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

Anonymous Author(s)

Affiliation

Address

email

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.

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 [62], Gemini models [72], and LLaVA [50] exhibit remarkable proficiency in visual understanding, language generation, and multi-step reasoning. These capabilities are driving transformative applications across fields like education, design, scientific discovery, and embodied intelligence.

Existing benchmarks for LVLMs [3, 55, 31, 87, 39, 38, 88, 17, 27, 53] 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. **Association**, a cornerstone of human cognition, enables creative thinking [57], underpins the integration of fragmented information into coherent knowledge and supports critical cognitive processes such as memory, perception, and rule discovery [5].

Current efforts, such as the Labyrinth of Links [41] 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

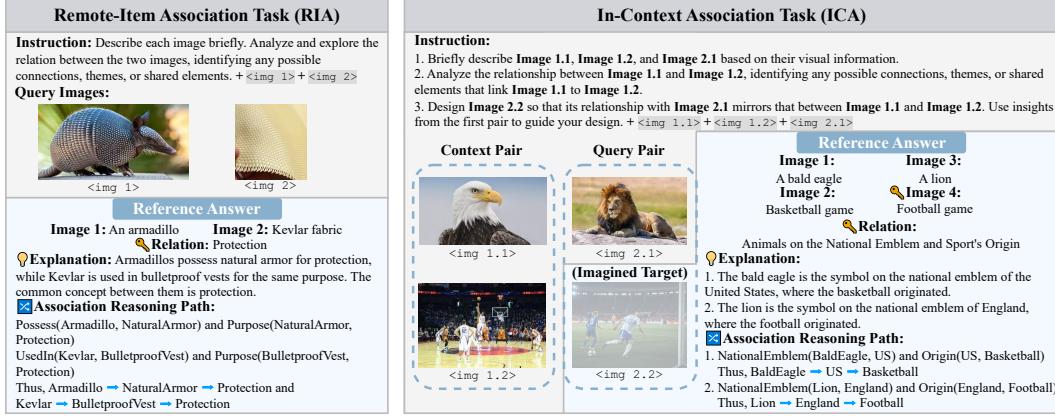


Figure 1: An overview of MM-OPERA. The RIA task challenges models to discover meaningful connections between unrelated elements, while the ICA task requires transferring relationship patterns from a context pair to a query item to generate an appropriate target. The reference answer represents just one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

37 guiding the model’s associative behavior and masking its true capacity for independent reasoning; (2)
38 The fixed-answer format struggles to evaluate complex, long-form association reasoning, limiting the
39 ability to challenge models on intricate, multi-step relational inference. These limitations motivate our
40 development of a new benchmark that prioritizes **open-endedness** to rigorously assess and ultimately
41 enhance LVLMs’ association reasoning capabilities.

42 In cognitive science, association emerges from the interplay of *convergent and divergent thinking*:
43 the former identifying meaningful connections and selecting optimal solutions; the latter generating
44 multiple unique ideas [75, 12, 56]. The Remote Associates Test (RAT) [57, 21, 16, 79, 2] exemplifies
45 this by requiring individuals to uncover links between distant concepts, a process vital for adaptive
46 problem-solving. To mirror this in LVLMs and address the shortcomings of prior work, we propose
47 **MM-OPERA** (**M**ulti-**M**odal **O**Pen-**E**nded **R**easoning-guided **A**sso*c*iation), a benchmark designed
48 to evaluate association reasoning without predefined constraints. It assesses how models identify
49 and express meaningful links across distant concepts (*i.e.* convergent thinking), expected to emerge
50 through diverse reasoning paths (*i.e.* divergent thinking). Table 1 highlights how MM-OPERA
51 diverges from The Labyrinth of Links by adopting open-ended tasks, more challenging reasoning
52 scenarios, and a broader scope of evaluation, enabling a deeper probe into LVLMs’ relational inference
53 abilities.

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

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

54 MM-OPERA comprises 11,497 instances across two core tasks (Figure 1): *Remote-Item Association*
55 (**RIA**), testing the ability to link distant concepts with structured reasoning, and *In-Context Associa*c*tion* (**ICA**), probing pattern recognition within in-context learning [26]. Spanning 13 associative
56 dimensions and diverse cultural, linguistic, and thematic contexts, it offers a comprehensive evalua*c*tion
57 framework. It prioritizes free-form responses, employing reference answers as heuristic quality
58 benchmarks rather than rigid correctness criteria. To evaluate open-ended outputs, we design tailored
59 LLM-as-a-Judge strategies with a cascading scoring rubric. Furthermore, by leveraging process-
60 reward principles to trace reasoning steps, our evaluation captures cognitive flow and knowledge
61 integration, surpassing traditional outcome-focused metrics.

63 Our contributions are threefold:

- 64 1. **MM-OPERA:** We introduce a benchmark of 10,000+ instances for evaluating LVLMs'
65 association reasoning, centered on Remote-Item Association (RIA) and In-Context Associa-
66 tion (ICA) tasks inspired by classic psychometric studies. It spans 13 analytical dimensions
67 to enable comprehensive assessment.
- 68 2. **LLM-as-a-Judge Strategies:** To support open-ended evaluation, we design tailored LLM-
69 as-a-Judge methods that assess both response quality and reasoning processes, enabling
70 fine-grained and reliable scoring.
- 71 3. **Profound Findings:** Our analysis reveals key limitations of current LVLMs and highlights
72 the critical role of association reasoning in advancing real-world, general-purpose AI.

73 2 Related Work

74 **Large Vision Language Models (LVLMs).** Early studies [71, 90, 66] established the foundations
75 of vision-language models. CoCa [86], Flamingo [1], and BLIP-2 [42], advanced performance with
76 enhanced architectures and large-scale multimodal pretraining. InstructBLIP [22], MiniGPT-4 [95],
77 and LLaVA [51], have refined multimodal instruction tuning and alignment strategies. Recent open-
78 source LVLMs, e.g., LLaVA-OneVision [40], mPLUG-Owl3 [83], and Qwen2-VL [7], have extended
79 these capabilities to multi-image and video understanding. Proprietary models like GPT-4V [62],
80 Gemini-Pro-V [72], and Qwen-VL-Max [8] have demonstrated state-of-the-art performance.

81 **LVLM Benchmarks.** The evaluation of LVLM has progressed from early benchmarks like
82 VQA [3, 31] and OK-VQA [55] to broader assessments such as SEED-Bench [39], LAMM [84],
83 LVLM-eHub [82], MMBench [53], MSCOCO [48], and MM-Vet [87], covering tasks like
84 Optical Character Recognition (OCR) [52], adversarial robustness [92], and hallucination de-
85 tection [20, 49, 43, 76]. Specialized benchmarks target various capabilities: MathVista [54],
86 CLEVR [35], CVR [89], ReMI [37], Encyclopedic VQA [58], LogicVista [81], SPACE [67],
87 BLINK [28], ZeroBench [68], MMMU [88], and Visual Riddles [14] each focus on different aspects
88 of reasoning and perception. Li et al. [41] propose an adjective-verb association benchmark, but it is
89 constrained to predefined categories, leaving open-ended associative reasoning largely unexplored.

90 **Psychometric Test for AI Evaluation.** Researchers have proposed psychometric frameworks to
91 assess AI cognition [34], ranging from personality and theory-of-mind benchmarks [44], latent
92 trait profiling [65], and reasoning evaluation via the Technology Acceptance Model (TAM) [45], to
93 broader construct-oriented approaches emphasizing underlying cognitive mechanisms over task-level
94 performance [78]. Adaptive testing further enhances efficiency by dynamically adjusting to model
95 responses [96]. Association reasoning has also been modeled and involved in AI evaluation [69, 41].
96 Mednick's Theory of Creativity defines creativity as forming remote connections [57], underpinning
97 associative creativity theories [64, 9] and the Remote Associates Test (RAT), adapted for semantic and
98 visual associations [15, 61, 10, 60] through convergent thinking tasks. Divergent thinking [73, 13]
99 is also assessed via tasks like the Alternate Uses Task (AUT) [32] and Divergent Association Task
100 (DAT) [59]. Studies on LLMs and LVLMs reveal mixed results: “leap-of-thought” tasks enhance
101 divergent reasoning [94], GPT models show varied creativity and even surpass humans [36, 19].

102 3 MM-OPERA: Dataset

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

106 3.1 Association Tasks: Motivation and Definition

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

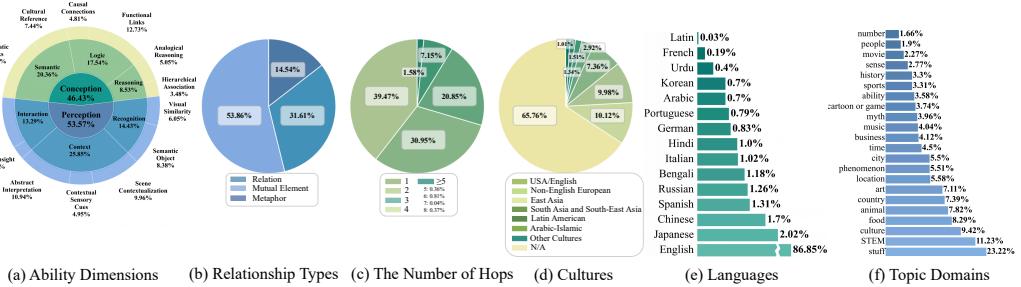


Figure 2: Statistics of MM-OPERA. (a) Hierarchical ability taxonomy consists of 3 levels, refining perceptual and conceptual associations. We report each ability’s frequency as a percentage of total label occurrences to better represent the dataset’s distribution. (b) Three relationship types capturing 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.

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

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

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

131 3.2 Dataset Statistics

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

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

142 **Hierarchical Ability Taxonomy.** Associative thinking operates on multiple cognitive levels, from
 143 perception to conception—both crucial for understanding complex environments [80, 70, 33, 11].
 144 Perception handles immediate sensory inputs, while conception deals with abstract, knowledge-driven
 145 associations. These fundamental processes form our Level-1 (*L-1*) associative ability. We further
 146 refine it into six *L-2* and thirteen *L-3* dimensions, creating a hierarchical framework that mirrors
 147 human cognition and enables systematic evaluation of LVLMs’ capabilities in processing both sensory
 148 input and abstract reasoning. Detailed definitions are in Appendix A.2.

149 **Types of Relationship.** To capture the nuanced ways entities or concepts connect, we identify three
 150 relationship categories: *Relation*, denoting general links between entities; *Mutual Element*, indicating
 151 shared traits; and *Metaphor*, connecting entities through abstract or symbolic meanings. This tripartite

152 classification enhances the benchmark’s ability to evaluate associative reasoning across both literal
153 and abstract dimensions, reflecting the multifaceted nature of human associative thinking.

