highlight how different architectural approaches influence cross-modal generalization in associative reasoning tasks. Additionally, the observations also showed that LVLMs struggle more with processing the raw visual input than with reasoning from a 'perfect' text description. It demonstrated how visually challenging our benchmark is.

Order Sensitivity Test. We examined the model's sensitivity to input order by reversing the image sequence. Appendix B.3.3 suggests that while GPT-4o processes image pairs in a more commutative manner, treating both ordering equally, Gemini models, particularly Gemini-1.5-Pro, appear to apply asymmetric reasoning processes that may prioritize the first image as context and the second as the target for association, highlighting architectural differences in how models approach bimodal associative reasoning.

5.5 LLM-as-a-Judge Strategy Validation for Reliable Evaluation

Full details of our LLM-as-a-Judge framework validation are in Appendix B.4.

Bias Analysis. We addressed **verbosity and position biases** [101]. Excluding short 1-point responses, the Pearson correlation between response length and scores was 0.376 for regular scoring and 0.291 for PR-Judge, indicating minimal verbosity bias. Permutation tests shuffling answer order showed mean score differences below 0.1 (regular) and 0.16 (PR-Judge), confirming negligible position bias.

Alignment with Human Judgment. We compared 300 sampled GPT-4o judgments with 8 human evaluators, yielding an average score difference of 0.077, with 78.33% perfect matches and no discrepancies exceeding 1 point. For PR-Judge, we evaluated 200 reasoning paths scored by 8 domain-expert judges on Reasonableness, Distinctiveness, and Knowledgeability. The average score difference was 0.1961, with 81% differing by less than 0.20 and none exceeding 0.60. Correlations were strong: $r = 0.72$ for Reasonableness, $r = 0.68$ for Distinctiveness, and 83.5% accuracy for Knowledgeability (Cohen's Kappa = 0.65), demonstrating robust alignment with human judgment.

Effectiveness of Process-Reward LLM-as-a-Judge. We compared it with outcome-based regular scoring on 100 paths. While outcome-based methods gave similar scores (e.g., 4) to correct answers, PR-Judge distinguished reasoning quality (e.g., 1.3 vs. 1.8), offering a more nuanced evaluation.

6 Conclusion

MM-OPERA introduces a novel framework for evaluating LVLM's association reasoning through open-ended tasks without predefined constraints. Drawing from cognitive psychology, it addresses traditional limitations while capturing diverse aspects of associative thinking. Results reveal that top LVLMs fail to achieve human performance, exposing task-specific patterns and a distinctiveness gap in robust conceptual reasoning. These insights underscore current limitations and provide direction for advancing human-like reasoning models.

Acknowledgments and Disclosure of Funding

We would like to express our sincere gratitude to the numerous volunteers for their invaluable contributions to the development of the MM-OPERA benchmark. Their diligent efforts in data collection and quality control were crucial to the success of this work. We would especially like to thank Xiaotong Fu, Dawei Zheng, Hanbo Zhang, Chao Tan, Jinfeng Liu, Jiale Wang, Qinglin Zeng. We are also grateful to Zechuan Chen, Quanlong Guan, Xinghe Cheng, and Jialong Xue for their significant help in organizing the human baseline evaluation.

The research was supported in part by Guangdong S&T Programme (Grant No. 2024B0101010003); the Open research fund of Pengcheng Laboratory (Grant 2025KF1B0050); the Major Key Project of PCL (Grant Nos. PCL2024A04, PCL2025A02); the National Natural Science Foundation of China (NSFC) (Grant Nos. 62206110, 62176103, 62377208, 62572498, and 62276114); and the Science and Technology Planning Project of Guangzhou (Grant Nos. 2024A04J9896, 2025A03J3565).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Paul Luc, Antoine Miech, Ian Barr, Yana Hasson, Lucas Leute, Katie Millican, Malcolm Reynolds, Roy Ring, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[2] 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.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[4] Zahra Ashktorab, Michael Desmond, Qian Pan, James M Johnson, Martin Santillan Cooper, Elizabeth M Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. Aligning human and llm judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. *arXiv preprint arXiv:2410.00873*, 2024.

[5] David P Ausubel. The psychology of meaningful verbal learning. 1963.

[6] Firstname1 Author1, Firstname2 Author2, and Firstname3 Author3. The claude 3 model family: Opus, sonnet, haiku. *Proceedings of the Conference/Journal Name*, Volume Number(Issue Number):Page Range, 2023.

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023.

[9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[10] Roger E Beaty, Alexander P Christensen, Mathias Benedek, Paul J Silvia, and Daniel L Schacter. Creative constraints: Brain activity and network dynamics underlying semantic interference during idea production. *NeuroImage*, 148:189–196, 2017.

[11] Maxi Becker and Roberto Cabeza. Assessing creativity independently of language: A normed language independent remote associate task (li-rat). *Europe PMC free article*, 2021.

[12] Maxi Becker and Roberto Cabeza. Assessing creativity independently of language: A language-independent remote associate task (li-rat). *Behavior Research Methods*, 55(1):85–102, 2023.

[13] Mathias Benedek and Aljoscha C. Neubauer. Revisiting mednick's model on creativity-related differences in associative hierarchies. evidence for a common path to uncommon thought. *The Journal of Creative Behavior*, 47:273 – 289, 2013.

[14] Théophile Bieth, Marcela Ovando-Tellez, Alizée Lopez-Persem, Béatrice Garcin, Laurent Hugueville, Katia Lehongre, Richard Levy, Nathalie George, and Emmanuelle Volle. Time course of eeg power during creative problem-solving with insight or remote thinking. *Human Brain Mapping*, 45(1):e26547, 2024.

[15] Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. *arXiv preprint arXiv:2407.19474*, 2024.

[16] Edward M Bowden and Mark Jung-Beeman. Normative data for 144 compound remote associate problems. *Behavior research methods, instruments, & computers*, 35:634–639, 2003.

[17] Denise J Cai, Sarnoff A Mednick, Elizabeth M Harrison, Jennifer C Kanady, and Sara C Mednick. Rem, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences*, 106(25):10130–10134, 2009.

[18] Wenhu Chen, Hongmin Wang, Junkun Song, Shiji Tang, Ming-Wei Chang, and William Yang Wang. Bongard-logo: A new benchmark for human-level concept learning and reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21505–21517, 2021.

[19] Ziliang Chen, Xin Huang, Quanlong Guan, Liang Lin, and Weiqi Luo. A retrospect to multi-prompt learning across vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22190–22201, 2023.

[20] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[21] Ziliang Chen, Xin Huang, Xiaoxuan Fan, Keze Wang, Yuyu Zhou, Quanlong Guan, and Liang Lin. Reproducible vision-language models meet concepts out of pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14701–14711, 2025.

[22] David Cropley. Is artificial intelligence more creative than humans?: Chatgpt and the divergent association task. *Learning Letters*, 2:13–13, 2023.

[23] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges. 2023.

[24] J Barton Cunningham, James N MacGregor, Jenny Gibb, and Jarrod Haar. Categories of insight and their correlates: An exploration of relationships among classic-type insight problems, rebus puzzles, remote associates and esoteric analogies. *The Journal of Creative Behavior*, 43(4):262–280, 2009.

[25] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[26] Google DeepMind. Gemini 2.5 flash. Website, 2025. `https://deepmind.google/technologies/gemini/flash/`.

[27] Google DeepMind. Gemini 2.5 pro. Website, 2025. `https://deepmind.google/technologies/gemini/pro/`.

[28] DeepSeek-AI. Deepseek-v3 technical report, 2024.

[29] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[30] Xiuliang Duan, Dating Tan, Liangda Fang, Yuyu Zhou, Chaobo He, Ziliang Chen, Lusheng Wu, Guanliang Chen, Zhiguo Gong, Weiqi Luo, et al. Reason-and-execute prompting: Enhancing multi-modal large language models for solving geometry questions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6959–6968, 2024.

[31] Chaoyou Fu, Peixian Chen, Yunhang Shen, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, 2024.

[32] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[33] Wentao Ge, Shunian Chen, Guiming Hardy Chen, et al. Mllm-bench: Evaluating multimodal llms with per-sample criteria. *Preprint*, 2024.

[34] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

[35] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, page 398–414, 2019.

[36] Joy P Guilford. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1):3–14, 1967.

[37] David J Hargreaves. Musical imagination: Perception and production, beauty and creativity. *Psychology of music*, 40(5):539–557, 2012.

[38] José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48:397–447, 2017.

[39] Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. *arXiv preprint arXiv:2503.11117*, 2025.

[40] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[41] Yennie Jun. Exploring creativity in large language models: From gpt-2 to gpt-4.

[42] Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, et al. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109, 2025.

[43] Bohao Li, Yuying Ge, Guangzhi Wang, et al. Seed-bench-2: Benchmarking multimodal large language models. *Preprint*, 2023.

[44] Bing Li, Qi Li, Tianle Yang, Bowen Zheng, Yufeng Xie, Dawei Qi, Yue Zhang, Xiaoyan Zhu, and Jie Tang. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[45] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[46] Hong Li, Nanxi Li, Yuanjie Chen, Jianbin Zhu, Qinlu Guo, Cewu Lu, and Yong-Lu Li. The labyrinth of links: Navigating the associative maze of multi-modal llms. *arXiv preprint arXiv:2410.01417*, 2024.

[47] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C.H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[48] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, WayneXin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models.

[49] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*, 2024.

[50] Yibai Li, Xiaolin Lin, Zhenghui Sha, Zhiye Jin, and Emily Lee. Ai psychometrics: Evaluating the psychological reasoning of large language models with psychometric validities. 2025.

[51] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[52] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.

[53] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[54] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models. 2023.

[55] Haotian Liu, Chunyuan Li, Qingyang Wu, YongJae Lee, Madison Madison, and Microsoft Research. Visual instruction tuning.

[56] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[57] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641, 2023.

[58] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, and Xiang Bai. On the hidden mystery of OCR in large multimodal models. *CoRR*, abs/2305.07895, 2023.

[59] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, pages 216–233. Springer, 2024.

[60] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 2025.

[61] Pan Lu, Rui Shi, Kun Zhao, Simiao Zuo, Michael Zeng, and Sijia Liu. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[62] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019.

[63] Sarnoff A. Mednick. The associative basis of the creative process. *Psychological review*, 69:220–32, 1962.

[64] Sarnoff A Mednick. The remote associates test. *The Journal of Creative Behavior*, 2(3):213–214, 1962.

[65] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. 2023.

[66] Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118, 2021.

[67] Ana-Maria Olteţeanu and Faheem Hassan Zunjani. A visual remote associates test and its validation. *Frontiers in psychology*, 11:26, 2020.

[68] Ana-Maria Olteţeanu, Mikkel Schöttner, and Susanne Schuberth. Computationally resurrecting the functional remote associates test using cognitive word associates and principles from a computational solver. *Knowledge-Based Systems*, 168:1–9, 2019.

[69] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[70] OpenAI. Introducing openai o3 and o4-mini. Website, 2025. `https://openai.com/index/introducing-o3-and-o4-mini/`.

[71] Marcela Ovando-Tellez, Yoed N Kenett, Mathias Benedek, Matthieu Bernard, Joan Belo, Benoit Beranger, Theophile Bieth, and Emmanuelle Volle. Brain connectivity-based prediction of combining remote semantic associates for creative thinking. *Creativity Research Journal*, 35(3):522–546, 2023.

[72] Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024.

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[74] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024.

[75] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

[76] Claudia Schon, Ulrich Furbach, and Marco Ragni. Modeling associative reasoning processes. *arXiv preprint arXiv:2201.00716*, 2022.

[77] Masaki Suwa and Barbara Tversky. Constructive perception: A metacognitive skill for coordinating perception and conception. In *Proceedings of the 25th Annual Cognitive Science Society*, pages 1140–1145. Psychology Press, 2013.

[78] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5100–5111, 2019.

[79] Gemini Team and Google Google. Gemini: A family of highly capable multimodal models.

[80] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[81] Classical Divergent Thinking. 19 associative and controlled cognition in divergent thinking: Theoretical, experimental, neuroimaging evidence, and new directions. *The Cambridge Handbook of the Neuroscience of Creativity*, page 333, 2018.

[82] Oshin Vartanian, Erin L Beatty, Ingrid Smith, Sarah Forbes, Emma Rice, and Jenna Crocker. Measurement matters: the relationship between methods of scoring the alternate uses task and brain activation. *Current Opinion in Behavioral Sciences*, 27:109–115, 2019.

[83] Deana Vitrano, Jeanette Altarriba, and Deniz Leblebici-Basar. Revisiting mednick's (1962) theory of creativity with a composite measure of creativity: The effect of stimulus type on word association production. *The Journal of Creative Behavior*, 2021.

[84] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

[85] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023.

[86] Xiting Wang, Liming Jiang, Jose Hernandez-Orallo, David Stillwell, Luning Sun, Fang Luo, and Xing Xie. Evaluating general-purpose ai with psychometrics. *arXiv preprint arXiv:2310.16379*, 2023.

[87] Jamie Ward, Daisy Thompson-Lake, Roxanne Ely, and Flora Kaminski. Synaesthesia, creativity and art: What is the link? *British Journal of Psychology*, 99(1):127–141, 2008.

[88] R.A. Wolf. Defining the concept of creativity, 2014.

[89] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[90] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.

[91] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2024.