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

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

167 4 LLM-as-a-Judge Strategies for MM-OPERA

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

175 4.1 Regular LLM-as-a-Judge Scoring

176 Since the open-ended responses and references answers presented in text, we adopt LVLMs as automatic
177 judge engine. Unlike prior benchmarks that use per-sample criteria [29], we adopt unified scoring
178 rubrics that evaluate the association quality of responses—prioritizing depth, coherence, and insight
179 over mere correctness. With regards to open-ended responses with multiple valid potential answers,
180 our regular judge engine assess the internal consistency and reasoning quality by the cascading
181 scoring rubric:

- 182 • **4 points:** Accurate, logically consistent, and insightful, matching the reference answer’s
183 intellectual rigor.
- 184 • **3 points:** Shows reasonable understanding but lacks key insights or completeness.
- 185 • **2 points:** Somewhat relevant but lacks depth, is overly broad, or omits critical reasoning.
- 186 • **1 point:** Vague, uncertain, or incomplete, failing to provide meaningful reasoning.
- 187 • **0 points:** Contains factual errors or fabrications that undermine validity.

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

196 4.2 Process-Reward LLM-as-a-Judge Scoring

197 The regular scoring rule in Section 4.1 is outcome-based and fail to distinguish and analyze models
198 that produce similar outcomes through divergent thinking with different reasoning paths. Drawing
199 inspiration from process reward models [77, 46], which qualify each intermediate reasoning step
200 based on its potential to reach the correct outcome, we propose a customized process-reward LLM-
201 as-a-Judge method (PR-Judge) to access each association reasoning step towards the final outcome
202 connections, offering insights of reasoning process that outcome-based metrics cannot capture.

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

213 These stepwise indicators are inspired from Guilford's Alternate Uses [74] that reflects the
 214 divergent thinking behaviors of human. R_t and D_t are scalar values in $[0, 1]$ while K_t is
 215 binary in $(0 \text{ or } 1)$.

- 216 3. **Stepwise Association Quality and Path Scoring:** With regards to the indicators, the
 217 association quality per step s_t is calculated as:

$$s_t = \alpha R_t D_t + (1 - \alpha) K_t, \quad (1)$$

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

$$S_r = \sum_{t=1}^n s_t \delta^t. \quad (2)$$

219 Among them, α balances internal reasoning coherence $R_t D_t$ against knowledge K_t ; δ
 220 serves as a cognitive decay factor resembling the spirit of self-supervised process reward
 221 model [77], inherently favoring efficient and precise reasoning paths.

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

224 5 Experiments and Analysis

225 5.1 Settings

226 **LVLM Baselines.** We evaluated both proprietary and open-source VLMs under zero-shot conditions
 227 with default temperature. Proprietary models¹ include GPT-4 Omni [62], o4-mini [63], Gemini-
 228 1.5-Pro [72], Gemini-1.5-Flash [72], Gemini-2.5-Pro-Preview [24], Gemini-2.0-Flash-Thinking-
 229 Experimental [23], Claude-3.5-Sonnet [6], Qwen-VL-Max [8], Qwen-VL-Plus [8], while open-source
 230 models consist of GLM-4V [30], Yi-VL-34B [85], InternVL-Chat-V1-2 [18], VILA1.5 [47], and
 231 InternLM-XComposer2.5-7B [91]. Experiments for locally deployed models were conducted using
 232 80 GB NVIDIA A800 GPUs.

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

238 **Judge Engine.** We employ GPT-4o (gpt-4o-2024-08-06) and DeepSeek-V3 [25] as the mixed basic
 239 LLM-as-a-Judge engine for scoring. The former is excluded to evaluate its LVLM variant to ensure
 240 the fairness and prevent self-enhancement bias.

241 5.2 Outcome Evaluation of Association Reasoning

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

¹The 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.

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	<u>37.88</u>	<u>28.14</u>
Gemini-2.5-Pro-Preview	<u>60.05</u>	23.89	41.75	17.86	63.09	12.85	41.15	28.30
o4-mini	60.33	<u>19.86</u>	<u>37.89</u>	<u>18.03</u>	<u>61.55</u>	<u>10.24</u>	36.60	26.36
GPT-4o	59.72	10.89	28.83	17.94	58.26	6.27	29.62	23.35
OpenSource LVLMs								
GLM-4V	26.92	0.49	4.73	4.24	43.63	0.20	3.67	3.47
InternVL-Chat-V1-2	36.41	<u>3.52</u>	<u>16.02</u>	12.5	34.30	0.62	9.59	8.97
InternLM-XComposer2.5-7B	50.21	2.21	14.39	12.18	<u>44.87</u>	1.41	<u>18.18</u>	<u>16.77</u>
VILA1.5	<u>46.72</u>	2.45	15.38	<u>12.93</u>	44.46	1.27	14.93	13.66
Yi-VL-34B	45.25	4.97	19.63	14.66	54.39	<u>1.30</u>	19.53	18.23
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.

244 **LVLMs Far Below Humans in Association Reasoning.** MM-OPERA reveals the formidable
 245 challenges of associative reasoning for current LVLMs. While latest models like o4-mini and latest
 246 Gemini models show improved performance, with SR approaching the human baseline, they still fall
 247 short in achieving high-quality associations. For instance, on the RIA task, o4-mini achieves an HR-4
 248 of 19.86% compared to humans’ 22.84%, and on the ICA task, Gemini-2.5-Pro-Preview reaches
 249 an HR-4 of 12.85% against humans’ 31.65%, which demonstrates that sophisticated associative
 250 reasoning remains at the cutting edge of LVLM capabilities. Case studies in Appendix C illuminate
 251 key limitations, such as cross-domain knowledge retrieval deficiencies and perceptual misalignments.
 252

252 **Creativity Gap in Divergent Thinking.** The △HR metric highlights divergent thinking, with most
 253 models scoring 12%–20%, showing their ability to generate reasonable yet non-optimal associations.
 254 Latest Gemini models lead among LVLMs (18.87% and 28.30% in two tasks), but humans outperform
 255 with both higher △HR (26.13% and 29.82%) and HR-3 scores, demonstrating a superior balance of
 256 creativity and accuracy—an area where LVLMs remain limited.
 257

257 **ICA: Dual Challenge of Pattern Abstraction and Transfer.** Most models perform better on RIA
 258 than ICA, highlighting the challenges of pattern-based associative reasoning. ICA requires not
 259 only connecting concepts but also abstracting and transferring these patterns to new contexts—a
 260 complex process demanding advanced meta-reasoning. Notably, latest Gemini models, Yi-VL-34B
 261 and GLM-4V outperform on ICA compared to RIA, suggesting that certain architectures excel
 262 in specific associative reasoning tasks. These distinctions may stem from more effective pattern
 263 extraction or transfer mechanisms, warranting further investigation.
 264

264 **Conservative Reasoning vs. Associative Flexibility.** Analysis shows an inverse correlation between
 265 model constraints and associative abilities. Gemini-1.5-Flash (55.86% SR on RIA), optimized for
 266 speed, outperforms Gemini-1.5-Pro (45.34% SR), despite Pro’s larger size and focus on detailed
 267 reasoning. Examination of 500 random RIA samples (Figure 3) shows Pro’s conservative behavior to
 268 reason the high-rate association, prioritizing factuality and ethics, led to 1 point scores on nearly 20%
 269 of RIA questions due to conservative responses like “unrelated”, versus Flash’s <10%. Flash tended
 270 to offer superficial connections where Pro declined. Thus, factuality checks and ethical considerations,
 271 while improving reliability for complex tasks, can limit performance on creative association.
 272

272 5.3 Process Evaluation of Association Reasoning

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

Model	RIA		ICA	
	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.

279 scoring ranges (Deepseek-V3 trending higher), the self-enhancement bias of GPT-4o and the impact
 280 on comparative rankings remained minimal (Appendix B.2).

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

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

299 5.4 Sensitivity Analysis of Task Instances

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

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

312 **Text-Image Substitution Test.** We evaluated cross-modal generalization by replacing images
 313 with text descriptions and comparing scores across conditions. Appendix B.3.2 suggests GPT-4o’s
 314 reliance on nuanced visual cues that text descriptions cannot fully capture, while Gemini models
 315 demonstrate stronger text-equivalence in their reasoning, potentially processing visual information
 316 through language-like internal representations. These findings highlight how different architectural
 317 approaches influence cross-modal generalization in associative reasoning tasks.

318 **Order Sensitivity Test.** We examined the model’s sensitivity to input order by reversing the image
 319 sequence. Appendix B.3.3 suggests that while GPT-4o processes image pairs in a more commutative
 320 manner, treating both ordering equally, Gemini models, particularly Gemini-1.5-Pro, appear to apply

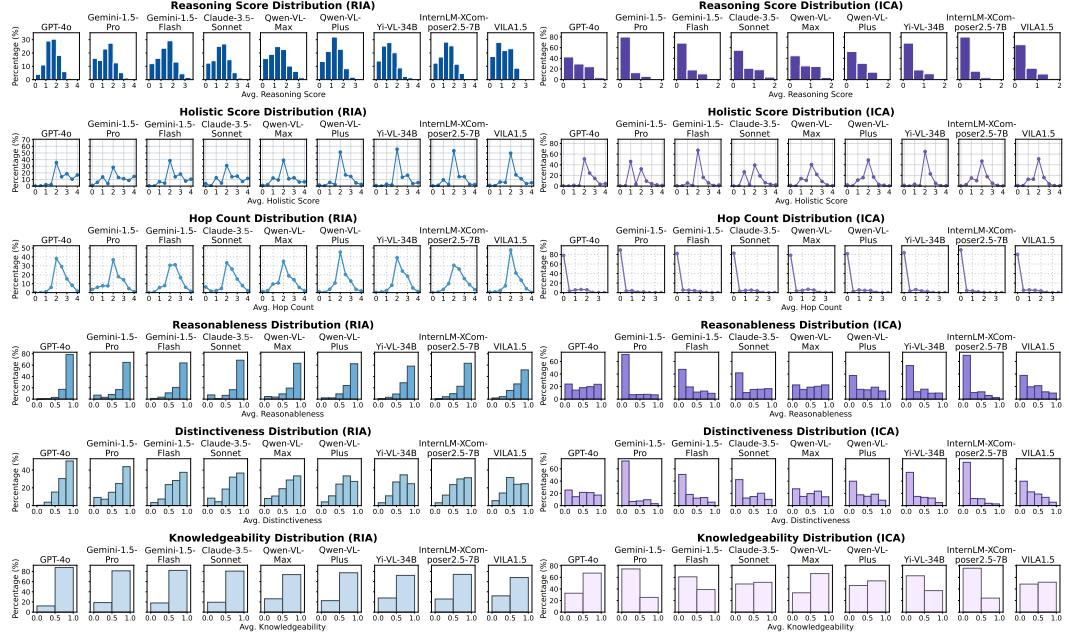


Figure 3: Fine-grained reasoning capability analysis of nine multimodal language models on RIA (left) and ICA tasks (right). From top to bottom: reasoning score distribution, holistic score distribution, reasoning path hop count distribution, Reasonableness distribution, Distinctiveness distribution,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

321 asymmetric reasoning processes that may prioritize the first image as context and the second as
 322 the target for association, highlighting architectural differences in how models approach bimodal
 323 associative reasoning.