[92] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, Jing Shao, and Wanli Ouyang. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[93] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

[94] Jiahui Yu, Xiaowei Chen, Jingren Shen, Lu Yuan, Wei Chang, and Thomas S Huang. Coca: Contrastive captioners are image-text foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[95] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities.

[96] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[97] Fengji Zhang, Linquan Wu, Huiyu Bai, Guancheng Lin, Xiao Li, Xiao Yu, Yue Wang, Bei Chen, and Jacky Keung. Humaneval-v: Evaluating visual understanding and reasoning abilities of large multimodal models through coding tasks. *arXiv preprint arXiv:2410.12381*, 2024.

[98] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Lijuan Yang, Lei Zhang, Yejin Wang, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, 2021.

[99] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

[100] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models.

[101] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[102] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257, 2024.

[103] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[104] Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Zachary A Pardos, Patrick C Kyllonen, Jiyun Zu, Qingyang Mao, Rui Lv, Zhenya Huang, et al. From static benchmarks to adaptive testing: Psychometrics in ai evaluation. *arXiv preprint arXiv:2306.10512*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately capture the our contributions and scope by outlining the MM-OPERA benchmark, its innovative evaluation approach, and the insights it provides into the limitations of existing Large Vision-Language Models.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we have included the discussion of limitations in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The benchmark work do not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, the paper provides model names, versions, complete experimental setups, and includes the dataset and code for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, the paper offers open access to data and code with detailed instructions in the supplemental material for reproducing results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper details data statistics, data curation process and hyperparameters (e.g. temperature) to ensure result comprehension.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No, error bars are not provided due to the inherent uncertainty in LLMs and LVLMs, making statistical significance challenging to report.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the experimental setup section specifies GPU types and memory sizes required for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, human subjects received appropriate compensation, data sources were publicly available with consent, experiments avoided harmful consequences, and data/code release included proper licenses.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we have included the discussion of broader impacts section in the appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Yes, we have described how we avoid unsafe images during the data curation process.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, original dataset authors are credited via citations, and all public data usage complies with specified licenses and terms.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, new assets are thoroughly documented, with documentation provided alongside them.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: No, due to the offline nature of the process and privacy concerns, full instructions and screenshots are not provided, but the process is compliant and adheres to scientific principles.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: No, the experiments involve no risks to participants, and all necessary approvals were obtained.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Yes, the use of LLM-as-a-Judge, a core and innovative method, is fully described in the paper. We ensure that all content is correct and original.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Technical Appendices

- Section A: More Benchmark Details
- Section B: Supplementary Results and Analysis
- Section C: Case Study 1: Why Do Models Perform Poorly?
- Section D: Case Study 2: Success Cases
- Section E: Test and Evaluation Prompts
- Section F: Limitations and Broader Impacts

A More Benchmark Details

A.1 Association Path

We define three types of association paths to systematically represent different patterns of associative reasoning.

Type 1: Sequential Association

- Structure: $A \to X_1 \to X_2 \to B$
- Format:

Predicate1(A, X_1) and Predicate2(X_1, X_2)
Predicate3(X_2, B)
$A \to X_1 \to X_2 \to B$

Type 2: Convergent Association

- Structure: $A \to X_1 \to X_2$ and $B \to X_2$
- Format:

Predicate1(A, X_1) and Predicate2(X_1, X_2)
Predicate3(B, X_2)
$A \to X_1 \to X_2$ and $B \to X_2$

Type 3: Metaphorical Association

- Structure: $A \land B \to X$
- Format: $A \land B \to X$

Notation Conventions: Entities and predicates follow PascalCase naming convention. The symbol 'and' connects separate relational clauses, while '$\land$' represents logical conjunction between entities. Each arrow ($\to$) represents one associative hop.

While the examples above demonstrate paths with one or three hops, the actual number of intermediate nodes (X_i) and associative steps may vary depending on the complexity of the reasoning process.

A.2 Hierarchical Association Annotation

We develop a hierarchical annotation framework to systematically evaluate multimodal associative reasoning abilities. The framework consists of three levels that progress from basic perception to complex conceptual reasoning:

Level-1 (L-1) divides associative abilities into two fundamental categories:

- *Perception*: Processes immediate sensory inputs, focusing on visual understanding and interpretation
- *Conception*: Handles abstract, knowledge-driven associations requiring higher-order cognitive processing

Level-2 (L-2) further refines these categories into six dimensions:

- Under *Perception*: Recognition, Context, and Interaction
- Under *Conception*: Logic, Semantic, and Reasoning

Level-3 (L-3) provides the most granular classification with thirteen specific dimensions. Each dimension captures a distinct aspect of associative reasoning. Table 4 presents detailed definitions for each dimension.

Table 4: Detailed Definitions of Hierarchical Association Dimensions

L-1	L-2	L-3
Perception	**Recognition**	**Visual Similarity** Associations based on visual features like shape, color, texture, and appearance.
		Semantic Object High-level semantic recognition of objects, including fine-grained identification in specific contexts.
	Context	**Contextual Sensory Cues** Perceptual associations based on visual details like tone, lighting, and spatial layout.
		Scene Contextualization Understanding of overall scene context, including atmosphere and purpose.
		Abstract Interpretation Recognition of abstract concepts and symbolic patterns.
	Interaction	**Social Insight** Understanding emotions and interactions between people in visual scenes.
		Relational Perception Comprehension of spatial and logical relationships between objects.
Conception	**Logic**	**Functional Links** Associations based on functional relationships between concepts.
		Causal Connections Associations based on cause-and-effect relationships.
	Semantic	**Thematic Links** Associations within the same theme or context.
		Cultural Reference Associations based on cultural knowledge and specific contexts.
	Reasoning	**Hierarchical Association** Vertical associations between abstract and concrete concepts.
		Analogical Reasoning Associations based on structural, feature, or pattern similarities.

This hierarchical framework enables systematic evaluation of LVLMs' associative abilities across different cognitive levels, from basic sensory processing to sophisticated abstract reasoning. The progression from L-1 to L-3 mirrors human cognitive development and provides a comprehensive structure for analyzing multimodal understanding capabilities.

A.3 Data Sources

The MM-OPERA-Bench dataset, consisting of images, reference answers, and fine-grained annotations, was manually curated by a group of volunteers. Of the total data, 33.35% of the questions and reference answers were sourced from the RAT [64], while 4.01% of the images, questions, and reference answers were sourced from the LI-RAT [11] datasets for human psychometric testing. The remaining images were sourced from the Internet, and all fine-grained annotations were manually constructed and revised to ensure consistency and accuracy.

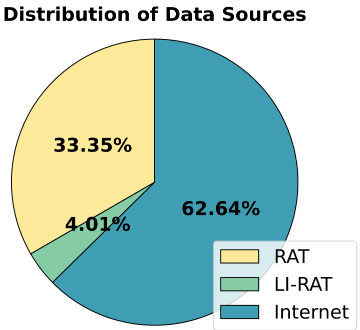


Figure 4: Distribution of data sources.

A.4 Data Collection and Curation Protocol

The MM-OPERA dataset, encompassing images, reference answers (including reasoning paths), and multifaceted annotations, was meticulously curated through a multi-stage, volunteer-driven process, adhering to ethical guidelines and scientific rigor.

A.4.1 Data Collection

Volunteers, primarily undergraduate and graduate students from diverse disciplines (STEM, humanities, social sciences, arts), were recruited via university channels to leverage their strong cognitive abilities and varied perspectives, enriching the dataset creation process with a broad range of knowledge and insights. They received training on project goals, task details, data privacy considerations (anonymized contributions), and time commitment. Guidelines covered associative attribute definitions, example generation, image sourcing (avoiding unsafe or inappropriate content), and annotation consistency. Participation was voluntary, with contributors acknowledged.

A core research team created 10–20 high-quality seed instances for each of the 13 Level-3 (L-3) associative attributes as exemplars. Volunteers expanded the dataset by sourcing images from public repositories (e.g., Wikimedia Commons, public domain archives) and adapting items from psychometric tests (e.g., RAT and LI-RAT). They were trained to exclude images depicting illegal, violent, or offensive content, using safe search filters and careful judgment. Sourced items underwent manual revision to ensure appropriateness, clear associative links, plausible reasoning paths, and diverse associations beyond original tests. Volunteers were also guided to create instances reflecting cultural contexts, linguistic nuances (English-based items testing concepts across 15 linguistic backgrounds), and thematic domains (22 topic domains to avoid biases). A tracking system ensured balanced coverage, prompting targeted collection if gaps were identified.

A.4.2 Quality Control

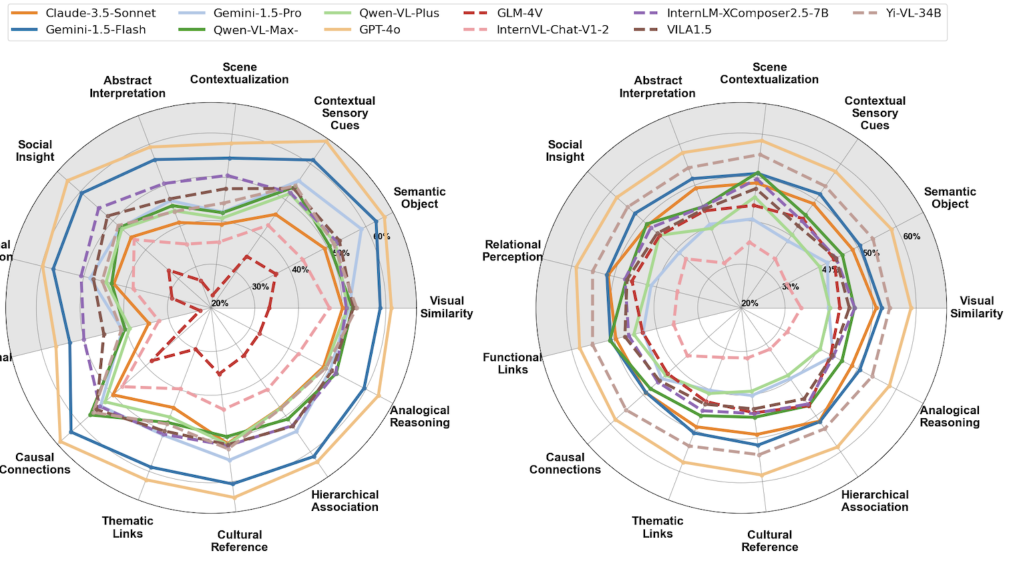
A multi-layered quality control process ensured accuracy, clarity, challenge, and safety of the MM-OPERA benchmark. Each instance underwent initial screening by the core team for guideline adherence, including checks for inappropriate images. A two-stage peer review followed: (1) **Cross-Review**: Two uninvolved volunteers assessed clarity, relevance, reasoning plausibility, formatting, and image safety, providing revision feedback. (2) **Expert Review**: Core researchers evaluated conceptual soundness, difficulty, biases, and safety, discarding or revising problematic items. Five core team members then assessed instance difficulty (Easy, Medium, Hard, Very Hard) based on association remoteness, reasoning complexity, and cue subtlety. Consensus was reached through discussion. Approximately 5% of instances (too trivial or obscure) were excluded to ensure meaningful challenges. Feedback from quality control refined guidelines and training.

Crucially, what sets our validation apart is the **structured Association Reasoning Path** included with every instance thanks to our Process-Reward LLM-as-a-Judge method. Reviewers validated the entire step-by-step logical chain, ensuring that the association is not only plausible but also coherently and correctly reasoned. This traceable reasoning provides a far more robust and objective measure of correctness than benchmarks with only a final label, significantly mitigating the risk of flawed or ambiguous examples.

The final dataset includes only instances passing all review and calibration stages, ensuring a high-quality, diverse, challenging, and safe benchmark for evaluating associative reasoning in LVLMs.

B Supplementary Results and Analysis

B.1 Multi-dimensional Analysis



(a) RIA

(b) ICA

Figure 5: Comparison of Model Performance in RIA and ICA across Different Conceptual (white background) and Perceptual (gray background) Dimensions. The radar charts illustrate the capabilities of various LVLMs in handling tasks related to relational perception, social insight, causal connections, abstract interpretation, and other cognitive functions. The left chart (RIA) exhibits greater variability in model performance, while the right chart (ICA) shows more consistent trends across models.

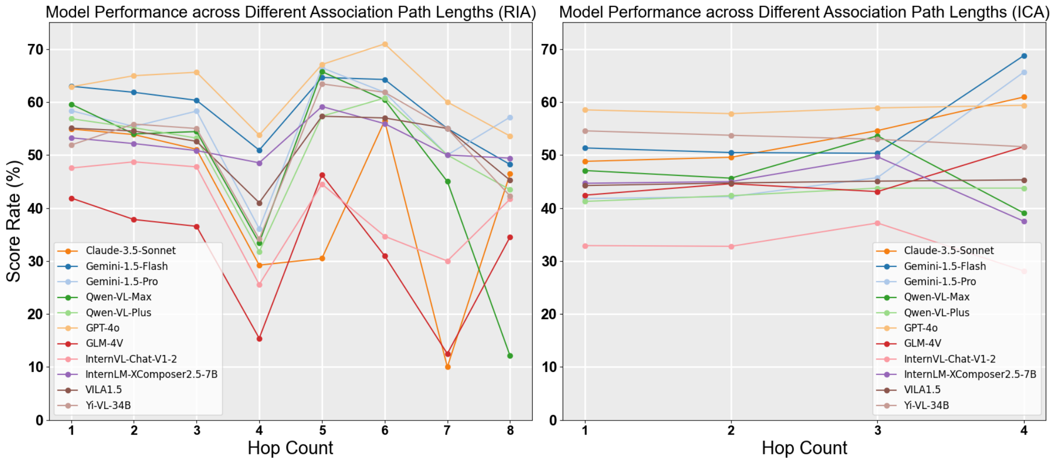


Figure 6: Model Performance Across Different Association Path Lengths in RIA and ICA tasks. The line graphs illustrate the score rates (%) of various LVLMs as the number of association path "hops" increases. The left chart represents RIA results, showing notable fluctuations in performance across different hop counts. The right chart represents ICA results, where models generally display more stable trends. This analysis highlights how different models handle varying levels of associative complexity.