324 5.5 LLM-as-a-Judge Strategy Validation for Reliable Evaluation

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

326 **Bias Analysis.** We addressed **verbosity and position biases** [93]. Excluding short 1-point responses,
 327 the Pearson correlation between response length and scores was 0.376 for regular scoring and 0.291
 328 for PR-Judge, indicating minimal verbosity bias. Permutation tests shuffling answer order showed
 329 mean score differences below 0.1 (regular) and 0.16 (PR-Judge), confirming negligible position bias.

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

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

340 6 Conclusion

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

347 **References**

- 348 [1] Jean-Baptiste Alayrac, Jeff Donahue, Paul Luc, Antoine Miech, Ian Barr, Yana Hasson, Lucas Leute, Katie
349 Millican, Malcolm Reynolds, Roy Ring, et al. Flamingo: a visual language model for few-shot learning.
350 In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- 351 [2] Pamela I Ansburg and Katherine Hill. Creative and analytic thinkers differ in their use of attentional
352 resources. *Personality and Individual Differences*, 34(7):1141–1152, 2003.
- 353 [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick,
354 and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference
355 on Computer Vision (ICCV)*, pages 2425–2433, 2015.
- 356 [4] Zahra Ashktorab, Michael Desmond, Qian Pan, James M Johnson, Martin Santillan Cooper, Elizabeth M
357 Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. Aligning human and
358 llm judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy
359 preferences. *arXiv preprint arXiv:2410.00873*, 2024.
- 360 [5] David P Ausubel. The psychology of meaningful verbal learning. 1963.
- 361 [6] Firstname1 Author1, Firstname2 Author2, and Firstname3 Author3. The claude 3 model family: Opus,
362 sonnet, haiku. *Proceedings of the Conference/Journal Name*, Volume Number(Issue Number):Page Range,
363 2023.
- 364 [7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han,
365 Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- 366 [8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou,
367 and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*,
368 abs/2308.12966, 2023.
- 369 [9] Roger E Beaty, Alexander P Christensen, Mathias Benedek, Paul J Silvia, and Daniel L Schacter. Creative
370 constraints: Brain activity and network dynamics underlying semantic interference during idea production.
371 *NeuroImage*, 148:189–196, 2017.
- 372 [10] Maxi Becker and Roberto Cabeza. Assessing creativity independently of language: A normed language
373 independent remote associate task (li-rat). *Europe PMC free article*, 2021.
- 374 [11] Maxi Becker and Roberto Cabeza. Assessing creativity independently of language: A language-independent
375 remote associate task (li-rat). *Behavior Research Methods*, 55(1):85–102, 2023.
- 376 [12] Mathias Benedek and Aljoscha C. Neubauer. Revisiting mednick’s model on creativity-related differences
377 in associative hierarchies. evidence for a common path to uncommon thought. *The Journal of Creative
378 Behavior*, 47:273 – 289, 2013.
- 379 [13] Théophile Bieth, Marcela Ovando-Tellez, Alizée Lopez-Persem, Béatrice Garcin, Laurent Hugueville,
380 Katia Lehongre, Richard Levy, Nathalie George, and Emmanuelle Volle. Time course of eeg power during
381 creative problem-solving with insight or remote thinking. *Human Brain Mapping*, 45(1):e26547, 2024.
- 382 [14] Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan
383 Szpektor, Amir Globerson, and Yuval Elovici. Visual riddles: a commonsense and world knowledge
384 challenge for large vision and language models. *arXiv preprint arXiv:2407.19474*, 2024.
- 385 [15] Edward M Bowden and Mark Jung-Beeman. Normative data for 144 compound remote associate problems.
386 *Behavior research methods, instruments, & computers*, 35:634–639, 2003.
- 387 [16] Denise J Cai, Sarnoff A Mednick, Elizabeth M Harrison, Jennifer C Kanady, and Sara C Mednick. Rem,
388 not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy
389 of Sciences*, 106(25):10130–10134, 2009.
- 390 [17] Wenhui Chen, Hongmin Wang, Junkun Song, Shiji Tang, Ming-Wei Chang, and William Yang Wang.
391 Bongard-logo: A new benchmark for human-level concept learning and reasoning. In *Advances in Neural
392 Information Processing Systems (NeurIPS)*, pages 21505–21517, 2021.
- 393 [18] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang,
394 Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision
395 foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- 396 [19] David Cropley. Is artificial intelligence more creative than humans?: Chatgpt and the divergent association
397 task. *Learning Letters*, 2:13–13, 2023.

- 398 [20] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic
 399 analysis of hallucination in gpt-4v(ision): Bias and interference challenges. 2023.
- 400 [21] J Barton Cunningham, James N MacGregor, Jenny Gibb, and Jarrod Haar. Categories of insight and their
 401 correlates: An exploration of relationships among classic-type insight problems, rebus puzzles, remote
 402 associates and esoteric analogies. *The Journal of Creative Behavior*, 43(4):262–280, 2009.
- 403 [22] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang
 404 Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with
 405 instruction tuning, 2023.
- 406 [23] Google DeepMind. Gemini 2.5 flash. Website, 2025. <https://deepmind.google/technologies/gemini/flash/>.
- 407 [24] Google DeepMind. Gemini 2.5 pro. Website, 2025. <https://deepmind.google/technologies/gemini/pro/>.
- 410 [25] DeepSeek-AI. Deepseek-v3 technical report, 2024.
- 411 [26] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong
 412 Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.
- 413 [27] Chaoyou Fu, Peixian Chen, Yunhang Shen, et al. Mme: A comprehensive evaluation benchmark for
 414 multimodal large language models. *Preprint*, 2024.
- 415 [28] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith,
 416 Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In
 417 *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- 418 [29] Wentao Ge, Shunian Chen, Guiming Hardy Chen, et al. Mllm-bench: Evaluating multimodal llms with
 419 per-sample criteria. *Preprint*, 2024.
- 420 [30] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng,
 421 Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie
 422 Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng
 423 Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam,
 424 Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue
 425 Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi
 426 Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang.
 427 Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.
- 428 [31] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh.
 429 Making the v in vqa matter: Elevating the role of image understanding in visual question answering.
 430 *International Journal of Computer Vision*, page 398–414, 2019.
- 431 [32] Joy P Guilford. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1):3–14,
 432 1967.
- 433 [33] David J Hargreaves. Musical imagination: Perception and production, beauty and creativity. *Psychology of
 434 music*, 40(5):539–557, 2012.
- 435 [34] José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measure-
 436 ment. *Artificial Intelligence Review*, 48:397–447, 2017.
- 437 [35] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross
 438 Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In
 439 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
- 440 [36] Yennie Jun. Exploring creativity in large language models: From gpt-2 to gpt-4.
- 441 [37] Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare
 442 Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, et al. Remi: A dataset for reasoning with multiple
 443 images. *Advances in Neural Information Processing Systems*, 37:60088–60109, 2025.
- 444 [38] Bohao Li, Yuying Ge, Guangzhi Wang, et al. Seed-bench-2: Benchmarking multimodal large language
 445 models. *Preprint*, 2023.
- 446 [39] Bing Li, Qi Li, Tianle Yang, Bowen Zheng, Yufeng Xie, Dawei Qi, Yue Zhang, Xiaoyan Zhu, and Jie
 447 Tang. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint
 448 arXiv:2307.16125*, 2023.

- 449 [40] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang,
 450 Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*,
 451 2024.
- 452 [41] Hong Li, Nanxi Li, Yuanjie Chen, Jianbin Zhu, Qinlu Guo, Cewu Lu, and Yong-Lu Li. The labyrinth of
 453 links: Navigating the associative maze of multi-modal llms. *arXiv preprint arXiv:2410.01417*, 2024.
- 454 [42] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C.H. Hoi. Blip-2: Bootstrapping language-image pre-
 455 training with frozen image encoders and large language models. In *Proceedings of the 40th International
 456 Conference on Machine Learning (ICML)*, 2023.
- 457 [43] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, WayneXin Zhao, and Ji-Rong Wen. Evaluating object
 458 hallucination in large vision-language models.
- 459 [44] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. Quantifying ai
 460 psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*,
 461 2024.
- 462 [45] Yibai Li, Xiaolin Lin, Zhenghui Sha, Zhiye Jin, and Emily Lee. Ai psychometrics: Evaluating the
 463 psychological reasoning of large language models with psychometric validities. 2025.
- 464 [46] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John
 465 Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.
- 466 [47] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad
 467 Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.
- 468 [48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár,
 469 and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014:
 470 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages
 471 740–755. Springer, 2014.
- 472 [49] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou.
 473 Hallusionbench: You see what you think? or you think what you see? an image-context reasoning
 474 benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models. 2023.
- 475 [50] Haotian Liu, Chunyuan Li, Qingyang Wu, YongJae Lee, Madison Madison, and Microsoft Research.
 476 Visual instruction tuning.
- 477 [51] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction
 478 tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages
 479 26296–26306, 2024.
- 480 [52] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezheng Peng, Mingyu Liu, Mingrui
 481 Chen, Chunyuan Li, Lianwen Jin, and Xiang Bai. On the hidden mystery of OCR in large multimodal
 482 models. *CoRR*, abs/2305.07895, 2023.
- 483 [53] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi
 484 Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an
 485 all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September
 486 29–October 4, 2024, Proceedings, Part VI*, pages 216–233. Springer, 2024.
- 487 [54] Pan Lu, Rui Shi, Kun Zhao, Simiao Zuo, Michael Zeng, and Sijia Liu. Mathvista: Evaluating mathematical
 488 reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- 489 [55] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question
 490 answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on
 491 Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019.
- 492 [56] Sarnoff A. Mednick. The associative basis of the creative process. *Psychological review*, 69:220–32, 1962.
- 493 [57] Sarnoff A. Mednick. The remote associates test. *The Journal of Creative Behavior*, 2(3):213–214, 1962.
- 494 [58] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha,
 495 Andre Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of
 496 fine-grained categories. 2023.
- 497 [59] Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. Naming unre-
 498 lated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118,
 499 2021.

- 500 [60] Ana-Maria Oltețeanu and Faheem Hassan Zunjani. A visual remote associates test and its validation.
 501 *Frontiers in psychology*, 11:26, 2020.
- 502 [61] Ana-Maria Oltețeanu, Mikkel Schöttner, and Susanne Schuberth. Computationally resurrecting the
 503 functional remote associates test using cognitive word associates and principles from a computational
 504 solver. *Knowledge-Based Systems*, 168:1–9, 2019.
- 505 [62] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.
- 506 [63] OpenAI. Introducing openai o3 and o4-mini. Website, 2025. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- 507
- 508 [64] Marcela Ovando-Tellez, Yoed N Kenett, Mathias Benedek, Matthieu Bernard, Joan Belo, Benoit Beranger,
 509 Theophile Bieth, and Emmanuelle Volle. Brain connectivity-based prediction of combining remote
 510 semantic associates for creative thinking. *Creativity Research Journal*, 35(3):522–546, 2023.
- 511 [65] Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai
 512 psychometrics: Assessing the psychological profiles of large language models through psychometric
 513 inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024.
- 514 [66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish
 515 Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from
 516 natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR,
 517 2021.
- 518 [67] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Krahenbuehl, and Vladlen Koltun. Does spatial
 519 cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024.
- 520 [68] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru,
 521 Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual
 522 benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.
- 523 [69] Claudia Schon, Ulrich Furbach, and Marco Ragni. Modeling associative reasoning processes. *arXiv
 524 preprint arXiv:2201.00716*, 2022.
- 525 [70] Masaki Suwa and Barbara Tversky. Constructive perception: A metacognitive skill for coordinating
 526 perception and conception. In *Proceedings of the 25th Annual Cognitive Science Society*, pages 1140–1145.
 527 Psychology Press, 2013.
- 528 [71] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers.
 529 In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,
 530 pages 5100–5111, 2019.
- 531 [72] Gemini Team and Google Google. Gemini: A family of highly capable multimodal models.
- 532 [73] Classical Divergent Thinking. 19 associative and controlled cognition in divergent thinking: Theoretical,
 533 experimental, neuroimaging evidence, and new directions. *The Cambridge Handbook of the Neuroscience
 534 of Creativity*, page 333, 2018.
- 535 [74] Oshin Vartanian, Erin L Beatty, Ingrid Smith, Sarah Forbes, Emma Rice, and Jenna Crocker. Measurement
 536 matters: the relationship between methods of scoring the alternate uses task and brain activation. *Current
 537 Opinion in Behavioral Sciences*, 27:109–115, 2019.
- 538 [75] Deana Vitranio, Jeanette Altarriba, and Deniz Leblebici-Basar. Revisiting mednick's (1962) theory of
 539 creativity with a composite measure of creativity: The effect of stimulus type on word association
 540 production. *The Journal of Creative Behavior*, 2021.
- 541 [76] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming
 542 Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models.
 543 *arXiv preprint arXiv:2308.15126*, 2023.
- 544 [77] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang
 545 Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint
 546 arXiv:2312.08935*, 2023.
- 547 [78] Xiting Wang, Liming Jiang, Jose Hernandez-Orallo, David Stillwell, Luning Sun, Fang Luo, and Xing Xie.
 548 Evaluating general-purpose ai with psychometrics. *arXiv preprint arXiv:2310.16379*, 2023.

- 549 [79] Jamie Ward, Daisy Thompson-Lake, Roxanne Ely, and Flora Kaminski. Synaesthesia, creativity and art:
 550 What is the link? *British Journal of Psychology*, 99(1):127–141, 2008.
- 551 [80] R.A. Wolf. Defining the concept of creativity, 2014.
- 552 [81] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning
 553 benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.
- 554 [82] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang,
 555 Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language
 556 models. *arXiv preprint arXiv:2306.09265*, 2023.
- 557 [83] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren
 558 Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models.
 559 In *The Thirteenth International Conference on Learning Representations*, 2024.
- 560 [84] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong
 561 Wang, Lu Sheng, Lei Bai, Jing Shao, and Wanli Ouyang. LAMM: language-assisted multi-modal
 562 instruction-tuning dataset, framework, and benchmark. In *Advances in Neural Information Processing
 563 Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New
 564 Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- 565 [85] Alex Young, Bei Chen, Chao Li, Chengan Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu,
 566 Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*,
 567 2024.
- 568 [86] Jiahui Yu, Xiaowei Chen, Jingren Shen, Lu Yuan, Wei Chang, and Thomas S Huang. Coca: Contrastive
 569 captioners are image-text foundation models. In *Advances in Neural Information Processing Systems
 570 (NeurIPS)*, 2022.
- 571 [87] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and
 572 Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities.
- 573 [88] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu
 574 Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding
 575 and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision
 576 and Pattern Recognition*, pages 9556–9567, 2024.
- 577 [89] Fengji Zhang, Linquan Wu, Huiyu Bai, Guancheng Lin, Xiao Li, Xiao Yu, Yue Wang, Bei Chen, and Jacky
 578 Keung. Humaneval-v: Evaluating visual understanding and reasoning abilities of large multimodal models
 579 through coding tasks. *arXiv preprint arXiv:2410.12381*, 2024.
- 580 [90] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Lijuan Yang, Lei Zhang, Yejin Wang, and Jianfeng Gao. Vinvl:
 581 Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference
 582 on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, 2021.
- 583 [91] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan,
 584 Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue
 585 Zhang, Wei Li, Jingwen Li, Wenhui Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng
 586 Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language
 587 model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.
- 588 [92] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On
 589 evaluating adversarial robustness of large vision-language models.
- 590 [93] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,
 591 Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena.
 592 *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- 593 [94] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan
 594 Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor
 595 generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
 596 pages 13246–13257, 2024.
- 597 [95] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing
 598 vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*,
 599 2023.
- 600 [96] Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Zachary A Pardos, Patrick C Kyllonen, Jiyun Zu,
 601 Qingyang Mao, Rui Lv, Zhenya Huang, et al. From static benchmarks to adaptive testing: Psychometrics
 602 in ai evaluation. *arXiv preprint arXiv:2306.10512*, 2023.

603 **Technical Appendices**

- 604 • Section A: More Benchmark Details
605 • Section B: Supplementary Results and Analysis
606 • Section C: Case Study: Why Do Models Perform Poorly?
607 • Section D: Test and Evaluation Prompts
608 • Section E: Limitations and Broader Impacts

609 **A More Benchmark Details**

610 **A.1 Association Path**

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

613 **Type 1: Sequential Association**

- 614 • Structure: $A \rightarrow X_1 \rightarrow X_2 \rightarrow B$
615 • Format:
 Predicate1(A, X_1) and Predicate2(X_1, X_2)
 Predicate3(X_2, B)
 $A \rightarrow X_1 \rightarrow X_2 \rightarrow B$

616 **Type 2: Convergent Association**

- 617 • Structure: $A \rightarrow X_1 \rightarrow X_2$ and $B \rightarrow X_2$
618 • Format:
 Predicate1(A, X_1) and Predicate2(X_1, X_2)
 Predicate3(B, X_2)
 $A \rightarrow X_1 \rightarrow X_2$ and $B \rightarrow X_2$

619 **Type 3: Metaphorical Association**

- 620 • Structure: $A \wedge B \rightarrow X$
621 • Format: $A \wedge B \rightarrow X$

622 **Notation Conventions:** Entities and predicates follow PascalCase naming convention. The symbol
623 ‘and’ connects separate relational clauses, while ‘ \wedge ’ represents logical conjunction between entities.
624 Each arrow (\rightarrow) represents one associative hop.

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

627 **A.2 Hierarchical Association Annotation**

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

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

- 632 • *Perception*: Processes immediate sensory inputs, focusing on visual understanding and
633 interpretation
634 • *Conception*: Handles abstract, knowledge-driven associations requiring higher-order cogni-
635 tive processing

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

- 637 • Under *Perception*: Recognition, Context, and Interaction
 638 • Under *Conception*: Logic, Semantic, and Reasoning
- 639 **Level-3 (L-3)** provides the most granular classification with thirteen specific dimensions. Each
 640 dimension captures a distinct aspect of associative reasoning. Table 4 presents detailed definitions for
 641 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.
		Social Insight Understanding emotions and interactions between people in visual scenes.
Conception	Interaction	Relational Perception Comprehension of spatial and logical relationships between objects.
		Functional Links Associations based on functional relationships between concepts.
	Logic	Causal Connections Associations based on cause-and-effect relationships.
		Thematic Links Associations within the same theme or context.
	Semantic	Cultural Reference Associations based on cultural knowledge and specific contexts.
		Hierarchical Association Vertical associations between abstract and concrete concepts.
		Analogical Reasoning Associations based on structural, feature, or pattern similarities.

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

646 A.3 Data Sources

647 The MM-OPERA-Bench dataset, consisting of images, reference answers, and fine-grained annotations,
 648 was manually curated by a group of volunteers. Of the total data, 33.35% of the questions
 649 and reference answers were sourced from the RAT [57], while 4.01% of the images, questions, and
 650 reference answers were sourced from the LI-RAT [10] datasets for human psychometric testing. The
 651 remaining images were sourced from the Internet, and all fine-grained annotations were manually
 652 constructed and revised to ensure consistency and accuracy.

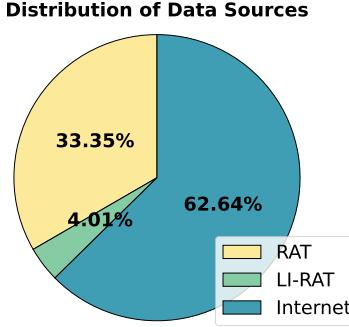


Figure 4: Distribution of data sources.

653 A.4 Data Collection and Curation Protocol

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

657 A.4.1 Data Collection

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

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

675 A.4.2 Quality Control

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

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

688 **B Supplementary Results and Analysis**

689 **B.1 Multi-dimensional Analysis**

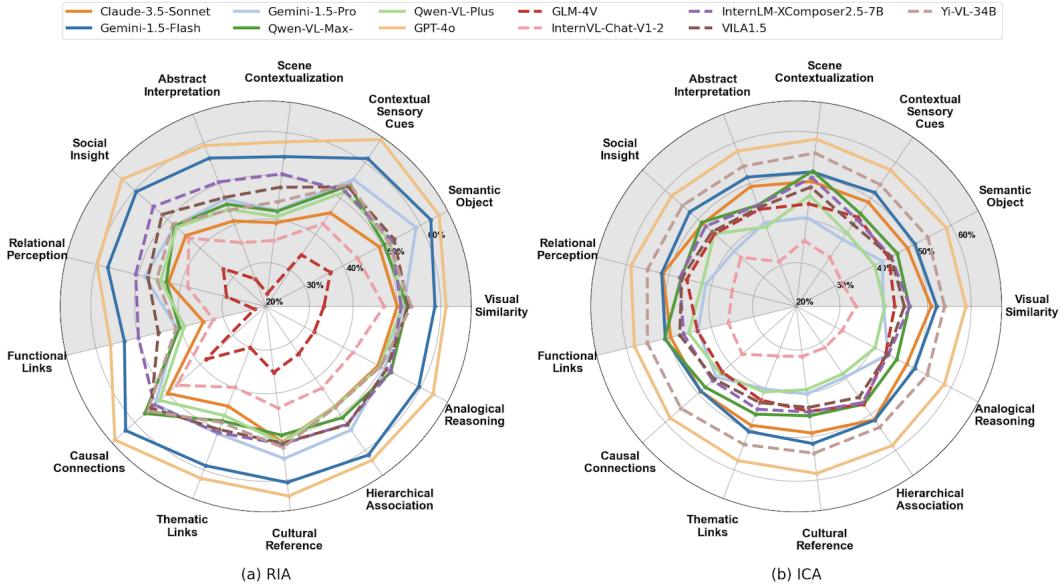


Figure 5: Comparison of Model Performance in RIA and ICA across Different Conceptual (white background) and Perceptual (gray background) Dimensions. The radar charts illustrate the capabilities of various MLLMs in handling tasks related to relational perception, social insight, causal connections, abstract 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.