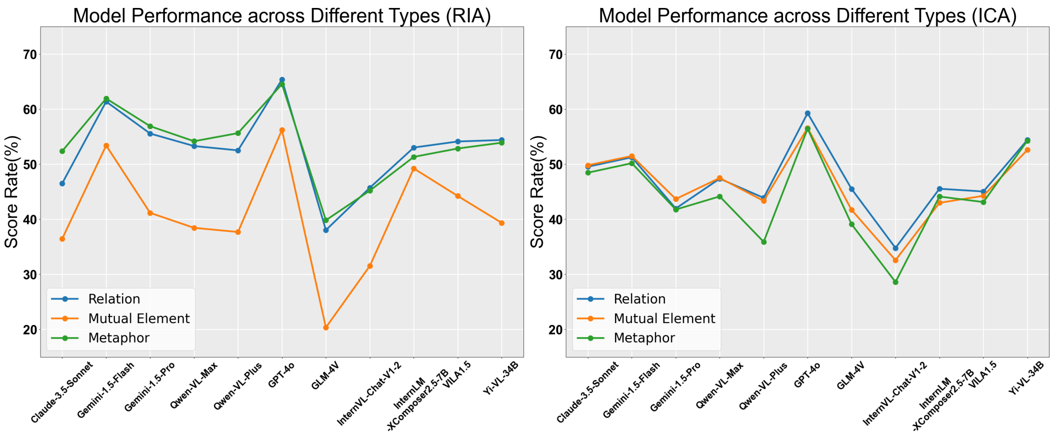
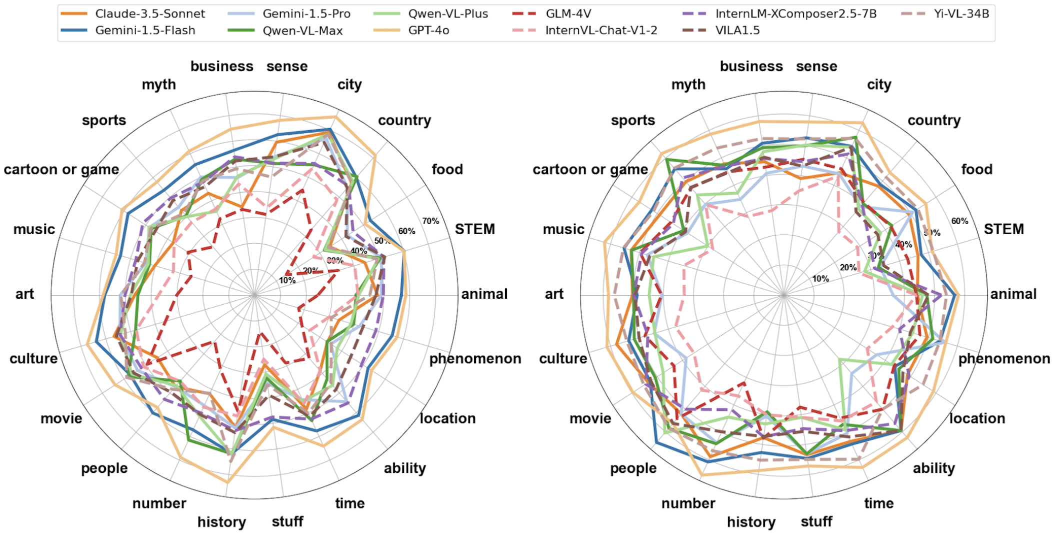


Figure 7: The figure presents the performance of various multimodal large models across different reasoning types in the RIA (left) and ICA (right) tasks. The three reasoning types—Relation, Mutual Element, and Metaphor—are represented by different colored lines. The vertical axis indicates the score rate (%), while the horizontal axis lists different models. The results show varying performance trends across reasoning types and tasks, highlighting differences in model capabilities in handling relational, compositional, and metaphorical understanding.



(a) RIA

(b) ICA

Figure 8: Radar Chart Comparison of Model Performance Across Domains in RIA and ICA tasks. The two radar charts display the performance of various LVLMs across different knowledge domains such as business, sports, music, STEM, history, and culture. The left chart (a) represents results from the RIA tasks, while the right chart (b) shows ICA results. The models exhibit varying performance across different domains, with some excelling in specific categories while struggling in others.

Multidimensional analysis reveals the complex landscape of associative reasoning capabilities in Large Vision Language Models (LVLMs). Most models perform better on RIA tasks compared to ICA tasks, with an average performance differential of approximately 5–7 percentage points. This suggests that identifying direct associations between unrelated items may be more tractable for current LVLMs than recognizing and extending associative patterns. The exception is Claude-3.5-Sonnet, which shows relatively consistent performance across both task types, indicating potentially more balanced associative reasoning capabilities. These findings underscore the multi-faceted nature of associative cognition and the importance of diverse task designs for comprehensive evaluation.

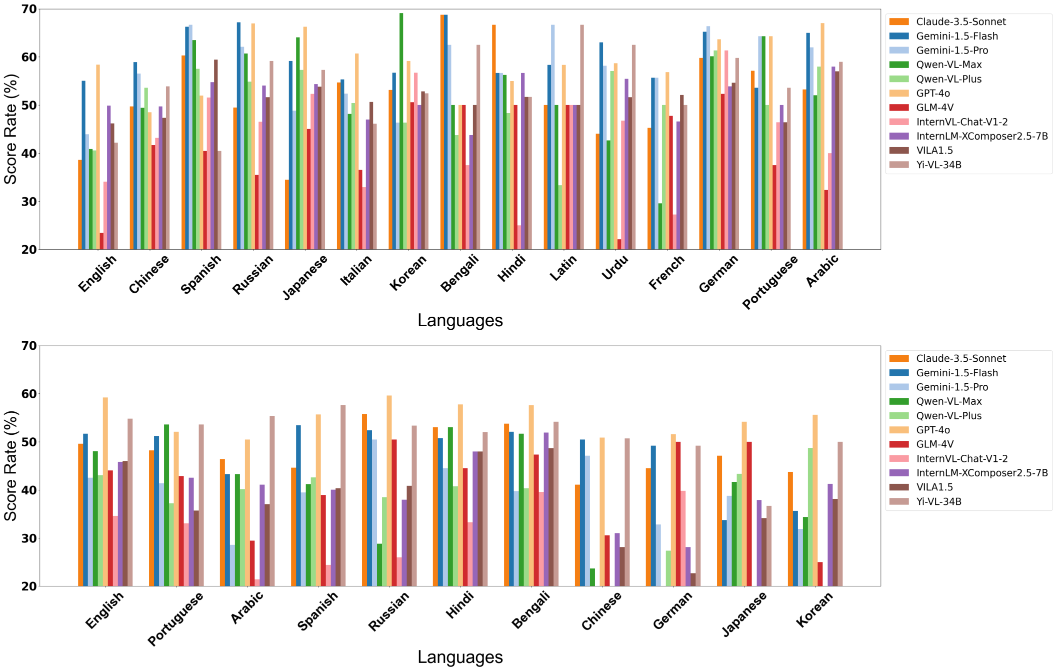


Figure 9: The figures compare the performance of different multimodal large models across multiple languages. The left figure represents the RIA tasks, while the right figure corresponds to the ICA tasks. The vertical axis indicates the score rate (%), and the horizontal axis lists various languages, including English, Chinese, Spanish, and others. The results highlight significant differences in model performance across tasks and languages, reflecting their varying capabilities in cross-linguistic understanding and reasoning.

Reasoning Complexity and Cognitive Abilities. Analysis of reasoning complexity reveals non-linear patterns in how models handle associative tasks. Figure 6 shows that while most models effectively manage simple 1–2 hop associations (with score rates around 50–60%), performance drops significantly for more complex 4-hop associations (29–53%). However, some models (e.g., GPT-4o, Gemini-1.5-Flash, and Qwen-VL-Plus) demonstrate relatively stable performance in very complex 5–8 hop associations, suggesting the emergence of new strategies in complex reasoning paths. This "complexity valley" phenomenon warrants further investigation as it may provide important insights into how LVLMs structure multi-step associative reasoning. Differences in perceptual and conceptual abilities are evident in Figure 5. "Semantic Object" in perceptual abilities and "Causal Connections" in conceptual abilities show stronger performance, while "Abstract Interpretation" and "Functional Links" remain challenging. Cross-task analysis indicates that models maintain consistent relative strengths across RIA and ICA tasks, but absolute performance levels are modulated by task demands, especially for perceptual abilities. This suggests that while underlying reasoning mechanisms remain stable, their expression is influenced by task requirements.

Relationship Types. Analyzing association types reveals distinctive performance patterns between Remote-Item Association (RIA) and In-Context Association (ICA) tasks. As shown in Figure 7, in RIA tasks, models demonstrate a pronounced hierarchy among association types, with Metaphor associations yielding the highest performance (52–65% for top performers), followed by Relation associations (46–65%), and Mutual Element associations showing the lowest scores (36–56%). This hierarchy is notably consistent across nearly all models. Interestingly, in ICA tasks, this performance stratification significantly diminishes, with much smaller performance gaps between association types. For instance, GPT-4o shows only a 2.82 percentage point difference between its highest (Relation: 59.29%) and lowest (Mutual Element: 56.59%) association type performance in ICA, compared to a 7.15 point gap in RIA. This convergence suggests that the contextual framework provided in ICA tasks may equalize the difficulty of recognizing different association types.

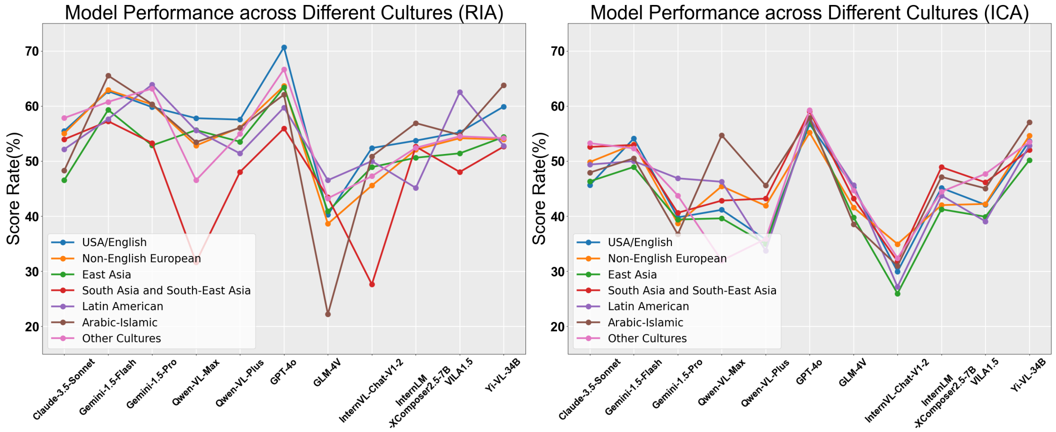


Figure 10: Comparison of Model Performance Across Different Cultures in RIA and ICA tasks. The line plots illustrate the score rates (%) of various LVLMs across cultural groups, including USA/English, Non-English European, East Asian, South Asian and South-East Asian, Latin American, Arabic-Islamic, and Other Cultures. The left graph represents RIA results, while the right graph shows ICA results

Domain and Cultural Dimensions. Domain knowledge differences are highly evident in model performance. As shown in Figure 8, urban-related associations consistently achieve higher performance (around 65–75% for top models), while everyday objects and food-related associations pose greater challenges (around 30–45%). These differences suggest inherent difficulties in forming associations within certain conceptual spaces. GPT-4o excels in history-related associations (73.07%), significantly outperforming other models, which may indicate superior historical knowledge representation or more effective temporal association retrieval mechanisms. Cultural background also significantly impacts model performance. Figure 10 reveals that most models show stronger associative reasoning when dealing with Western cultural references compared to East Asian, South Asian, or Arabic-Islamic contexts. In Figure 9, similar asymmetries are observed in language performance, with models generally performing better in Spanish, German, and Russian associations than in East Asian languages. These cultural and linguistic disparities may reflect imbalances in multilingual pretraining or fundamental differences in how associations manifest across different linguistic and cultural structures.

Multilingual Capability. The current language distribution (in Figure 2.e) is a principled design choice to robustly evaluate distinct aspects of associative reasoning. On one hand, English serves as the dominant language. Current LVLMs are primarily trained on English-centric data, so using English as the primary language offers a fair and stable baseline for evaluating core capabilities. On the other hand, to achieve a more comprehensive evaluation, we additionally include non-English samples, which serve as targeted probes for culturally nuanced associative phenomena. These include challenges such as linguistic wordplay (e.g., homophones or puns unique to a given language) and cultural knowledge (e.g., proverbs, historical references, or artistic expressions that require deep, language-specific world knowledge). The design of non-English samples requires the cultural context associated with each language, resulting in more complex construction logic and thus a more limited pool of suitable examples. This is why non-English samples follow a long-tail distribution.