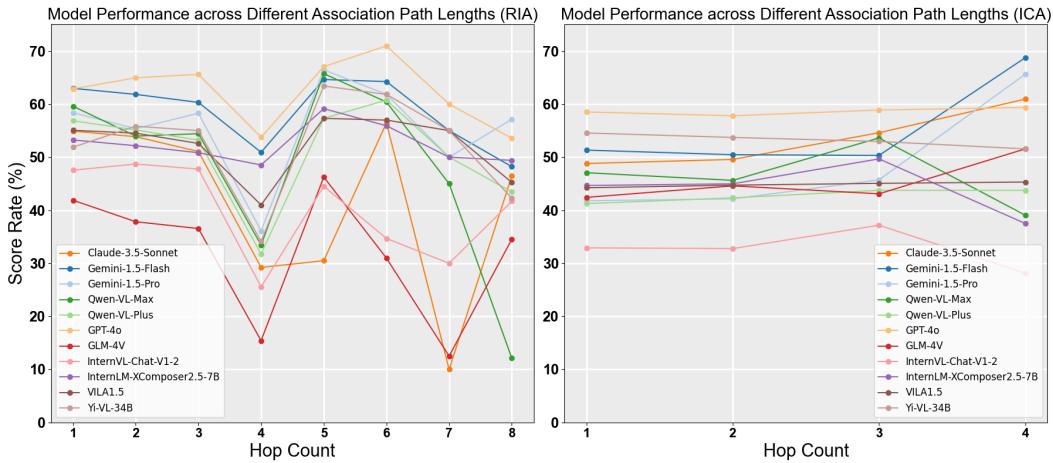


Figure 6: Model Performance Across Different Association Path Lengths in RIA and ICA tasks. The line graphs illustrate the score rates (%) of various MLLMs as the number of association path "hops" increases. The left chart represents RIA results, showing notable fluctuations in performance across different 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.

690 Multidimensional analysis reveals the complex landscape of associative reasoning capabilities in
 691 Multimodal Language Models (MLLMs). Most models perform better on RIA tasks compared to
 692 ICA tasks, with an average performance differential of approximately 5–7 percentage points. This

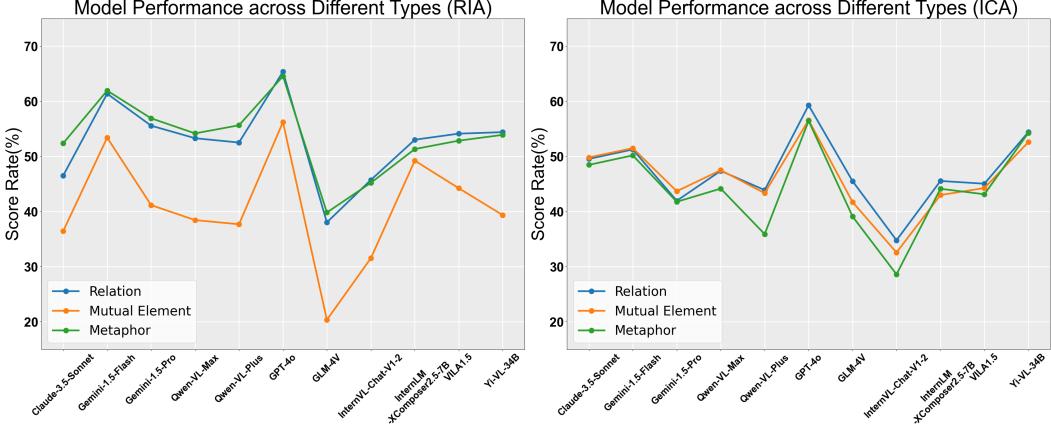


Figure 7: The figure presents the performance of various multimodal large models across different reasoning types in the RIA (left) and ICA (right) tasks. The three reasoning types—Relation, Mutual Element, and Metaphor—are represented by different colored lines. The vertical axis indicates the score 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.

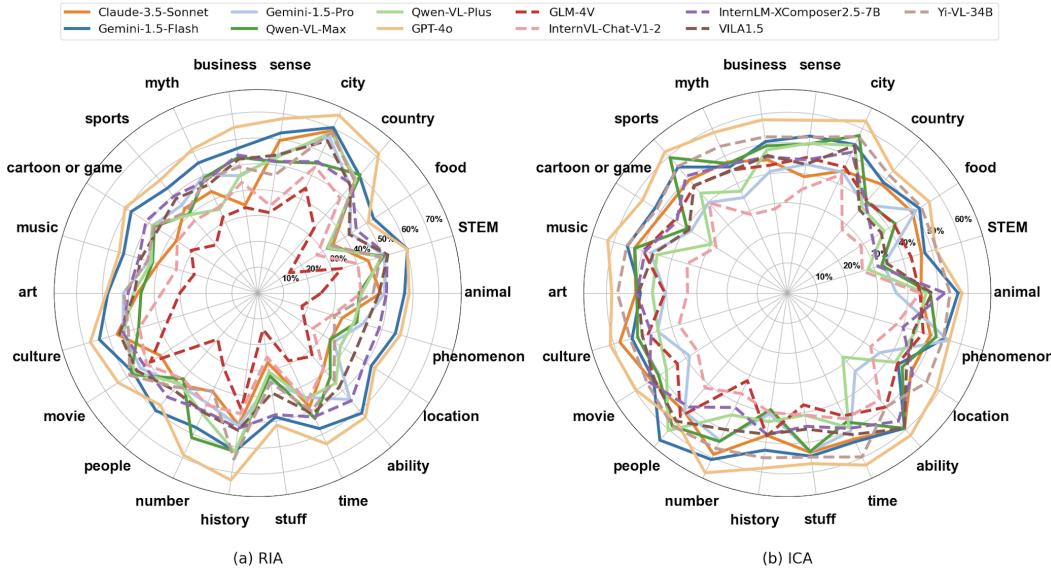


Figure 8: Radar Chart Comparison of Model Performance Across Domains in RIA and ICA tasks. The two radar charts display the performance of various MLLMs across different knowledge domains such as business, sports, music, STEM, history, and culture. The left chart (a) represents results from the RIA tasks, 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.

693 suggests that identifying direct associations between unrelated items may be more tractable for current
694 MLLMs than recognizing and extending associative patterns. The exception is Claude-3.5-Sonnet,
695 which shows relatively consistent performance across both task types, indicating potentially more
696 balanced associative reasoning capabilities. These findings underscore the multi-faceted nature of
697 associative cognition and the importance of diverse task designs for comprehensive evaluation.

698 **Reasoning Complexity and Cognitive Abilities.** Analysis of reasoning complexity reveals non-
699 linear patterns in how models handle associative tasks. Figure 6 shows that while most models
700 effectively manage simple 1–2 hop associations (with score rates around 50–60%), performance
701 drops significantly for more complex 4-hop associations (29–53%). However, some models (e.g.,

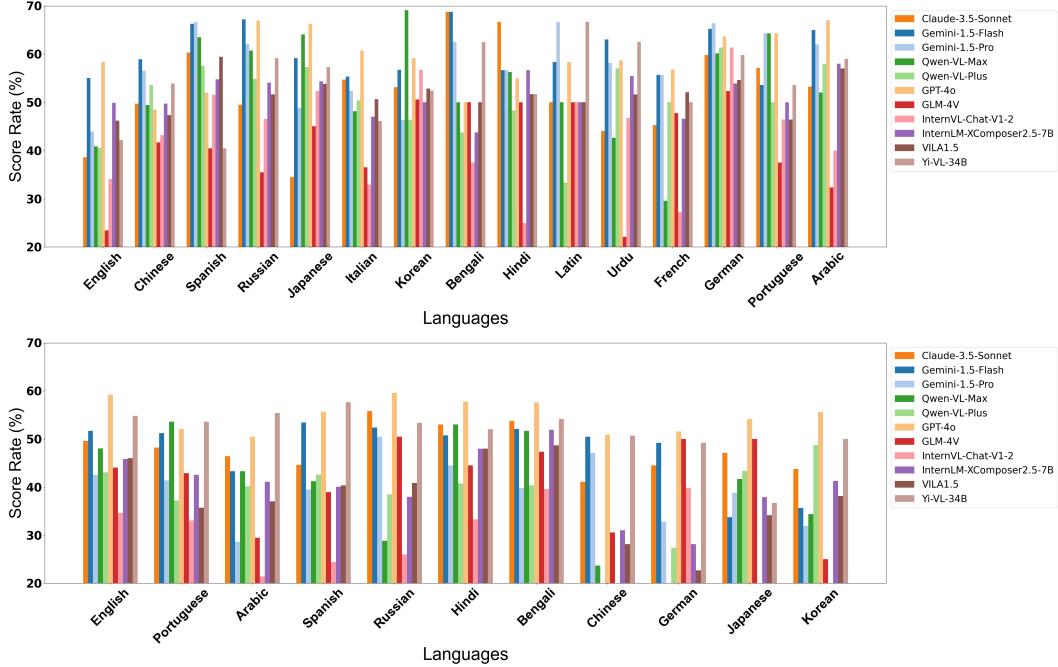


Figure 9: The figures compare the performance of different multimodal large models across multiple languages. The left figure represents the RIA tasks, while the right figure corresponds to the ICA tasks. The vertical axis indicates the score rate (%), and the horizontal axis lists various languages, 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.

702 GPT-4o, Gemini-1.5-Flash, and Qwen-VL-Plus) demonstrate relatively stable performance in very
 703 complex 5–8 hop associations, suggesting the emergence of new strategies in complex reasoning paths.
 704 This “complexity valley” phenomenon warrants further investigation as it may provide important
 705 insights into how MLLMs structure multi-step associative reasoning. Differences in perceptual and
 706 conceptual abilities are evident in Figure 5. “Semantic Object” in perceptual abilities and “Causal
 707 Connections” in conceptual abilities show stronger performance, while “Abstract Interpretation” and
 708 “Functional Links” remain challenging. Cross-task analysis indicates that models maintain consistent
 709 relative strengths across RIA and ICA tasks, but absolute performance levels are modulated by
 710 task demands, especially for perceptual abilities. This suggests that while underlying reasoning
 711 mechanisms remain stable, their expression is influenced by task requirements.

712 **Relationship Types.** Analyzing association types reveals distinctive performance patterns between
 713 Remote-Item Association (RIA) and In-Context Association (ICA) tasks. As shown in Figure 7, in
 714 RIA tasks, models demonstrate a pronounced hierarchy among association types, with Metaphor
 715 associations yielding the highest performance (52–65% for top performers), followed by Relation
 716 associations (46–65%), and Mutual Element associations showing the lowest scores (36–56%). This
 717 hierarchy is notably consistent across nearly all models. Interestingly, in ICA tasks, this performance
 718 stratification significantly diminishes, with much smaller performance gaps between association types.
 719 For instance, GPT-4o shows only a 2.82 percentage point difference between its highest (Relation:
 720 59.29%) and lowest (Mutual Element: 56.59%) association type performance in ICA, compared to a
 721 7.15 point gap in RIA. This convergence suggests that the contextual framework provided in ICA
 722 tasks may equalize the difficulty of recognizing different association types.

723 **Domain and Cultural Dimensions.** Domain knowledge differences are highly evident in model
 724 performance. As shown in Figure 8, urban-related associations consistently achieve higher per-
 725 formance (around 65–75% for top models), while everyday objects and food-related associations
 726 pose greater challenges (around 30–45%). These differences suggest inherent difficulties in forming
 727 associations within certain conceptual spaces. GPT-4o excels in history-related associations (73.07%),

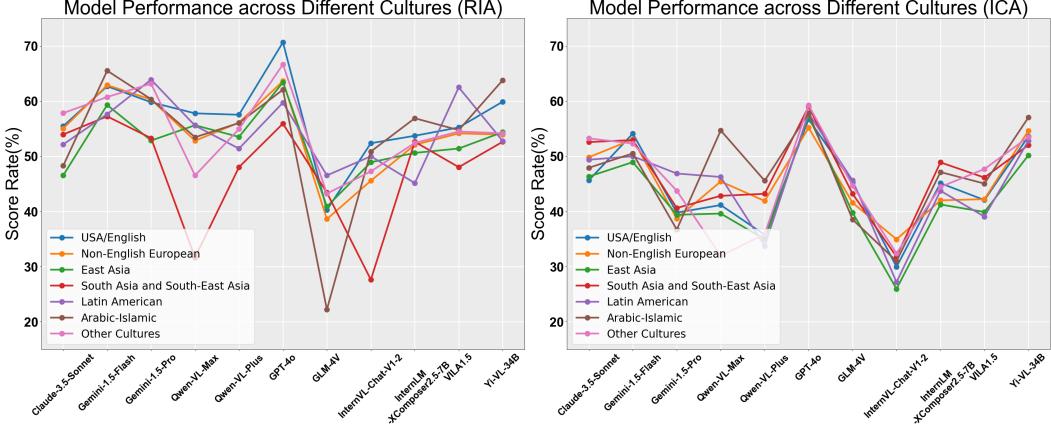


Figure 10: Comparison of Model Performance Across Different Cultures in RIA and ICA tasks. The line plots illustrate the score rates (%) of various MLLMs across cultural groups, including USA/English, Non-English European, East Asian, South Asian and South-East Asian, Latin American, Arabic-Islamic,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

Model	Remote-Item Association Task				In-Context Association Task			
	Holistic SR (%)		Avg. Reasoning Score		Holistic SR (%)		Avg. Reasoning Score	
	4o	V3	4o	V3	4o	V3	4o	V3
Claude-3.5-Sonnet	56.20	60.10	<u>1.2838</u>	1.5457	50.4	48.15	0.4159	0.6039
Gemini-1.5-Flash	63.25	60.65	1.2701	<u>1.5684</u>	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	<u>56.15</u>	<u>0.5584</u>	<u>0.7107</u>
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	<u>52.85</u>	53.35	0.3107	0.4027

Table 5: 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.

728 significantly outperforming other models, which may indicate superior historical knowledge representation
 729 or more effective temporal association retrieval mechanisms. Cultural background also
 730 significantly impacts model performance. Figure 10 reveals that most models show stronger associative
 731 reasoning when dealing with Western cultural references compared to East Asian, South Asian,
 732 or Arabic-Islamic contexts. In Figure 9, similar asymmetries are observed in language performance,
 733 with models generally performing better in Spanish, German, and Russian associations than in East
 734 Asian languages. These cultural and linguistic disparities may reflect imbalances in multilingual
 735 pretraining or fundamental differences in how associations manifest across different linguistic and
 736 cultural structures.

737 B.2 Evaluation by Different Judges

738 The comparison between the two judges (GPT-4o and Deepseek-V3) highlights notable differences
 739 in their scoring tendencies (see Table 5). Judge Deepseek-V3 consistently assigns higher average
 740 reasoning scores across most models, suggesting a more lenient evaluation of reasoning depth or
 741 quality. However, for holistic score rate (SR), the differences are less consistent, with some models
 742 (e.g., Claude-3.5-Sonnet and Qwen-VL-Max) receiving higher SR from V3, while others (e.g.,
 743 GPT-4o) show near parity between the two judges. These disparities underscore the importance of
 744 employing multiple evaluators to mitigate individual judgment bias and ensure robust evaluation of
 745 model performance.

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

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

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

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

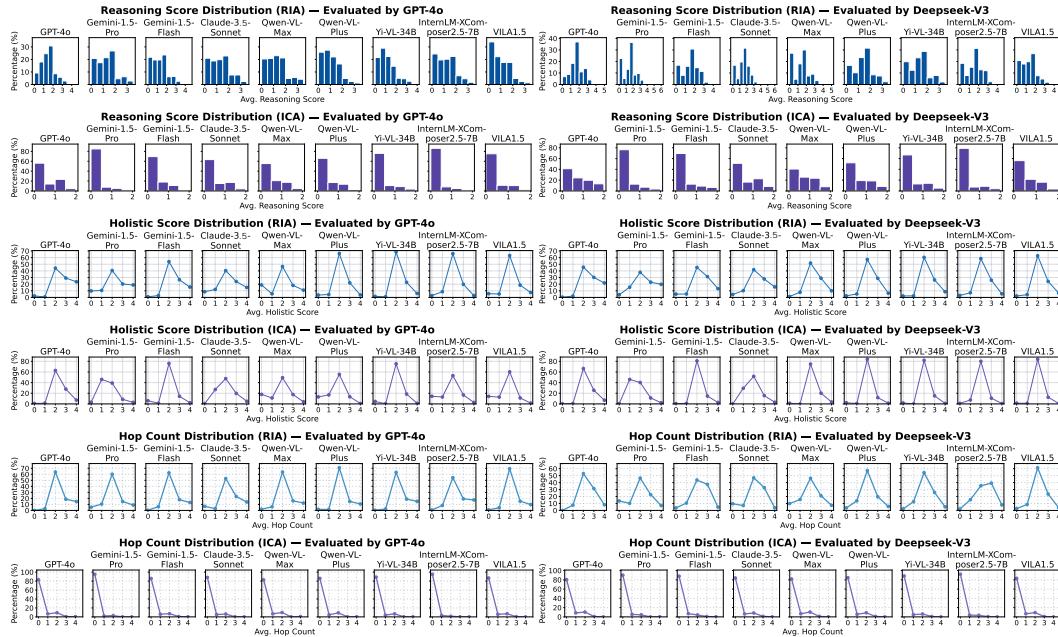


Figure 11: Fine-grained reasoning capability analysis of nine multimodal language models on RIA (blue) and ICA (purple) tasks judged by GPT-4o (left) and Deepseek-V3 (right). From top to bottom: reasoning score distribution, holistic score distribution, reasoning path hop count distribution. Each task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

776 **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 6: 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).

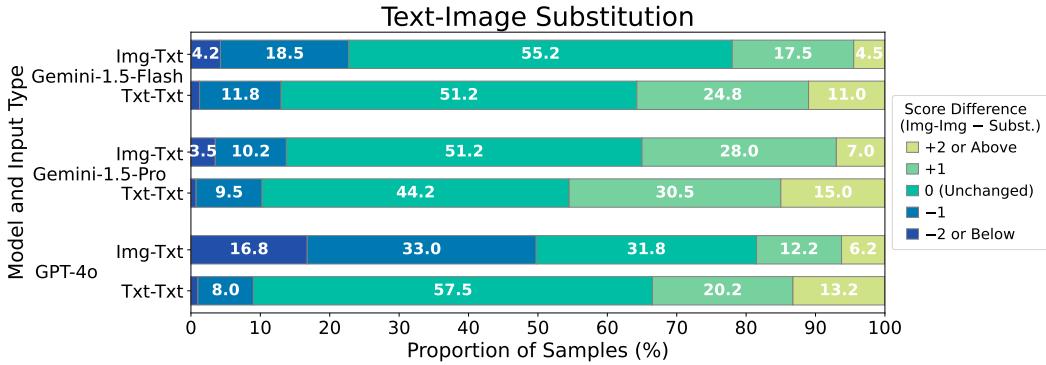


Figure 12: Score difference distribution for the Text-Image Substitution test across models in RIA. Bars show proportions of samples with varying score differences (original - substitution).

777 **B.3.1 Multi-Image Substitution Test**

778 To assess robustness, we conducted sensitivity tests to measure how MLLMs’ responses varied
 779 with different visual representations of the same concepts. Results in Table 6 revealed significant
 780 visual sensitivity across models. GPT-4o demonstrates exceptional consistency, showing the lowest
 781 intra-group score range (0.44) and standard deviation (0.18) while maintaining the highest score rate
 782 (0.59). In contrast, models like VILA1.5 and InternVL-Chat-V1-2 exhibit substantial variability (IG
 783 Ranges of 1.34 and 1.31, respectively) despite moderate performance, indicating that their associative
 784 reasoning is heavily influenced by specific visual features rather than robust concept understanding.
 785 This visual dependency suggests that most current MLLMs still associate at a surface feature level
 786 rather than at a deeper conceptual level—a critical limitation for real-world applications requiring
 787 consistent reasoning across variable visual inputs.

788 **B.3.2 Text-Image Substitution Test**

789 Results in Figure 12 reveals distinct cross-modal generalization patterns across models. GPT-4o
 790 experiences the most significant performance drop when images are replaced with text descriptions
 791 (Image-Text), with 49.8% of samples showing decreased scores (16.8% with severe drops of ≤ 2
 792 points). Conversely, Gemini models maintain more consistent performance across modalities, with
 793 Gemini-1.5-Flash showing unchanged scores in 55.2% of Image-Text substitutions. This pattern
 794 suggests fundamental differences in cross-modal processing strategies. GPT-4o appears more reliant
 795 on visual information for associative reasoning, extracting nuanced visual cues that text descriptions

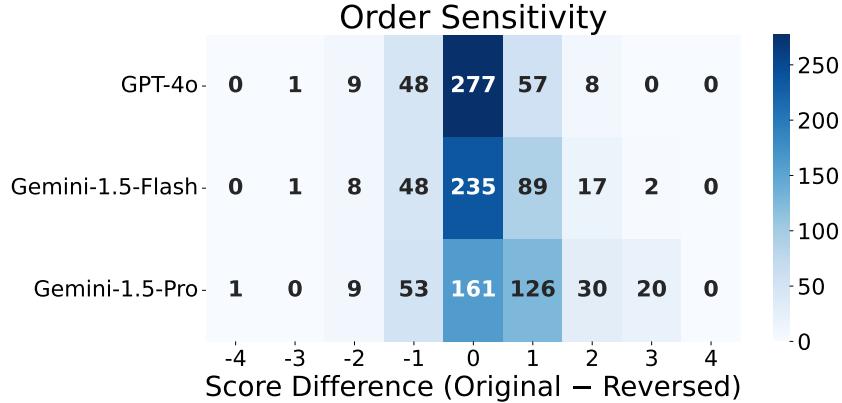


Figure 13: Order Sensitivity heatmap showing models versus score difference between original and reversed input order. Cell darkness indicates instance count.

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 MLLMs’ associative reasoning capabilities.

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.

842 **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.

851 **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.

865 **C Case Study: 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 MLLM 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.