To provide a more balanced view of multilingual capability, we report the harmonic mean (H_{SR}) of the score rate (SR) on English and non-English samples in Table 5. This metric mitigates the dominance of the larger English subset and is sensitive to performance disparities across languages. The H_{SR} confirms that top models like Gemini-2.5-Pro and o4-mini demonstrate strong, balanced capabilities. This reinforces our conclusions while offering a more nuanced view of multilingual performance.

We also observe an interesting fact that in the RIA task, most models achieve a higher SR on non-English samples. We hypothesize this is because these instances often test specific, well-defined cultural knowledge (e.g., proverbs). A successful association relies on retrieving a precise factual link, which models with broad world knowledge can do effectively. Conversely, in the ICA task, many top

	Remote-Item Association (RIA) Task			In-Context Association (ICA) Task		
Model	H_{SR} (%)	Eng SR (%)	Non-English SR (%)	H_{SR} (%)	Eng SR (%)	Non-English SR (%)
Proprietary LVLMs						
Claude-3.5-Sonnet	52.92	48.37	58.41	48.94	49.66	48.25
Gemini-1.5-Flash	58.12	55.28	61.27	50.13	51.72	48.64
Gemini-1.5-Pro	49.70	44.17	56.80	41.63	42.55	40.74
Qwen-VL-Max	51.05	44.84	59.27	46.22	51.50	41.92
Qwen-VL-Plus	47.47	41.24	55.90	42.93	46.05	40.21
Gemini-2.0-Flash-Thinking-Exp	60.93	58.58	63.47	59.56	62.72	56.71
Gemini-2.5-Pro-Preview	**63.53**	58.96	68.87	**61.05**	64.53	57.94
o4-mini	62.39	59.74	65.28	59.90	62.72	57.32
GPT-4o	62.02	59.27	65.04	56.94	59.21	54.83
OpenSource LVLMs						
GLM-4V	32.48	25.00	46.35	42.68	44.32	41.15
InternVL-Chat-V1-2	40.72	35.14	48.41	31.08	36.37	27.14
InternLM-XComposer2.5-7B	**51.36**	49.98	52.82	43.52	45.83	41.44
VILA1.5	49.10	46.14	52.45	42.17	46.00	38.92
Yi-VL-34B	50.26	43.98	58.65	**53.51**	55.04	52.06

Table 5: The harmonic mean of the score rate (SR) on English and non-English samples of various LVLMs on the Remote-Item Association (RIA) and In-Context Association (ICA) tasks. The best-performing model in each sub-category is highlighted in bold.

	Remote-Item Association Task				In-Context Association Task			
	Holistic SR (%)		Avg. Reasoning Score		Holistic SR (%)		Avg. Reasoning Score	
Model	4o	V3	4o	V3	4o	V3	4o	V3
Claude-3.5-Sonnet	56.20	60.10	1.2838	1.5457	50.4	48.15	0.4159	0.6039
Gemini-1.5-Flash	63.25	60.65	1.2701	1.5684	51.35	54.55	0.3507	0.3985
Gemini-1.5-Pro	56.90	59.80	1.2701	1.4908	40.55	42.20	0.1742	0.2674
Qwen-VL-Max	49.30	59.60	1.2587	1.3733	44.60	56.15	0.5584	0.7107
Qwen-VL-Plus	54.50	57.90	1.0511	1.4212	43.00	52.35	0.4011	0.5791
GPT-4o	**67.80**	**67.75**	**1.4676**	**1.7459**	**59.80**	**59.60**	**0.5611**	**0.7180**
InternLM-XComposer2.5-7B	52.80	55.95	1.1902	1.5345	45.15	50.75	0.1560	0.2727
VILA1.5	54.25	57.20	0.9979	1.2788	43.30	52.65	0.3122	0.5259
Yi-VL-34B	57.65	59.20	1.1424	1.3502	52.85	53.35	0.3107	0.4027

Table 6: Performance comparison of models on 500 sampled Remote-Item Association and In-Context Association instances as evaluated by two judges (GPT-4o and Deepseek-V3). Metrics include holistic score rate (SR) and average reasoning score. The highest values for each metric are bolded, while the second-highest are underlined.

models perform better on English samples. ICA demands abstracting and transferring a relational pattern, a meta-reasoning skill. We posit this capability is more robustly developed for English, the primary language in pre-training data, where such abstract logical structures are more prevalent.

Note that although some items were designed with specific cultural or cultural contexts in mind, models may still generate alternative but valid associations without explicitly relying on those cues. Our open-ended evaluation rewards any well-justified reasoning path, so scores may reflect reasoning flexibility or general knowledge rather than direct cultural or linguistic awareness. Thus, performance differences cannot be solely attributed to language proficiency.

B.2 Evaluation by Different Judges

The comparison between the two judges (GPT-4o and Deepseek-V3) highlights notable differences in their scoring tendencies (see Table 6). Judge Deepseek-V3 consistently assigns higher average reasoning scores across most models, suggesting a more lenient evaluation of reasoning depth or quality. However, for holistic score rate (SR), the differences are less consistent, with some models (e.g., Claude-3.5-Sonnet and Qwen-VL-Max) receiving higher SR from V3, while others (e.g., GPT-4o) show near parity between the two judges. These disparities underscore the importance of employing multiple evaluators to mitigate individual judgment bias and ensure robust evaluation of model performance.

The visualized score distributions in Figure 11 further highlight key differences in evaluation tendencies and scoring patterns between GPT-4o and Deepseek-V3 across different models and tasks.

Calibration Pattern Comparison. The visualized reasoning score distributions reveal distinctive evaluation tendencies between GPT-4o and Deepseek-V3 across models and tasks. While both evaluators maintain similar distribution shapes for each model, Deepseek-V3 consistently demonstrates a broader scoring range, particularly on RIA tasks where it occasionally assigns scores of 5–6 to top-performing models like GPT-4o and Claude-3.5-Sonnet—scores beyond GPT-4o's 0–4 scale. This suggests Deepseek-V3 employs a more granular assessment framework with higher ceiling effects. Additionally, GPT-4o shows more concentrated distributions with sharper peaks, while Deepseek-V3 exhibits more dispersed distributions, particularly in the mid-range scores. Despite these calibration differences, both evaluators converge on identifying the same relative performance hierarchy across models and consistently highlight the challenging nature of ICA tasks, where all models receive predominantly low scores (0-1) regardless of which system performs the evaluation.

Evaluator Consistency and Minimal Self-Enhancement Bias. The holistic score distributions reveal remarkable consistency between GPT-4o and Deepseek-V3 as evaluators, providing strong evidence against significant self-enhancement bias. Despite GPT-4o evaluating its own outputs, both evaluators produce strikingly similar distribution patterns across all models for both RIA and ICA tasks. Notably, GPT-4o does not disproportionately favor its own responses—its self-evaluation distribution closely mirrors Deepseek-V3's independent assessment, with both showing peaks at similar score points. This alignment is particularly evident in the ICA tasks, where both evaluators produce nearly identical bell-shaped distributions centered around scores 2–3 for all models. The consistency across different evaluators suggests that our evaluation framework successfully mitigates potential self-enhancement effects, reinforcing the reliability of our findings even when using an LLM to evaluate its own outputs. This methodological robustness strengthens confidence in the comparative analysis of associative reasoning capabilities across different LVLMs.

Reliable Path Complexity Analysis. Both GPT-4o and Deepseek-V3 extract nearly identical hop count distributions from the same model outputs, reinforcing the reliability of our path analysis methodology. This consistency in path complexity evaluation across different judges provides strong evidence that the observed patterns reflect genuine differences in associative reasoning strategies between tasks rather than evaluator bias.

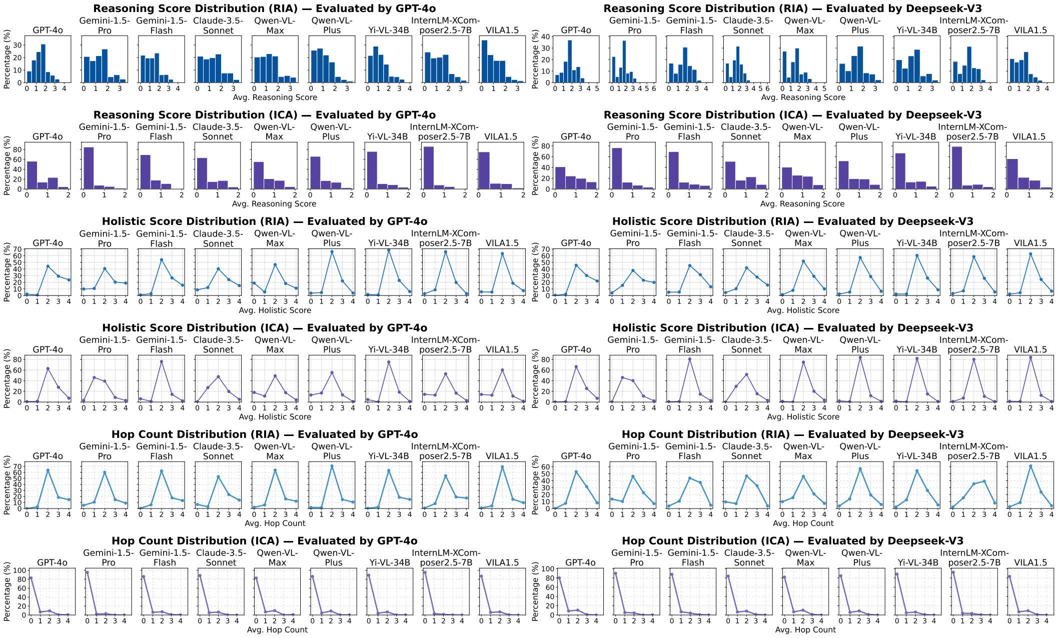


Figure 11: Fine-grained reasoning capability analysis of nine multimodal language models on RIA (blue) and ICA (purple) tasks judged by GPT-4o (left) and Deepseek-V3 (right). From top to bottom: reasoning score distribution, holistic score distribution, reasoning path hop count distribution. Each task includes 500 sampled questions.

B.3 Sensitivity Test Results

Model	IG Range↓	IG SR↑	IG SD↓
Claude-3.5-Sonnet	1.00	0.47	0.41
Gemini-1.5-Flash	0.68	0.55	0.27
Gemini-1.5-Pro	1.22	0.44	0.49
Qwen-VL-Max	0.99	0.43	0.38
Qwen-VL-Plus	1.06	0.41	0.41
GPT-4o	**0.44**	**0.59**	**0.18**
GLM-4V	0.79	0.25	0.30
InternVL-Chat-V1-2	1.31	0.35	0.49
InternLM-XComposer2.5-7B	0.93	0.50	0.36
VILA1.5	1.34	0.46	0.54
Yi-VL-34B	0.99	0.44	0.38

Table 7: Performance of models on the Multi-Image Substitution Test in RIA. We grouped multiple-image variants with identical concept pairs in RIA and measure score variability using *IG Range* (intra-group score range), *IG SR* (average intra-group score rate), and *IG SD* (intra-group standard deviation).

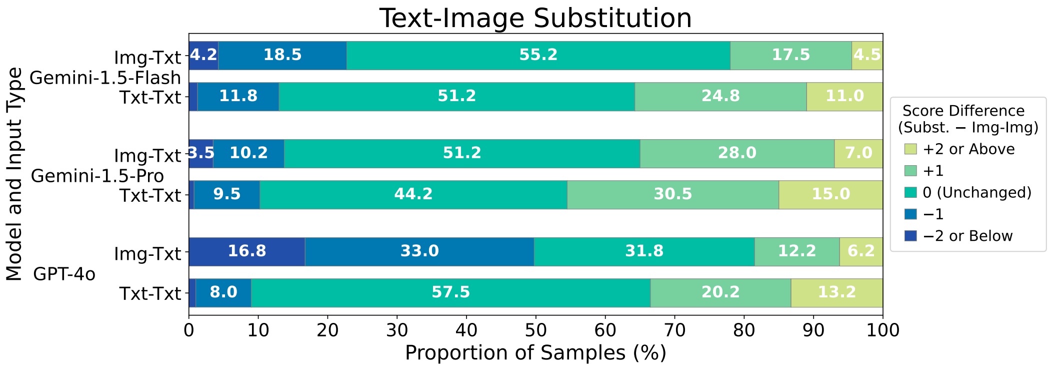


Figure 12: Score difference distribution for the Text-Image Substitution test across models in RIA. Bars show proportions of samples with varying score differences (substitution - original).

B.3.1 Multi-Image Substitution Test

To assess robustness, we conducted sensitivity tests to measure how LVLMs' responses varied with different visual representations of the same concepts. Results in Table 7 revealed significant visual sensitivity across models. GPT-4o demonstrates exceptional consistency, showing the lowest intra-group score range (0.44) and standard deviation (0.18) while maintaining the highest score rate (0.59). In contrast, models like VILA1.5 and InternVL-Chat-V1-2 exhibit substantial variability (IG Ranges of 1.34 and 1.31, respectively) despite moderate performance, indicating that their associative reasoning is heavily influenced by specific visual features rather than robust concept understanding. This visual dependency suggests that most current LVLMs still associate at a surface feature level rather than at a deeper conceptual level—a critical limitation for real-world applications requiring consistent reasoning across variable visual inputs.

B.3.2 Text-Image Substitution Test

Results in Figure 12 reveals distinct cross-modal generalization patterns across models. GPT-4o experiences the most significant performance drop when one image is replaced with the text description (Image-Text), with 49.8% of samples showing decreased scores (16.8% with severe drops of $\geq$2 points). Intriguingly, in the Txt-Txt setting, GPT-4o's performance is much more robust, with scores dropping for around 8.0% of samples. This reveals a deeper insight: GPT-4o may struggle with cross-modal fusion. It performs well when reasoning over vision-only or text-only inputs, but its performance falters when forced to integrate information from disparate modalities (Img-Txt).

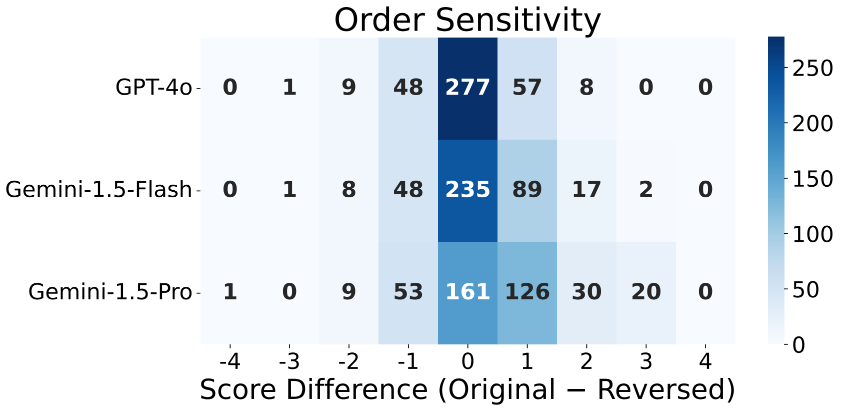


Figure 13: Order Sensitivity heatmap showing models versus score difference between original and reversed input order. Cell darkness indicates instance count.

In contrast, Gemini-1.5-Flash and Pro are far more resilient to modality substitution. In the Img-Txt setting, Gemini-1.5-Flash sees a performance drop in only 22.7% (4.2% + 18.5%) of cases, and scores are unchanged for a majority (55.2%) of samples. This suggests Gemini models may employ a different internal strategy, perhaps by more effectively converting visual inputs into a modality-agnostic, language-like representation.

These patterns also suggest fundamental differences in cross-modal processing strategies. GPT-4o appears more reliant on visual information for associative reasoning, extracting nuanced visual cues that text descriptions cannot fully capture. Meanwhile, Gemini models demonstrate stronger text-equivalence in their reasoning processes, suggesting they may process visual information by internally converting it to language-like representations. This finding highlights the importance of modality-specific evaluation when assessing LVLMs' associative reasoning capabilities.

These experiments have also demonstrated **how visually challenging the benchmark is**. When images are replaced with their text descriptions, the LVLMs' performances stayed the same or improved for over 87% of samples across all tested models (improved by $\geq$1 for 33-46% samples and by $\geq$2 for 11-15% samples). The observations can only be explained by LVLMs struggle more with processing the raw visual input than with reasoning from a 'perfect' text description. It leads to the evidence of the benchmark's visual challenge.

B.3.3 Order Sensitivity Test

Results in Figure 13 reveals varying degrees of input order sensitivity across models. GPT-4o demonstrates the highest stability, with 277 of 400 instances (69.25%) showing no score change when input order is reversed. Gemini-1.5-Flash shows moderate consistency (58.75% unchanged), while Gemini-1.5-Pro exhibits notably lower order invariance (only 40.25% unchanged) with a significant rightward shift toward positive score differences, indicating better performance on the original order. This suggests that while GPT-4o processes image pairs in a more commutative manner, treating both ordering equally, Gemini models—particularly Gemini-1.5-Pro—appear to apply asymmetric reasoning processes that may prioritize the first image as context and the second as the target for association, highlighting architectural differences in how models approach bimodal associative reasoning.

B.4 LLM-as-a-Judge Strategy Validation

For **human alignment** validation, We compared 300 randomly sampled scoring results of the model with those of 8 human evaluators. Furthermore, we analyzed potential biases in the LLM-as-a-Judge evaluation, focusing on **verbosity bias** by examining the correlation between response length and scores, and **position bias** through answer order permutation tests. Both analyses aimed to ensure an objective and consistent evaluation across models and human responses.

B.4.1 Alignment with Human Judgment for Regular LLM-as-a-Judge Scoring

We compared 300 sampled GPT-4o's regular scoring results with those of human evaluators, finding an average score difference of 0.077. Notably, 78.33% of the model's scores perfectly matched those of human judges, with 21.67% of responses aligning within a 1-point difference. Critically, there were no instances of disagreement exceeding a 1-point margin, indicating strong calibration between our automated evaluation and human judgment. This high level of agreement demonstrates the reliability of our LLM-as-a-Judge framework for evaluating open-ended associative responses, effectively balancing the efficiency of automated assessment with the nuanced judgment characteristic of human evaluators. The absence of large scoring discrepancies further validates our approach as a robust proxy for human evaluation in this complex reasoning domain, addressing a key challenge in the assessment of open-ended multimodal tasks.

For the Process-Reward LLM-as-a-Judge (PR-Judge), we randomly selected 200 reasoning paths generated by the models and had them evaluated by 8 human judges with domain expertise. The human judges scored each reasoning step based on the same criteria used by the PR-Judge: Reasonableness R_t, Distinctiveness D_t, and Knowledgeability K_t. The overall reasoning score S_r for each path was then calculated. Our results show that the average reasoning score difference between the human judges and the PR-Judge (GPT-4o) was 0.1961. Specifically, 81% of the paths received scores differed by no more than 0.20 from the PR-Judge and human judges, while 16% differed by no more than 0.50 points, and none had a difference more than 0.60, indicating a high level of agreement between the automated and human evaluations. We also observed strong positive correlations between the PR-Judge's scores and the average human scores: Pearson's $r = 0.72$ for Reasonableness, $r = 0.68$ for Distinctiveness. For the binary Knowledgeability indicator, the PR-Judge achieved an accuracy of 83.5% (Cohen's Kappa = 0.65) compared to the majority human vote. These findings suggest that the PR-Judge effectively captures human-like nuances in assessing the quality of individual reasoning steps.

B.4.2 Effectiveness of Process-Reward LLM-as-a-Judge

To justify the introduction of the Process-Reward LLM-as-a-Judge, we compared its performance with a traditional outcome-based scoring method using the same 100 reasoning paths. We found that the outcome-based method often assigned similar scores to models that produced correct outcomes but through different reasoning processes. For instance, two models might both receive a score of 4 based on their final answers, but the Process-Reward method revealed differences in their reasoning quality, with one model scoring 1.3 and the other 1.8, reflecting the latter's superior reasoning process. This demonstrates that the Process-Reward approach provides a more nuanced evaluation of reasoning quality compared to traditional methods.

B.4.3 Bias Analysis

We investigated potential biases in LLM-based evaluation.

Verbosity bias. Since 1-point responses are significantly shorter due to their vague or uncertain nature, we excluded them and compared the correlation between response length and performance. Our analysis yielded a Pearson Correlation coefficient of 0.376 for regular scoring and 0.291 for PR-Judge. This moderate positive correlation is acceptable, as high-quality responses often require more detailed explanations. The correlation is not strong enough to suggest that the LLM judge is primarily influenced by response length rather than content quality.

Position bias. We performed permutation tests on 500 samples each on RIA and ICA tasks by randomly shuffling the order of the standard and model-generated answers in the judging prompt. The results showed no systematic advantage for any position, with mean score differences across permutations averaging 0.0871 for regular scoring and 0.1563 for Process-Reward scoring. These findings indicate that the evaluation process remains relatively objective and not significantly affected by response length or ordering.

C Case Study 1: Why Do Models Perform Poorly?

To gain deeper insights into the challenges of MM-OPERA-Bench tasks, we analyzed the low-scoring answers provided by GPT-4o, Gemini-1.5-Pro, and Gemini-1.5-Flash. This analysis serves a dual purpose: identifying current limitations of these models and informing future advancements in LVLM design and training methodologies. Specifically, we examined 50 randomly selected low-scoring instances (holistic score ≤ 2) on both the RIA and ICA tasks for each model, investigating the underlying causes of suboptimal performance. It is noteworthy that, due to the inherent complexity of the tasks, a single response may exhibit multiple limitations, resulting in a cumulative contribution of factors exceeding 100%. Furthermore, we present five illustrative case studies, accompanied by detailed analyses, to facilitate further exploration.

Perceptual Misalignment (45%). Models frequently demonstrate an inability to accurately detect salient visual features or to appropriately interpret their significance within the broader associative context. This fundamental perceptual limitation manifests in two primary forms: complete omission of critical visual elements (as exemplified in Case 1, where GPT-4o failed to recognize the QR code embedded within the castle image) or inadequate conceptual abstraction from correctly perceived elements (as illustrated in Case 4, where the model identified visual components but failed to abstract the linguistic concept of "See" from an image depicting an act of looking). These perceptual errors initiate cascading reasoning failures that fundamentally compromise the associative process. More specifically, limitations in image resolution, the presence of visual noise, or a lack of sensitivity to certain visual attributes can lead to perceptual inaccuracies. Furthermore, biases in understanding spatial relationships, relative sizes, and interactions between objects within an image can impede accurate scene interpretation.

Knowledge Retrieval Gap (48%). Despite possessing encyclopedic knowledge within their parameters, LVLMs exhibit difficulty in activating relevant information during multimodal association tasks, particularly across cultural, linguistic, and domain boundaries. Case 3 exemplifies this challenge, wherein Gemini-1.5-Flash failed to retrieve cross-cultural knowledge pertaining to "Sanmao," leading to the generation of spurious connections rather than the identification of the genuine linguistic homonym linking Chinese literature and cartoons. Similarly, in Case 5, Gemini-1.5-Pro was unable to access historical knowledge regarding peach baskets as the original basketball hoops, resulting in erroneous pattern identification. This suggests that knowledge activation, rather than mere knowledge possession, represents a significant bottleneck in multimodal associative reasoning. This can be attributed to inefficient knowledge indexing, fragmented knowledge representation, or delayed knowledge updates. Furthermore, inadequate confidence assessment and source attribution mechanisms can hinder the effective utilization of retrieved knowledge.

Overgeneralization (53%). When confronted with complex or ambiguous associations, models frequently resort to overly broad and imprecise relationships that lack meaningful specificity. This tendency is clearly demonstrated in Case 1, where GPT-4o defaulted to a generic "creativity" association when unable to identify the more specific "hidden symbols" relationship. Similarly, in Case 4, the model proposed an abstract theme of "emphasis and clarity" rather than recognizing the homophonic relationship between musical notes and their verbal counterparts. This pattern reveals a tendency to prioritize plausibility over clarity when faced with challenging associative tasks. This may be influenced by biases in the training data distribution, favoring frequently occurring association patterns. Furthermore, inaccuracies in assessing association strength and calibrating association confidence can lead to an over-reliance on generalized association patterns.

Limited Insight and Excessive Caution (23%). A notable subset of failures stemmed from the models' reluctance to venture beyond superficial observations or to propose connections that require conceptual leaps. Case 2 illustrates this limitation, where Gemini-1.5-Pro correctly identified individual elements (centaur and calendar) but declared them "unrelated" rather than exploring potential symbolic associations through astrological knowledge. This cautious approach restricts the models' ability to discover non-obvious but meaningful connections, a cornerstone of human-like associative thinking. This may be due to pre-programmed constraints that limit the exploration of unconventional reasoning paths. Furthermore, a low tolerance for uncertainty and a high aversion to risk can lead to the adoption of conservative reasoning strategies.

Additionally, we observed a small percentage of cases (approximately 1%) where models declined to engage with certain prompts due to safety or ethical considerations, and technical failures (approximately 2%) where models were unable to properly access all image inputs. These findings collectively underscore the multifaceted challenges inherent in open-ended multimodal association tasks, highlighting the need for advancements in visual perception, cross-domain knowledge activation, and reasoning flexibility to achieve more human-like associative capabilities.

Case 1 (Remote-Item Association Task)

Instruction

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

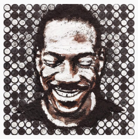


Figure 14: Image 1

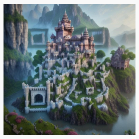


Figure 15: Image 2

Reference Answer

- **Image 1**: A human face made from Oreo cookies
- **Image 2**: A QR code represented by the arrangement of a castle's elements
- **Relation**: Hidden symbols
- **Explanation**: These images use common objects to represent hidden symbols, making Oreo cookies and castle elements serve as symbols for a face and a QR code, respectively.
- **Association Reasoning Path**:
  Symbolize(Oreo, HumanFace) and Is(HumanFace, HiddenSymbols)
  Symbolize(Castle, QRCode) and Is(QRCode, HiddenSymbols)
  Thus, Oreo → HumanFace → HiddenSymbols and Castle → QRCode → HiddenSymbols

LVLM Output

GPT-4o:

- First image: Portrait created with Oreo cookies.
- Second image: Fantasy castle landscape.
- Relation: Creativity.
- Explanation: The first image showcases creativity through an unconventional medium, using Oreo cookies to form a portrait. The second image features a creatively designed fantasy castle, illustrating imaginative architecture and a scenic landscape. Both images highlight artistic expression and inventiveness in their respective forms.