875 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

881 the linguistic concept of “See” from an image depicting an act of looking). These perceptual errors
882 initiate cascading reasoning failures that fundamentally compromise the associative process. More
883 specifically, limitations in image resolution, the presence of visual noise, or a lack of sensitivity to
884 certain visual attributes can lead to perceptual inaccuracies. Furthermore, biases in understanding
885 spatial relationships, relative sizes, and interactions between objects within an image can impede
886 accurate scene interpretation.

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

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

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

918 Additionally, we observed a small percentage of cases (approximately 1%) where models declined to
919 engage with certain prompts due to safety or ethical considerations, and technical failures (approximate-
920 ly 2%) where models were unable to properly access all image inputs. These findings collectively
921 underscore the multifaceted challenges inherent in open-ended multimodal association tasks, high-
922 lighting the need for advancements in visual perception, cross-domain knowledge activation, and
923 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

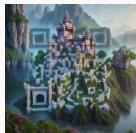


Figure 15: Image 2

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

MLM 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



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

MLLM 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 MLLM 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

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

MLLM 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 MLLM includes incorrect facts, referring to an imaginary story, thus leading to a mistaken chain of reasoning.

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: Im-
age 1



Figure 21: Im-
age 2



Figure 22: Im-
age 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

MLLM 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 MLLM 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.

928

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: Im-
age 1



Figure 24: Im-
age 2



Figure 25: Im-
age 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

MLLM 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: UContainer 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 MLLM’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

930

931 D Testing and Evaluation Prompts

932 We report our prompts for testing and evaluation.

933 D.1 Testing Prompts

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

Prompt: <Instruction> + <Format>

937

938 The specific implementation of our prompt structure is presented below.

939 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]

939

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]

940

941 D.2 Prompts for Regular LLM-as-a-Judge Scoring

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

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

951

952 The specific implementation of the prompt structure is presented below.

Scoring Rules for RIA Task

Evaluate the quality of responses provided by MLLMs 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.

953

- **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.

954

Rating Examples for RIA Task

Rating examples:

Example 1

MLLM 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 MLLM’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]

955

Scoring Rules for ICA Task

Background: MLLMs 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 MLLM 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.

956

- **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 MLLM's explanation. That is, even if MLLM'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.)

957

Rating Examples for ICA Task

Rating examples:

Example 1

MLLM 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 MLLM 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 MLLM’s explanation was clear and logical, with no errors.”

}

[4 more examples of different reasoning omitted for brevity]

958

I/O Format for Both RIA and ICA Tasks

You will receive multiple independent questions in a numbered format:

1. MLLM Output: []

Reference Answer: []
2. MLLM Output: []

Reference Answer: []

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

{

959

```

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

```

960

961 D.3 Prompts for LLM Judging in MM-OPERA Reasoning

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

```

Prompt: <Role Definition> + <Cross-Structured Instructions> +  

<Annotative Framework and I/O Protocol> + <Rating Examples>

```

973

974 The specific implementation of our prompt structure is presented below. And the sections marked
 975 with ellipses share a similar structure and content with the surrounding context and are therefore
 976 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
  }
  mllm_output: {
    description: [ImageA_desc, ImageB_desc],
    relation: string,
    explanation: string
  }
}

```

2. Path Standardization for mllm_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}"
}

```

977

```

(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": {...}}
```

978

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 → Eucharist → TheLastSupper and 12 → TheLastSupper

MLLM'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:

979

```
{
  "path": "Represent(BreadAndWine, Eucharist) and Symbolize(Eucharist, Christian)
Number(12, ApostlesOfJesus) and Related(ApostlesOfJesus, Christian)
Thus, BreadAndWine → Eucharist → Christian and 12 → ApostlesOfJesus → Christian",
  "shop_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 MLLM 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]

980

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

MLLMs 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 MLLM'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
  },
  mllm_output: {
    description: [Image1_desc, Image2_desc, Image3_desc, Image4_desc],
    relation: string,
    explanation: string
  }
}
```

2. Path Standardization for mllm_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
  ]
}
```

981

```

        - '$\land$' for entity connection
        - '→' for each association hop
    }
}

3. Quality Assessment for mllm_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 mllm_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": {...}}
```

982

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

MLLM'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"

983

- **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 MLLM'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]

984

985 E Limitations and Broader Impacts

986 E.1 Limitation

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

- 989 • **Limited Exploration of Temporal Association Reasoning**: MM-OPERA's static task
990 design (RIA and ICA) does not fully capture temporal or sequential association reasoning,
991 a key aspect of human cognition in dynamic contexts like decision-making, restricting its
992 evaluation scope.
- 993 • **High Cost and Scalability Challenges for Open-Ended Evaluation**: Evaluating 11,497
994 open-ended tasks with a resource-intensive LLM-as-a-Judge and cascading scoring rubric in-
995 curs high computational costs (due to increased token usage) and limits scalability, hindering
996 rapid or large-scale testing of LVLMs.
- 997 • **Challenges in Systematic Task Creation**: Although association is common in human
998 cognition, systematically collecting and converting ideas or existing data into task instances
999 is challenging, especially given LLMs' weaknesses in this area, leading to high human effort
1000 costs for data expansion.

1001 These limitations underscore the need for continued innovation to enhance MM-OPERA's robustness,
1002 scalability, and applicability in advancing AI research.

1003 **E.2 Broader Impacts**

1004 MM-OPERA is an **evaluation** benchmark for associative reasoning in Large Vision-Language Models
1005 (LVLMs), not a training set. It aims to deepen understanding and guide AI development. However,
1006 evaluation standards carry societal implications.

1007 **Potential Societal Considerations:**

- 1008 • **Guiding Development and Bias Risks:** Benchmarks shape research. Any unaddressed
1009 gaps or subtle biases within MM-OPERA could inadvertently steer development towards
1010 a narrow or skewed form of associative intelligence, impacting real-world fairness and
1011 applicability.
- 1012 • **Perception of Capabilities and Misuse Potential:** By identifying models with advanced
1013 associative abilities, MM-OPERA may elevate perceptions of their power. Such identified
1014 capacities, even if not developed via this benchmark, could be leveraged for sophisticated
1015 misuse (e.g., disinformation) if not responsibly managed.
- 1016 • **Deployment Risks from Identified Limitations:** MM-OPERA reveals model weaknesses.
1017 Overlooking these identified limitations during deployment in critical systems could lead to
1018 erroneous and harmful outcomes.

1019 **Mitigation and Responsible Use of Insights:** Transparency in design and responsible interpretation
1020 of MM-OPERA's results are crucial. We advocate for:

- 1021 • Continuous community scrutiny and refinement of MM-OPERA to address potential biases
1022 and representational gaps.
- 1023 • Using evaluation insights to understand fundamental AI limitations and guide research
1024 towards robust, safe, and aligned systems, beyond mere model ranking.
- 1025 • Informed deployment decisions by developers and deployers, using benchmarked strengths
1026 and weaknesses to assess suitability and mitigate risks.

1027 Our goal is for MM-OPERA to foster rigorous evaluation, contributing to more capable and societally
1028 beneficial AI.

1029 **NeurIPS Paper Checklist**

1030 **1. Claims**

1031 Question: Do the main claims made in the abstract and introduction accurately reflect the
1032 paper's contributions and scope?

1033 Answer: [Yes]

1034 Justification: The abstract and introduction accurately capture the our contributions and
1035 scope by outlining the MM-OPERA benchmark, its innovative evaluation approach, and the
1036 insights it provides into the limitations of existing Large Vision-Language Models.

1037 Guidelines:

- 1038 • The answer NA means that the abstract and introduction do not include the claims
1039 made in the paper.
- 1040 • The abstract and/or introduction should clearly state the claims made, including the
1041 contributions made in the paper and important assumptions and limitations. A No or
1042 NA answer to this question will not be perceived well by the reviewers.
- 1043 • The claims made should match theoretical and experimental results, and reflect how
1044 much the results can be expected to generalize to other settings.
- 1045 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
1046 are not attained by the paper.

1047 **2. Limitations**

1048 Question: Does the paper discuss the limitations of the work performed by the authors?

1049 Answer: [Yes]

1050 Justification: Yes, we have included the discussion of limitations in the appendix.

1051 Guidelines:

- 1052 • The answer NA means that the paper has no limitation while the answer No means that
1053 the paper has limitations, but those are not discussed in the paper.
- 1054 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 1055 • The paper should point out any strong assumptions and how robust the results are to
1056 violations of these assumptions (e.g., independence assumptions, noiseless settings,
1057 model well-specification, asymptotic approximations only holding locally). The authors
1058 should reflect on how these assumptions might be violated in practice and what the
1059 implications would be.
- 1060 • The authors should reflect on the scope of the claims made, e.g., if the approach was
1061 only tested on a few datasets or with a few runs. In general, empirical results often
1062 depend on implicit assumptions, which should be articulated.
- 1063 • The authors should reflect on the factors that influence the performance of the approach.
1064 For example, a facial recognition algorithm may perform poorly when image resolution
1065 is low or images are taken in low lighting. Or a speech-to-text system might not be
1066 used reliably to provide closed captions for online lectures because it fails to handle
1067 technical jargon.
- 1068 • The authors should discuss the computational efficiency of the proposed algorithms
1069 and how they scale with dataset size.
- 1070 • If applicable, the authors should discuss possible limitations of their approach to
1071 address problems of privacy and fairness.
- 1072 • While the authors might fear that complete honesty about limitations might be used by
1073 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
1074 limitations that aren't acknowledged in the paper. The authors should use their best
1075 judgment and recognize that individual actions in favor of transparency play an impor-
1076 tant role in developing norms that preserve the integrity of the community. Reviewers
1077 will be specifically instructed to not penalize honesty concerning limitations.

1078 **3. Theory assumptions and proofs**

1079 Question: For each theoretical result, does the paper provide the full set of assumptions and
1080 a complete (and correct) proof?

1081 Answer: [NA]

1082 Justification: The benchmark work do not include theoretical results.

1083 Guidelines:

- 1084 • The answer NA means that the paper does not include theoretical results.
- 1085 • All the theorems, formulas, and proofs in the paper should be numbered and cross-
- 1086 referenced.
- 1087 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 1088 • The proofs can either appear in the main paper or the supplemental material, but if
- 1089 they appear in the supplemental material, the authors are encouraged to provide a short
- 1090 proof sketch to provide intuition.
- 1091 • Inversely, any informal proof provided in the core of the paper should be complemented
- 1092 by formal proofs provided in appendix or supplemental material.
- 1093 • Theorems and Lemmas that the proof relies upon should be properly referenced.

1094 4. Experimental result reproducibility

1095 Question: Does the paper fully disclose all the information needed to reproduce the main ex-

1096 perimental results of the paper to the extent that it affects the main claims and/or conclusions

1097 of the paper (regardless of whether the code and data are provided or not)?

1098 Answer: [Yes]

1099 Justification: Yes, the paper provides model names, versions, complete experimental setups,

1100 and includes the dataset and code for reproducibility.

1101 Guidelines:

- 1102 • The answer NA means that the paper does not include experiments.
- 1103 • If the paper includes experiments, a No answer to this question will not be perceived
- 1104 well by the reviewers: Making the paper reproducible is important, regardless of
- 1105 whether the code and data are provided or not.
- 1106 • If the contribution is a dataset and/or model, the authors should describe the steps taken
- 1107 to make their results reproducible or verifiable.
- 1108 • Depending on the contribution, reproducibility can be accomplished in various ways.
- 1109 For example, if the contribution is a novel architecture, describing the architecture fully
- 1110 might suffice, or if the contribution is a specific model and empirical evaluation, it may
- 1111 be necessary to either make it possible for others to replicate the model with the same
- 1112 dataset, or provide access to the model. In general, releasing code and data is often
- 1113 one good way to accomplish this, but reproducibility can also be provided via detailed
- 1114 instructions for how to replicate the results, access to a hosted model (e.g., in the case
- 1115 of a large language model), releasing of a model checkpoint, or other means that are
- 1116 appropriate to the research performed.
- 1117 • While NeurIPS does not require releasing code, the conference does require all submis-
- 1118 sions to provide some reasonable avenue for reproducibility, which may depend on the
- 1119 nature of the contribution. For example
 - 1120 (a) If the contribution is primarily a new algorithm, the paper should make it clear how
 - 1121 to reproduce that algorithm.
 - 1122 (b) If the contribution is primarily a new model architecture, the paper should describe
 - 1123 the architecture clearly and fully.
 - 1124 (c) If the contribution is a new model (e.g., a large language model), then there should
 - 1125 either be a way to access this model for reproducing the results or a way to reproduce
 - 1126 the model (e.g., with an open-source dataset or instructions for how to construct
 - 1127 the dataset).
 - 1128 (d) We recognize that reproducibility may be tricky in some cases, in which case
 - 1129 authors are welcome to describe the particular way they provide for reproducibility.
 - 1130 In the case of closed-source models, it may be that access to the model is limited in
 - 1131 some way (e.g., to registered users), but it should be possible for other researchers
 - 1132 to have some path to reproducing or verifying the results.

1133 5. Open access to data and code

1134 Question: Does the paper provide open access to the data and code, with sufficient instruc-
1135 tions to faithfully reproduce the main experimental results, as described in supplemental
1136 material?

1137 Answer: [Yes]

1138 Justification: Yes, the paper offers open access to data and code with detailed instructions in
1139 the supplemental material for reproducing results.

1140 Guidelines:

- 1141 • The answer NA means that paper does not include experiments requiring code.
- 1142 • Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- 1143 • While we encourage the release of code and data, we understand that this might not be
1144 possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not
1145 including code, unless this is central to the contribution (e.g., for a new open-source
1146 benchmark).
- 1147 • The instructions should contain the exact command and environment needed to run to
1148 reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- 1149 • The authors should provide instructions on data access and preparation, including how
1150 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 1151 • The authors should provide scripts to reproduce all experimental results for the new
1152 proposed method and baselines. If only a subset of experiments are reproducible, they
1153 should state which ones are omitted from the script and why.
- 1154 • At submission time, to preserve anonymity, the authors should release anonymized
1155 versions (if applicable).
- 1156 • Providing as much information as possible in supplemental material (appended to the
1157 paper) is recommended, but including URLs to data and code is permitted.

1158 6. Experimental setting/details

1159 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
1160 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
1161 results?

1162 Answer: [Yes]

1163 Justification: Yes, the paper details data statistics, data curation process and hyperparameters
1164 (e.g. temperature) to ensure result comprehension.

1165 Guidelines:

- 1166 • The answer NA means that the paper does not include experiments.
- 1167 • The experimental setting should be presented in the core of the paper to a level of detail
1168 that is necessary to appreciate the results and make sense of them.
- 1169 • The full details can be provided either with the code, in appendix, or as supplemental
1170 material.

1171 7. Experiment statistical significance

1172 Question: Does the paper report error bars suitably and correctly defined or other appropriate
1173 information about the statistical significance of the experiments?

1174 Answer: [No]

1175 Justification: No, error bars are not provided due to the inherent uncertainty in LLMs and
1176 LVLMs, making statistical significance challenging to report.

1177 Guidelines:

- 1178 • The answer NA means that the paper does not include experiments.
- 1179 • The authors should answer "Yes" if the results are accompanied by error bars, confi-
1180 dence intervals, or statistical significance tests, at least for the experiments that support
1181 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.

1235 Guidelines:

- 1236 • The answer NA means that there is no societal impact of the work performed.
- 1237 • If the authors answer NA or No, they should explain why their work has no societal
- 1238 impact or why the paper does not address societal impact.
- 1239 • Examples of negative societal impacts include potential malicious or unintended uses
- 1240 (e.g., disinformation, generating fake profiles, surveillance), fairness considerations
- 1241 (e.g., deployment of technologies that could make decisions that unfairly impact specific
- 1242 groups), privacy considerations, and security considerations.
- 1243 • The conference expects that many papers will be foundational research and not tied
- 1244 to particular applications, let alone deployments. However, if there is a direct path to
- 1245 any negative applications, the authors should point it out. For example, it is legitimate
- 1246 to point out that an improvement in the quality of generative models could be used to
- 1247 generate deepfakes for disinformation. On the other hand, it is not needed to point out
- 1248 that a generic algorithm for optimizing neural networks could enable people to train
- 1249 models that generate Deepfakes faster.
- 1250 • The authors should consider possible harms that could arise when the technology is
- 1251 being used as intended and functioning correctly, harms that could arise when the
- 1252 technology is being used as intended but gives incorrect results, and harms following
- 1253 from (intentional or unintentional) misuse of the technology.
- 1254 • If there are negative societal impacts, the authors could also discuss possible mitigation
- 1255 strategies (e.g., gated release of models, providing defenses in addition to attacks,
- 1256 mechanisms for monitoring misuse, mechanisms to monitor how a system learns from
- 1257 feedback over time, improving the efficiency and accessibility of ML).

1258 **11. Safeguards**

1259 Question: Does the paper describe safeguards that have been put in place for responsible
1260 release of data or models that have a high risk for misuse (e.g., pretrained language models,
1261 image generators, or scraped datasets)?

1262 Answer: [Yes]

1263 Justification: Yes, we have described how we avoid unsafe images during the data curation
1264 process.

1265 Guidelines:

- 1266 • The answer NA means that the paper poses no such risks.
- 1267 • Released models that have a high risk for misuse or dual-use should be released with
- 1268 necessary safeguards to allow for controlled use of the model, for example by requiring
- 1269 that users adhere to usage guidelines or restrictions to access the model or implementing
- 1270 safety filters.
- 1271 • Datasets that have been scraped from the Internet could pose safety risks. The authors
- 1272 should describe how they avoided releasing unsafe images.
- 1273 • We recognize that providing effective safeguards is challenging, and many papers do
- 1274 not require this, but we encourage authors to take this into account and make a best
- 1275 faith effort.

1276 **12. Licenses for existing assets**

1277 Question: Are the creators or original owners of assets (e.g., code, data, models), used in
1278 the paper, properly credited and are the license and terms of use explicitly mentioned and
1279 properly respected?

1280 Answer: [Yes]

1281 Justification: Yes, original dataset authors are credited via citations, and all public data usage
1282 complies with specified licenses and terms.

1283 Guidelines:

- 1284 • The answer NA means that the paper does not use existing assets.
- 1285 • The authors should cite the original paper that produced the code package or dataset.
- 1286 • The authors should state which version of the asset is used and, if possible, include a
- 1287 URL.

- 1288 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
1289 • For scraped data from a particular source (e.g., website), the copyright and terms of
1290 service of that source should be provided.
1291 • If assets are released, the license, copyright information, and terms of use in the
1292 package should be provided. For popular datasets, paperswithcode.com/datasets
1293 has curated licenses for some datasets. Their licensing guide can help determine the
1294 license of a dataset.
1295 • For existing datasets that are re-packaged, both the original license and the license of
1296 the derived asset (if it has changed) should be provided.
1297 • If this information is not available online, the authors are encouraged to reach out to
1298 the asset's creators.

1299 **13. New assets**

1300 Question: Are new assets introduced in the paper well documented and is the documentation
1301 provided alongside the assets?

1302 Answer: [Yes]

1303 Justification: Yes, new assets are thoroughly documented, with documentation provided
1304 alongside them.

1305 Guidelines:

- 1306 • The answer NA means that the paper does not release new assets.
1307 • Researchers should communicate the details of the dataset/code/model as part of their
1308 submissions via structured templates. This includes details about training, license,
1309 limitations, etc.
1310 • The paper should discuss whether and how consent was obtained from people whose
1311 asset is used.
1312 • At submission time, remember to anonymize your assets (if applicable). You can either
1313 create an anonymized URL or include an anonymized zip file.

1314 **14. Crowdsourcing and research with human subjects**

1315 Question: For crowdsourcing experiments and research with human subjects, does the paper
1316 include the full text of instructions given to participants and screenshots, if applicable, as
1317 well as details about compensation (if any)?

1318 Answer: [No]

1319 Justification: No, due to the offline nature of the process and privacy concerns, full instruc-
1320 tions and screenshots are not provided, but the process is compliant and adheres to scientific
1321 principles.

1322 Guidelines:

- 1323 • The answer NA means that the paper does not involve crowdsourcing nor research with
1324 human subjects.
1325 • Including this information in the supplemental material is fine, but if the main contribu-
1326 tion of the paper involves human subjects, then as much detail as possible should be
1327 included in the main paper.
1328 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
1329 or other labor should be paid at least the minimum wage in the country of the data
1330 collector.

1331 **15. Institutional review board (IRB) approvals or equivalent for research with human
1332 subjects**

1333 Question: Does the paper describe potential risks incurred by study participants, whether
1334 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
1335 approvals (or an equivalent approval/review based on the requirements of your country or
1336 institution) were obtained?

1337 Answer: [No]

1338 Justification: No, the experiments involve no risks to participants, and all necessary approvals
1339 were obtained.

1340 Guidelines:

- 1341 • The answer NA means that the paper does not involve crowdsourcing nor research with
1342 human subjects.
- 1343 • Depending on the country in which research is conducted, IRB approval (or equivalent)
1344 may be required for any human subjects research. If you obtained IRB approval, you
1345 should clearly state this in the paper.
- 1346 • We recognize that the procedures for this may vary significantly between institutions
1347 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
1348 guidelines for their institution.
- 1349 • For initial submissions, do not include any information that would break anonymity (if
1350 applicable), such as the institution conducting the review.

1351 **16. Declaration of LLM usage**

1352 Question: Does the paper describe the usage of LLMs if it is an important, original, or
1353 non-standard component of the core methods in this research? Note that if the LLM is used
1354 only for writing, editing, or formatting purposes and does not impact the core methodology,
1355 scientific rigorousness, or originality of the research, declaration is not required.

1356 Answer: [Yes]

1357 Justification: Yes, the use of LLM-as-a-Judge, a core and innovative method, is fully
1358 described in the paper. We ensure that all content is correct and original.

1359 Guidelines:

- 1360 • The answer NA means that the core method development in this research does not
1361 involve LLMs as any important, original, or non-standard components.
- 1362 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1363 for what should or should not be described.