Judge (GPT-4o) Output

- **Holistic Score**: 2
- **Reason**: The response is somewhat relevant but lacks depth. It correctly identifies the creative aspect of the images but fails to recognize the key theme of 'hidden symbols.' The focus on 'creativity' is too broad and does not align with the intended reasoning.

Further Analysis

- **Key Words**: Perceptual Misalignment, Overgeneralization
- **Analysis**: GPT-4o's response exhibits both perceptual misalignment and overgeneralization. The model completely misses the QR code hidden within the castle elements (perceptual misalignment), failing to detect the critical visual pattern that would establish the valid association with the Oreo face. This initial perception failure leads to overgeneralization, where the model retreats to a broadly applicable but imprecise "creativity" association rather than identifying the "hidden symbols" that connects both images. This demonstrates how perception failures lead to reasoning limitations, preventing the model from discovering the more sophisticated, intentional symbolic relationships embedded in the visual content.

Annotation

- **L-3 Perception**: Relational Perception
- **L-3 Conception**: Causal Connections, Thematic Links, Hierarchical Association
- **Relationship Type**: Mutual Element
- **Culture**: N/A
- **Language**: English
- **Topic Domain**: Sense

Case 2 (Remote-Item Association Task)

Instruction

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

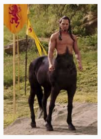
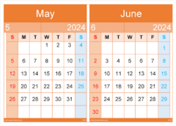


Figure 16: Image 1



May

5 2024

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June

6 2024

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

Figure 17: Image 2

Reference Answer

- **Image 1**: The half human and half horse.
- **Image 2**: May and Jun.
- **Relation**: Sagittarius is opposite to Gemini.
- **Explanation**: Sagittarius is the ninth astrological sign in the Western zodiac, represented by the half human and half horse. Gemini is the third astrological sign in the Western zodiac, represented by the twins. Sagittarius and Gemini are opposite signs in the zodiac.
- **Association Reasoning Path**:
  Symbol(HalfHumanHalfHorse, Sagittarius) and Opposite(Sagittarius, Gemini)
  Period(Gemini, MayJun)
  Thus, HalfHumanHalfHorse → Sagittarius → Gemini → MayJun

LVLM Output

Gemini-1.5-pro:

- First image: A centaur, a mythical creature with the upper body of a human and the lower body of a horse.
- Second image: Calendar for the months of May and June 2024.
- Relation: Unrelated
- Explanation: There's no clear connection between the image of a centaur from mythology and a calendar displaying specific months. They originate from entirely different contexts.

Judge (GPT-4o) Output

- **Holistic Score**: 1
- **Reason**: The LVLM expresses uncertainty and leaves a lack of conceptual connection between the astrological symbols of Sagittarius and the intended zodiac connection.

Further Analysis

- **Key Words**: Knowledge Retrieval Gap, Limited Insight and Excessive Caution
- **Analysis**: This response demonstrates Limited Insight and Excessive Caution in Gemini-1.5-Pro's associative reasoning. The model fails to activate the deeper astrological knowledge connecting these concepts through zodiac relationships, and retreats to the overly cautious conclusion that the images are "unrelated," reflecting a fundamental unwillingness to venture beyond surface-level observations and preventing the model from making the conceptual leap to recognize that the centaur represents Sagittarius.

Annotation

- **L-3 Perception**: Relational Perception
- **L-3 Conception**: Causal Connections, Thematic Links, Hierarchical Association
- **Relationship Type**: Relation
- **Culture**: Non-English European
- **Language**: English
- **Topic Domain**: Myth

Case 3 (Remote-Item Association Task)

Instruction

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

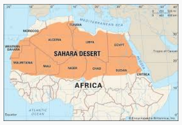


Figure 18: Image 1

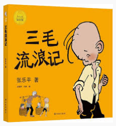


Figure 19: Image 2

LVLM Output

Gemini-1.5-flash:

- First image: Map of the Sahara Desert in Africa
- Second image: Book cover of “Three Hair Travels”
- Relation: A fictional story set in the Sahara Desert
- Explanation: The second image is the book cover of a children’s book, “Three Hair Travels,” which tells a story in the setting of the Sahara Desert. The first image shows the location and size of the Sahara Desert, which is likely the setting for the story. The book’s cover also features imagery that might suggest a desert environment.

Judge (GPT-4o) Output

- **Holistic Score**: 0
- **Reason**: The LVLM includes incorrect facts, referring to an imaginary story, thus leading to a mistaken chain of reasoning.

Reference Answer

- **Image 1**: Sahara map.
- **Image 2**: *Winter Of Three Hairs.*
- **Relation**: Sanmao (in Chinese).
- **Explanation**: San Mao is a Taiwanese writer who wrote *The Stories of the Sahara*. The story is about her life in the Sahara Desert. The Sahara map is related to the story. Sanmao (Three Hairs) is also a cartoon character created by Zhang Leping in 1935. They share the same Chinese name.
- **Association Reasoning Path**:
  Related(The Stories Of Sahara, Sahara Map) and Write By(SanMao, The Stories Of Sahara)
  SameName(SanMao, Sanmao)
  Thus, SaharaMap → TheStoriesOfSahara → SanMao → Sanmao

Further Analysis

- **Key Words**: Knowledge Retrieval Gap
- **Analysis**: This response illustrates a significant Knowledge Retrieval Gap in Gemini-1.5-Flash’s reasoning process. The model fabricates an entirely fictitious connection by claiming *Winter Of Three Hairs* tells a story set in the Sahara Desert—a complete misrepresentation of this iconic Chinese cartoon about a homeless child’s struggles. When directly questioned about *Winter Of Three Hairs* (especially in Chinese), the model fails to activate this knowledge during multimodal association tasks. This disconnect highlights a critical limitation in cross-modal, cross-cultural knowledge retrieval: the model cannot effectively bridge visual perception with cultural knowledge across domains, instead confabulating artificial connections.

Annotation

- **L-3 Perception**: Semantic Object
- **L-3 Conception**: Cultural Reference
- **Relationship Type**: Relation
- **Culture**: East Asia
- **Language**: Chinese
- **Topic Domain**: Art

Case 4 (In-Context Association Task)

Instruction

1. Briefly describe **Image 1**, **Image 2**, and **Image 3** based on their visual information.
2. Analyze the relationship between **Image 1** and **Image 2**, identifying any possible connections, themes, or shared elements that link **Image 1** to **Image 2**.
3. design **Image 4** so that its relationship with **Image 3** mirrors that between **Image 1** and **Image 2**. Use insights from the first pair to guide your design.

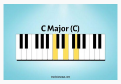


Figure 20: Image 1

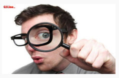


Figure 21: Image 2

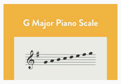


Figure 22: Image 3

Reference Answer

- **Image 1**: C Major.
- **Image 2**: See.
- **Image 3**: G Major.
- **Image 4**: Glee.
- **Relation**: Homophones related to musical notes.
- **Explanation**:
  1. The pronunciation of the letter ‘C’ sounds exactly like the word ‘See’.
  2. The pronunciation of ‘G’ is homophonous with the word ‘Glee,’ which signifies joy or happiness.
- **Association Reasoning Path**:
  1. Homophone(C, See)
     Thus, C → See
  2. Homophone(G, Glee)
     Thus, G → Glee

LVLM Output

GPT-4o:

- Image 1: A piano keyboard illustration highlighting the C Major chord.
- Image 2: A person holding a magnifying glass up to their eye, creating a humorous enlarged eye effect.
- Image 3: Musical notation for the G Major scale.
- Image 4: A humorous image of a magnifying glass enlarging musical notes or sheet music.
- Relation: Emphasis and Clarity
- Explanation: Image 1 emphasizes specific notes clearly within a piano chord, while Image 2 uses a magnifying glass to humorously emphasize and enlarge a part of a person. Image 4 mirrors this by using a magnifying glass to emphasize specific musical notes from Image 3, establishing a connection through the theme of emphasis and clarity in both musical and visual elements.

Judge (GPT-4o) Output

- **Holistic Score**: 2
- **Reason**: The LVLM identified a loose visual connection (emphasis and clarity) rather than the intended homophone relationship. While the explanation is somewhat reasonable, it does not align with the reference answer's wordplay-based reasoning. The connection between Image 3 and Image 4 is vague and does not capture the intended linguistic homophony.

Further Analysis

- **Key Words**: Perceptual Misalignment, Overgeneralization
- **Analysis**: The model fails at the perceptual level by describing Image 2 as "a person holding a magnifying glass up to their eye" without abstracting the critical concept of "See," missing its homophonic link to "C Major." This misalignment leads to an overgeneralized interpretation of the pattern as "Emphasis and Clarity." As a result, instead of recognizing "Glee" as the homophonic pair for "G Major," the model suggests "a magnifying glass enlarging musical notes." This illustrates how poor conceptual abstraction undermines pattern recognition, especially in subtle cross-domain associations like linguistic and musical wordplay.

Annotation

- **L-3 Perception**: Abstract Interpretation, Semantic Object
- **L-3 Conception**: Causal Connections, Analogical Reasoning
- **Relationship Type**: Relation
- **Culture**: N/A
- **Language**: English
- **Topic Domain**: Sense, Music

Case 5 (In-Context Association Task)

Instruction

1. Briefly describe **Image 1**, **Image 2**, and **Image 3** based on their visual information.
2. Analyze the relationship between **Image 1** and **Image 2**, identifying any possible connections, themes, or shared elements that link **Image 1** to **Image 2**.
3. design **Image 4** so that its relationship with **Image 3** mirrors that between **Image 1** and **Image 2**. Use insights from the first pair to guide your design.

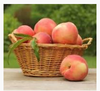


Figure 23: Image 1

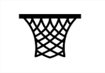


Figure 24: Image 2

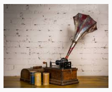


Figure 25: Image 3

Reference Answer

- **Image 1**: A peach basket.
- **Image 2**: A modern basketball hoop.
- **Image 3**: A phonograph.
- **Image 4**: A modern speaker.
- **Relation**: Evolution of Equipment.
- **Explanation**:
  1. Basketball was originally played with a peach basket as a goal, which evolved into the modern basketball hoop.
  2. The phonograph was an early device for playing recorded sound, which evolved into modern speakers.
- **Association Reasoning Path**:
  1. Evolve(PeachBasket, BasketballHoop)
     Thus, PeachBasket → BasketballHoop
  2. Evolve(Phonograph, Speaker)
     Thus, Phonograph → Speaker

LVLM Output

Gemini-1.5-pro:

- First image: Peaches in a basket.
- Second image: Basketball hoop.
- Third image: Vintage gramophone.
- Forth image: Vinyl records in a case.
- Relation: Container and its contents
- Explanation: Image 1 shows peaches held within a basket, representing a container and its contents. Similarly, Image 2 depicts a basketball hoop, which "contains" a basketball during play. Applying this relationship to Image 3, a gramophone, Image 4 would then be vinyl records, as they are the components "contained" and used by the gramophone to produce sound.

Judge (GPT-4o) Output

- **Holistic Score**: 2
- **Reason**: While the LVLM's reasoning follows a 'container and contents' theme, the connection between a peach in a basket and a basketball in a hoop is weak. A basketball hoop does not function as a container in the same way a basket does. This weakens the overall logical consistency of the response, making it somewhat reasonable but lacking a strong foundation.

Further Analysis

- **Key Words**: Overgeneralization, Perceptual Misalignment
- **Analysis**: The model fails to retrieve critical knowledge that peach baskets were the original basketball goals. This leads to a "container and contents" relationship, which makes the connection between a peach in a basket and a basketball in a hoop is weak—a basketball hoop does not function as a container in the same way a basket does. This weakens the logical consistency. This illustrates how early perceptual misalignment can cascade, with superficial pattern matching overriding deeper conceptual understanding in multimodal reasoning.

Annotation

- **Relationship Type**: Relation
- **L-3 Perception**: Relational Perception, Semantic Object
- **L-3 Conception**: Causal Reasoning, Functional Links
- **Culture**: N/A
- **Language**: English
- **Topic Domain**: Sports, Music

D Case Study 2: Success Cases

A close examination of high-performing instances offers invaluable insights into the model's strengths and the specific characteristics of a high-quality response. These cases demonstrate the model's practical effectiveness and serve as a benchmark for its optimal behavior. We report three such high-scoring cases (HR >= 3).

Case 1 (Remote-Item Association Task)

Instruction

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

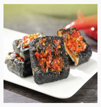


Figure 26: Image 1

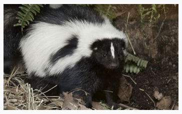


Figure 27: Image 2

Reference Answer

- **Image 1**: Stinky tofu.
- **Image 2**: A skunk.
- **Relation**: Unpleasant smell.
- **Explanation**: Stinky tofu and skunks are both known for their strong, unpleasant smell.
- **Association Reasoning Path**:
  Smell(StinkyTofu, UnpleasantSmell)
  Smell(Skunk, UnpleasantSmell)
  Thus, StinkyTofu → UnpleasantSmell and Skunk → UnpleasantSmell

LVLM Output

GPT-4o:

- First image: Spicy stinky tofu
- Second image: Skunk in natural habitat
- Relation: Odor
- Explanation: Both images feature subjects known for their strong, distinctive smells. Stinky tofu is a fermented dish famous for its pungent aroma, while skunks are notorious for their powerful defensive spray. Despite their different contexts, both are associated with unique and often polarizing scents.

Judge (GPT-4o) Output

- **Holistic Score**: 4
- **Reason**: The MLLM correctly identifies the relation as ‘odor’ and explains it in terms of strong, distinctive smells, consistent with the reference answer’s focus on ‘unpleasant smell.’

Further Analysis

- **Analysis**: This case reflects a relatively straightforward instance of associative reasoning. The model correctly identifies each image and retrieves the salient fact that both are known for their distinctive smells, which are frequently emphasized in cultural and zoological contexts. It highlights the model’s strength in capturing explicit thematic links from its training data.

Annotation

- **L-3 Perception**: Semantic Object
- **L-3 Conception**: Thematic Links, Cultural Reference
- **Relationship Type**: Mutual Elements
- **Culture**: Chinese
- **Language**: English
- **Topic Domain**: Sense

Case 2 (Remote-Item Association Task)

Instruction

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

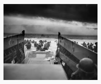


Figure 28: Image 1

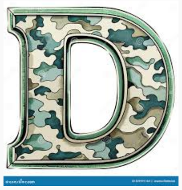


Figure 29: Image 2

Reference Answer

- **Image 1**: Omaha Beach.
- **Image 2**: Letter D.
- **Relation**: D-Day.
- **Explanation**: Omaha Beach, as before, represents one of the principal landing beaches of the D-Day invasion. The letter D symbolizes the day of the invasion, emphasizing the significance of the operation and the turning point it marked in World War II.
- **Association Reasoning Path**:
  LandingSite(OmahaBeach, DDay)
  Symbol(LetterD, DDay)
  Thus, OmahaBeach → DDay and LetterD → DDay

LVLM Output

Gemini-1.5-Pro:

- First image: Soldiers landing on a beach
- Second image: Letter D in camouflage pattern
- Relation: D-Day
- Explanation: The first image depicts the historical event of D-Day, the Allied invasion of Normandy during World War II. The second image shows the letter "D" in a camouflage pattern, symbolizing the military operation and its association with the letter "D" in "D-Day."

Judge (GPT-4o) Output

- **Holistic Score**: 4
- **Reason**: The MLLM output is accurate and logically consistent with the reference answer. It correctly identifies the relation as 'D-Day' and provides an explanation that mirrors the standard reasoning by associating the images with the historical event and its symbolism.

Further Analysis

- **Analysis**: This case requires a more sophisticated form of reasoning, moving beyond shared physical attributes to connect a historical event with its symbolic representation. The model must not only recognize the historical context of the first image (the D-Day landings) but also interpret the second image as a symbolic component of the event's name. The camouflage pattern on the letter 'D' is a crucial cue that reinforces the military theme. This success highlights the model's capacity to bridge concrete visual representations with abstract symbolic knowledge, demonstrating a strong integration of visual analysis with specific historical world knowledge.

Annotation

- **L-3 Perception**: Contextual Sensory Cues
- **L-3 Conception**: Cultural Reference
- **Relationship Type**: Mutual Elements
- **Culture**: USA/English
- **Language**: English
- **Topic Domain**: History

Case 3 (In-Context Association Task)

Instruction

1. Briefly describe **Image 1**, **Image 2**, and **Image 3** based on their visual information.
2. Analyze the relationship between **Image 1** and **Image 2**, identifying any possible connections, themes, or shared elements that link **Image 1** to **Image 2**.
3. design **Image 4** so that its relationship with **Image 3** mirrors that between **Image 1** and **Image 2**. Use insights from the first pair to guide your design.

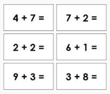


Figure 30: Image 1

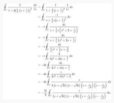


Figure 31: Image 2

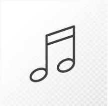


Figure 32: Image 3

Reference Answer

- **Image 1**: Basic addition equations.
- **Image 2**: An integral expression.
- **Image 3**: A musical note.
- **Image 4**: A symphony score.
- **Relation**: From Simple to Complex.
- **Explanation**:
  1. Basic addition is a simple, foundational operation, while integration represents a higher-level, complex concept in calculus that builds upon it.
  2. A single note is a basic musical element, while a symphony score is a complex composition made up of many such elements.
- **Association Reasoning Path**:
  1. Addition $\wedge$ Integration $\rightarrow$ From Simple to Complex
  2. MusicalNote $\wedge$ SymphonyScore $\rightarrow$ From Simple to Complex

LVLM Output

GPT-4o:

- First image: Simple math addition problems.
- Second image: A complex mathematical integration problem with detailed steps.
- Third image: A basic musical note symbol.
- Forth image: A complex sheet of musical notes, possibly including a variety of symbols, rhythms, and notations.
- Relation: Complexity progression
- Explanation: The relationship between the first pair of images is based on the complexity of mathematical problems. Image 1 shows basic arithmetic while Image 2 depicts a complex integration problem. Applying this relationship to the second pair, Image 3 shows a simple musical note, so Image 4 should represent a more complex musical composition with a detailed arrangement.

Judge (GPT-4o) Output

- **Holistic Score**: 4
- **Reason**: The MLLM's response correctly identifies the complexity progression relationship between the images and suggests an appropriate complex sheet of musical notes for Image 4. The explanation logically extends the complexity theme from simple math to complex math and from a simple musical note to a complex sheet, aligning well with the reference answer.

Further Analysis

- **Analysis**: This case showcases a clear instance of cross-domain associative reasoning based on structural progression. The model demonstrates accurate perception, appropriate domain knowledge, and effective abstraction of the "complexity progression" theme. While the reasoning is not particularly challenging due to the explicit visual and semantic cues, it highlights the model's ability to generalize a structural concept across distinct domains.

Annotation

- **Relationship Type**: Relation
- **L-3 Perception**: Abstract Interpretation
- **L-3 Conception**: Hierarchical Association, Analogical Reasoning
- **Culture**: N/A
- **Language**: English
- **Topic Domain**: STEM, Music

E Testing and Evaluation Prompts

We report our prompts for testing and evaluation.

E.1 Testing Prompts

Our testing prompt comprises two principal components. The first component constitutes the instruction for our tasks, and the second delineates the output format for the LVLM. The overall organizational structure is as follows:

Prompt: `<Instruction>` + `<Format>`

The specific implementation of our prompt structure is presented below.

Testing Prompt for RIA Task

Describe each image briefly.
Analyze and explore the relation between the two images, identifying any possible connections, themes, or shared elements.
Formulate the output as follows:
- First image: [image concept]
- Second image: [image concept]
- Relation: [one keyword, phrase or sentence]
- Explanation: [1-5 sentences]

Testing Prompt for ICA Task

1. Briefly describe **Image 1**, **Image 2**, and **Image 3** based on their visual information.
2. Analyze the relationship between **Image 1** and **Image 2**, identifying any possible connections, themes, or shared elements that link **Image 1** to **Image 2**.
3. Design **Image 4** so that its relationship with **Image 3** mirrors that between **Image 1** and **Image 2**. Use insights from the first pair to guide your design.
4. Present your analysis in this format:

- **Image 1**: [image concept]
- **Image 2**: [image concept]
- **Image 3**: [image concept]
- **Image 4**: [image concept that you design]
- **Relation**: [a keyword, phrase, or sentence describing the connection]
- **Explanation**: [1-5 sentences detailing the reasoning and its application to both pairs]

E.2 Prompts for Regular LLM-as-a-Judge Scoring

Prompts for Regular LLM-as-a-Judge Scoring comprises three principal components. The first component constitutes the scoring rules, which provide the LLM judge with a five-level scoring gradient (0–4) for evaluating the quality of LVLM responses. The second component delineates the input/output format; in this section, we supply the LLM judge with input references and constrain its output format, thereby enhancing the standardization of the information flow. The third component consists of exemplars. Here, we employ a Few-Shot approach to furnish the LLM judge with concrete examples of scoring criteria, effectively mitigating the high scoring redundancy associated with One-Shot approaches and enhancing scoring diversity, while simultaneously reinforcing the judge's accurate comprehension of the evaluation standards. The overall organizational structure is as follows:

Prompt: `<Scoring Rules>` + `<I/O Format>` + `<Rating Examples>`

The specific implementation of the prompt structure is presented below.

Scoring Rules for RIA Task

Evaluate the quality of responses provided by LVLMs using the following criteria:

- **4 points**: Accurate and logically consistent. Convey the same level of thoughtfulness and insight, the same caliber of intellectual rigor and comprehension with reference answer. (Not required to be literally identical to reference answers)
- **3 points**: Shows reasonable understanding but may be incomplete or lack key insights, such as mentioning key facts but failing to effectively summarize the relationship. However, what can be directly got from the images or a common theme with little similarity with the reference answer are not key facts.

- **2 points**: Somewhat relevant but lacks depth, is overly broad, or misses critical elements needed for a coherent explanation; does not address or align with the intended standard reasoning at all.
- **1 point**: Displays uncertainty, offering vague or incomplete reasoning, or concedes with phrases like "I don't know."
- **0 points**: Factuality hallucination in the explanation, i.e. contains factual inaccuracies or fabricated information that undermines its reasoning.

Rating Examples for RIA Task

Rating examples:
Example 1
LVLM Output:[
- First image: Chicago skyline with Willis Tower and glass buildings.
- Second image: Cloud Gate ("The Bean") sculpture in Millennium Park.
- Relation: Chicago landmarks
- Explanation: Both are iconic Chicago structures - the Willis Tower defining the skyline and Cloud Gate serving as a renowned public sculpture, representing the city's architectural heritage.
]
Reference Answer: [
"relation": "Chicago",
"explanation": "Both Skyscraper and Bean are located in Chicago. Skyscrapers are a prominent feature of Chicago's skyline, and Cloud Gate ("The Bean") is a famous public sculpture located in Millennium Park, Chicago.",
"reasoning": "LocatedIn(Skyscraper, Chicago)
LocatedIn(Bean, Chicago)
Thus, Skyscraper → Chicago and Bean → Chicago"
]
{
"score_judge": "4",
"score_reason": "The LVLM's output captures the essence of both images as notable Chicago landmarks, consistent with the reference answer, although its summary 'Chicago landmarks' differs from the reference answer 'Chicago.''
}

[4 more examples of different scores omitted for brevity]

Scoring Rules for ICA Task

Background: LVLMs analyze three images, identify relationships between Images 1 and 2, and propose Image 4 that maintains the same relationship with Image 3.
Your task: Evaluate LVLM responses focusing on Image 4 description, Relation, and Explanation against reference answers using these criteria:

- **4 points**: Accurate and logically consistent. Correctly matches reference answer's logic (not necessarily identical). Accurate relation identification and logical Image 4 description.
- **3 points**: Generally correct but incomplete. Shows reasonable understanding but may be incomplete or lack key insights, such as mentioning key facts but failing to effectively summarize the relationship.
- **2 points**: Somewhat reasonable but lacks depth, is overly broad, or misses critical elements needed for a coherent explanation; does not address or align with the intended standard reasoning at all; completely unrelated to the reference answer's focus; completely misses the thematic connection; does not address the intended

theme; a completely different interpretation. Vague relation or loose connection between image 3 and 4.

- **1 point**: Responds with “I don’t know,” “I can’t help,” “Unknown connection,” etc., or expresses its uncertainty cautiously.
- **0 points**: Contains factual errors or illogical content. Factuality hallucination in the explanation, i.e. contains factual inaccuracies or fabricated information that undermines its reasoning. (Note that hallucination is only related to the LVLM’s explanation. That is, even if LVLM’s answer does not address or align with the intended standard reasoning at all, it must not get 0 points unless it refers to erroneous information. Divergent opinions on the image theme do not lead to factual hallucination, but wrong image content does.)

Rating Examples for ICA Task

Rating examples:
Example 1
LVLM Output:[
- Image 1: film Amélie
- Image 2: Eiffel Tower
- Image 3: From Vegas to Macau II
- Image 4: Macau Tower
- Relation: Setting/Location
- Explanation: Image 1 (Amélie) is closely tied to Paris, France, where the film was set and shot. Image 2 shows the Eiffel Tower, a symbol of Paris. Likewise, Image 3 (From Vegas to Macau II) highlights Macau in both its title and visuals. Thus, Image 4 should feature a notable Macau landmark, such as the Macau Tower.

]
Reference Answer: [
“Image 4”: “the new Lisboa Hotel”, “relation”: “Films Associated with Iconic Locations”,
“explanation”: “Amélie is a film that captures the essence of Paris, showcasing its charm, streets, and culture. Similarly, From Vegas to Macau is centered around the gambling and nightlife culture in Macau, with the New Lisboa Hotel being a significant landmark in that context.”,
“reasoning path 1”: “FilmSetting(Amelie, Paris)
CulturalSymbol(EiffelTower, Paris)
Thus, Amelie → Paris”
“reasoning path 2”: “FilmSetting(MacauStorm, Macau)
CulturalSymbol(NewLisboaHotel, Macau)
Thus, MacauStorm → NewLisboaHotel”
]
{
“score_judge”: “4”,
“score_reason”: “The LVLM accurately linked ‘Amélie’ with Paris via the Eiffel Tower and ‘From Vegas to Macau II’ with Macau via the Macau Tower. Although the reference answer highlighted the New Lisboa Hotel for Macau, the Macau Tower is also a valid and recognizable symbol. The LVLM’s explanation was clear and logical, with no errors.”
}

[4 more examples of different reasoning omitted for brevity]

I/O Format for Both RIA and ICA Tasks

You will receive multiple independent questions in a numbered format:

```
1. LVLM Output: []
   Reference Answer: []
```

```
2. LVLM Output: []
   Reference Answer: []
```

Provide your response in JSON format where each key is the question number and the value is your answer:

```
{
    "1": {"score_judge": "", "score_reason": ""},
    "2": {"score_judge": "", "score_reason": ""}
}
```

E.3 Prompts for LLM Judging in MM-OPERA Reasoning

Our prompt implementation adopts a cross-structured architectural framework and comprises four principal components. The first component establishes the evaluative role, instituting the foundational operational parameters for the LLM judge. The second component formalizes the assessment methodology by constructing a cross-structured prompt that simultaneously provides the LLM judge with both the evaluative task specifications and output format requirements, effectively optimizing the prompt structure and enhancing the consistency of intentional conveyance within the linguistic framework. The third component comprises detailed annotations and format delineations, enabling the LLM judge to integrate task-specific analytical elements while further reinforcing the input-output structural protocol. The fourth component presents calibrated exemplars through a Few-Shot approach to further elucidate the assessment criteria and standardize evaluation procedures. The comprehensive organizational structure of our prompt is as follows:

Prompt: `<Role Definition>` + `<Cross-Structured Instructions>` + `<Annotative Framework and I/O Protocol>` + `<Rating Examples>`

The specific implementation of our prompt structure is presented below. And the sections marked with ellipses share a similar structure and content with the surrounding context and are therefore omitted for brevity.

Role Definition, Cross-Structured Instructions, Annotative Framework and I/O Protocol for RIA Reasoning-guided Evaluation Task

You are an expert judge evaluating association paths between two image concepts. Your task:
1. Analysis Input

```
input:{
    concepts: [ImageA_desc, ImageB_desc],
    reference_answer: {
    relation: string,
    explanation: string,
    path: string
}
lvlm_output: {
    description: [ImageA_desc, ImageB_desc],
    relation: string,
    explanation: string
    }
}
```

2. Path Standardization for lvlm_output

```
rules:
   {
     type1_sequential: "Predicate_{11}(A, X_{11}) and ...
     and Predicate_{1i}(X_{1i}, X)
     and Predicate_{21}(X, X_{21}) and ...
```

```
  and Predicate_{2j}(X_{2j}, B)\n
  A → X_{11} → ... → X_{1i} → X
  → X_{21} → ... → X_{2j} → B"
  type2_convergent: "Predicate_{11}(A, X_{11}) and ...
  and Predicate_{1i}(X_{1i}, X)\n
  Predicate_{21}(B, X_{21}) and ... and Predicate_{2j}
  (X_{2j}, X)\n
  A → X_{11} → ... X_{1i} → X and B → X_{21} → ... → X_{2j} → X"
  type3_metaphorical: "A $\land$ B → X"

  format: {
    - PascalCase for entities/predicates
    - ’and’ for clause connection
    - ’$\land$’ for entity connection
    - ’→’ for each association hop
  }
}
```

3. Quality Assessment

```
for each hop in path:
  score: [
    float[0-1],
    // Reasonableness: logical soundness to the ideal relation X
    float[0-1],
    // Distinctiveness: The clarity of concept boundaries
    int[0 or 1]
    // Knowledgeability: Domain knowledge depth
  ]
```

Note: The path don’t need to be identical to the reference answer but should demonstrate equivalent quality of reasoning.
You will receive multiple independent problem-answer pairs in a numbered format:
1. Problem: ...
2. Problem: ...
Provide your response in JSON format where each key is the problem-answer pair number and the value is your evaluation:

`{"1": {...},"2": {...}}`

Rating Examples for RIA Reasoning-guided Evaluation Task

Examples 1:
Problem:
- First image: Bread and Wine
- Second image: 12

Reference Answer:
- Relation: The Last Supper
- Explanation: The Last Supper is a religious painting depicting the final meal Jesus shared with his disciples before his crucifixion. The bread and wine represent the elements of the Eucharist, a central sacrament in Christianity. The number 12 refers to the twelve apostles present at the Last Supper.
- Association Path: Depict(TheLastSupper, Eucharist) and Represent(BreadAndWine, Eucharist) Number(TheLastSupper, 12)

Thus, BreadAndWine $\rightarrow$ Eucharist $\rightarrow$ TheLastSupper and 12 $\rightarrow$ TheLastSupper

LVLM’s Output:
- First image: A loaf of bread and a chalice
- Second image: The number 12 in gold
- Relation: Christian symbolism

- Explanation: The first image is a visual representation of the Eucharist, a central sacrament in Christianity, where bread and wine are consecrated and consumed as the body and blood of Jesus Christ. The number 12 holds significant symbolism in Christianity, representing the 12 apostles of Jesus. Together, these images allude to a key aspect of Christian faith and tradition.

Your evaluation:
{
“path”: “Represent(BreadAndWine, Eucharist) and Symbolize(Eucharist, Christian)
Number(12, ApostlesOfJesus) and Related(ApostlesOfJesus, Christian)
Thus, BreadAndWine → Eucharist → Christian and 12 → ApostlesOfJesus → Christian”,
“hop_quality”: {
“BreadAndWine → Eucharist”: [1.00, 1.00, 1],
“Eucharist → Christian”: [1.00, 1.00, 1],
“12 → ApostlesOfJesus”: [1.00, 1.00, 1],
“ApostlesOfJesus → Christian”: [1.00, 1.00, 1]
}
“explanation”: “The LVLM constructed a well-structured convergent path through Christian symbolism. Both paths (BreadAndWine→Eucharist→Christian and 12→ApostlesOfJesus→Christian) demonstrate perfect scores across all dimensions, showing deep theological understanding and precise use of religious concepts.”
}

[2 more examples of different scores omitted for brevity]

Role Definition, Cross-Structured Instructions, Annotative Framework and I/O Protocol for ICA Reasoning-guided Evaluation Task

LVLMs analyze three images, identify relationships between Images 1 and 2 (Pair 1), and propose Image 4 that maintains the same relationship with Image 3 (Pair 2).
You are an expert judge evaluating LVLM’s association path. Your task:
1. Analysis Input

```
input: {
 concepts: [Image1_desc, Image2_desc, Image3_desc],
 reference_answer: {
   image4: Image4_desc
   relation: string,
   explanation: string,
   path: string
 },
 lvlm_output: {
   description: [Image1_desc, Image2_desc, Image3_desc, Image4_desc],
   relation: string,
   explanation: string
 }
}
```

2. Path Standardization for lvlm_output

```
   rules:
   {
     type1_sequential: "Predicate_{11}(A, X_{11}) and ...
     and Predicate_{1i}(X_{1i}, X) and Predicate_{21}(X, X_{21})
     and ...
     and Predicate_{2j}(X_{2j}, B)\n
     A → X_{11} → ... → X_{1i} → X → X_{21} → ... → X_{2j} → B"
     type2_convergent: "Predicate_{11}(A, X_{11}) and ...
     and Predicate_{1i}(X_{1i}, X)\n
     Predicate_{21}(B, X_{21}) and ...
```

```
    and Predicate_{2j}(X_{2j}, X)\n
    A → X_{11} → ... X_{1i} → X and B → X_{21} → ... → X_{2j} → X"
    type3_metaphorical: "A $\land$ B → X"
    format: {
      - PascalCase for entities/predicates
      - 'and' for clause connection
      - '$\land$' for entity connection
      - '→' for each association hop
    }
  }
```

3. Quality Assessment for lvlm_output

```
  for each hop in path:
    score: [
      float[0-1],
      // Reasonableness: logical soundness to the ideal relation X
      float[0-1],
      // Distinctiveness: The clarity of concept boundaries
      int[0 or 1]
      // Knowledgeability: Domain knowledge depth
    ]
```

Note: The path don't need to be identical to the reference answer but should demonstrate equivalent quality of reasoning. If the lvlm_output is irrelevant or incorrect, provide a low score for each hop.
You will receive multiple independent problem-answer pairs in a numbered format:
1. Problem: ...
2. Problem: ...
Provide your response in JSON format where each key is the problem-answer pair number and the value is your evaluation:

```
{"1": {...},"2": {...}}
```

Rating Examples for ICA Reasoning-guided Evaluation Task

Examples 1:
Problem:
- Image 1: Burj Khalifa
- Image 2: Canton Tower
- Image 3: Statue of Liberty

Reference Answer:
- Image 4: Eiffel Tower
- Relation: Built in the Same Year
- Explanation 1: The Burj Khalifa in Dubai and Canton Tower in Guangzhou were both completed in 2010, marking a significant year for architectural achievements.
- Association Path 1: BuiltIn(BurjKhalifa,2010)

BuiltIn(CantonTower,2010)
Thus, BurjKhalifa → 2010 and CantonTower → 2010
- Explanation 2: The Statue of Liberty and the Eiffel Tower were both inaugurated in 1886, symbolizing important cultural landmarks in their respective countries.
- Association Path 2: BuiltIn(StatueOfLiberty,1886)

BuiltIn(EiffelTower, 2010)
Thus, StatueOfLiberty → 1886 and EiffelTower → 1886
LVLM's Output:
- **Image 1**: The Burj Khalifa, the tallest building in the world, located in Dubai, United Arab Emirates.
- **Image 2**: Canton Tower, also the tallest tower in China, located in Guangzhou,

Guangdong Province, China.
- **Image 3**: The Statue of Liberty, a colossal neoclassical sculpture on Liberty Island in New York Harbor, New York City.
- **Image 4**: The Eiffel Tower, a wrought iron lattice tower on the Champ de Mars in Paris, France.
- **Relation**: "Tallest Structures"
- **Explanation**: All four images depict some of the tallest structures in the world, each representing a different architectural style and cultural significance. Image 1 and Image 2 are skyscrapers, while Image 3 is a statue, and Image 4 is an iconic tower. This common theme allows for a comparison of how these structures have become symbols of their respective cities and countries.

Your evaluation:
{
"path1": "Is(BurjKhalifa, Tallest)
Is(CantonTower, Tallest)
BurjKhalifa → Tallest and CantonTower → Tallest", "path2": "Is(StatueOfLiberty, Tallest)
Is(EiffelTower, Tallest)
StatueOfLiberty → Tallest and EiffelTower → Tallest",
"hop_quality_path1": {
"BurjKhalifa → Tallest": [0.95, 0.86, 1],
"CantonTower → Tallest": [0.95, 0.85, 1]
},
"hop_quality_path2": {
"StatueOfLiberty → Tallest": [0.55, 0.45, 1],
"EiffelTower → Tallest": [0.83, 0.85, 1]
},
"explanation": "In the LVLM's output, the first path shows consistently high hop quality scores as both the Burj Khalifa and Canton Tower are indeed among the tallest structures. The second path shows more varied scores, with the Statue of Liberty receiving lower scores as it's not typically categorized among the world's tallest structures, while the Eiffel Tower maintains high scores."
}

[1 more examples of different reasoning omitted for brevity]

F Limitations and Broader Impacts

F.1 Limitation

While MM-OPERA represents a significant advancement in evaluating association reasoning in Large Vision-Language Models, several limitations highlight areas for future refinement.

- **Limited Exploration of Temporal Association Reasoning**: MM-OPERA's static task design (RIA and ICA) does not fully capture temporal or sequential association reasoning, a key aspect of human cognition in dynamic contexts like decision-making, restricting its evaluation scope.
- **High Cost and Scalability Challenges for Open-Ended Evaluation**: Evaluating 11,497 open-ended tasks with a resource-intensive LLM-as-a-Judge and cascading scoring rubric incurs high computational costs (due to increased token usage) and limits scalability, hindering rapid or large-scale testing of LVLMs.
- **Challenges in Systematic Task Creation**: Although association is common in human cognition, systematically collecting and converting ideas or existing data into task instances is challenging, especially given LLMs' weaknesses in this area, leading to high human effort costs for data expansion.

These limitations underscore the need for continued innovation to enhance MM-OPERA's robustness, scalability, and applicability in advancing AI research.

F.2 Broader Impacts

MM-OPERA is an **evaluation** benchmark for associative reasoning in Large Vision-Language Models (LVLMs), not a training set. It aims to deepen understanding and guide AI development. However, evaluation standards carry societal implications.

Potential Societal Considerations:

- **Guiding Development and Bias Risks:** Benchmarks shape research. Any unaddressed gaps or subtle biases within MM-OPERA could inadvertently steer development towards a narrow or skewed form of associative intelligence, impacting real-world fairness and applicability.
- **Perception of Capabilities and Misuse Potential:** By identifying models with advanced associative abilities, MM-OPERA may elevate perceptions of their power. Such identified capacities, even if not developed via this benchmark, could be leveraged for sophisticated misuse (e.g., disinformation) if not responsibly managed.
- **Deployment Risks from Identified Limitations:** MM-OPERA reveals model weaknesses. Overlooking these identified limitations during deployment in critical systems could lead to erroneous and harmful outcomes.

Mitigation and Responsible Use of Insights: Transparency in design and responsible interpretation of MM-OPERA's results are crucial. We advocate for:

- Continuous community scrutiny and refinement of MM-OPERA to address potential biases and representational gaps.
- Using evaluation insights to understand fundamental AI limitations and guide research towards robust, safe, and aligned systems, beyond mere model ranking.
- Informed deployment decisions by developers and deployers, using benchmarked strengths and weaknesses to assess suitability and mitigate risks.

Our goal is for MM-OPERA to foster rigorous evaluation, contributing to more capable and societally beneficial AI.