HIDDEN MEANINGS IN PLAIN SIGHT: REBUSBENCH FOR EVALUATING COGNITIVE VISUAL REASONING

Seyed Amir Kasaei, Arash Marioriyad, Mahbod Khaleti,
MohammadAmin Fazli, Mahdieh Soleymani Baghshah & Mohammad Hossein Rohban
Department of Computer Engineering
Sharif University of Technology
a.kasaei@me.com, {arashmarioriyad, mahbod.kh2005}@gmail.com
{fazli, soleymani, rohban}@sharif.edu

ABSTRACT

Large Vision–Language Models (LVLMs) have achieved remarkable proficiency in explicit visual recognition, effectively describing what is directly visible in an image. However, a critical cognitive gap emerges when the visual input serves only as a clue rather than the answer. We identify that current models struggle with the complex, multi-step reasoning required to solve problems where information is not explicitly depicted. Successfully solving a rebus puzzle requires a distinct cognitive workflow: the model must extract visual and textual attributes, retrieve linguistic prior knowledge (such as idioms), and perform abstract mapping to synthesize these elements into a meaning that exists outside the pixel space. To evaluate this neurosymbolic capability, we introduce **RebusBench**, a benchmark of 1,164 puzzles designed to test this specific integration of perception and knowledge. Our evaluation of state-of-the-art models (including Qwen, InternVL, and LLaVA) shows a severe deficiency: performance saturates below 10% Exact Match and 20% semantic accuracy, with no significant improvement observed from model scaling or In-Context Learning (ICL). These findings suggest that while models possess the necessary visual and linguistic components, they lack the cognitive reasoning "glue" to connect them. The project page is available at this URL

1 INTRODUCTION

Visual reasoning extends beyond perception, requiring the synthesis of visual inputs with abstract knowledge. Large Vision–Language Models (LVLMs) have recently demonstrated remarkable progress in this domain. Systems ranging from foundational architectures like Flamingo and BLIP Alayrac et al. (2022); Li et al. (2022; 2023b) to recent instruction-tuned and scaled models such as LLaVA, InternVL, and Qwen-VL Liu et al. (2023b); Dai et al. (2023); Chen et al. (2024b); Bai et al. (2025b;a) have set new standards in visual question answering and spatial grounding. Together with proprietary models like GPT-5.2 OpenAI (2025) and Gemini 3 Google DeepMind (2025), these advancements suggest that modern LVLMs are increasingly capable of bridging the gap between pixel-level processing and semantic understanding.

The advancement of the field relies heavily on the quality of its benchmarks, yet existing evaluation suites often fail to probe the cognitive depth of these models. Standard datasets such as VQA v2, GQA, and CLEVR Antol et al. (2015); Hudson & Manning (2019); Johnson et al. (2017) predominantly assess referential grounding—the capacity to map linguistic tokens to explicit visual features. From a cognitive perspective, this mirrors rapid, "System 1" perceptual processing rather than deep, deliberate reasoning. These tasks rarely require the model to engage in modal entanglement, where visual and textual elements must be creatively recombined to form new concepts. Consequently, current benchmarks struggle to differentiate between shallow pattern matching and the multi-step, open-ended symbolic manipulation required for true human-level intelligence.

To this end, we introduce **RebusBench**, a novel benchmark based on rebus puzzles—visual riddles that serve as a rigorous testbed for deep, cognitively-inspired reasoning. Unlike standard tasks

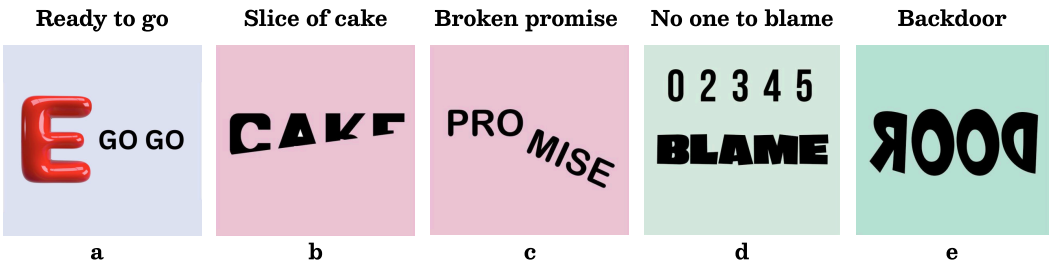

Figure 1: **Sample puzzles from RebusBench.** Solving these requires transcending literal perception ("System 1") to perform abstract "System 2" reasoning, where models must entangle visual cues (position, color, style) with linguistic knowledge to reconstruct the hidden idiom.

where the answer is explicitly present in the image, a rebus puzzle demands a constructive inference process. As illustrated in Figure 1, consider a puzzle displaying a red letter "E" alongside two instances of the word "GO". To arrive at the correct solution, "*Ready to go*," a model cannot merely describe the visible elements. It must execute a complex cognitive chain: (i) extract the visual attributes (color red, repetition of "GO"); (ii) retrieve external phonological and semantic knowledge (mapping "Red E" $\rightarrow$ "Ready" and "Two Gos" $\rightarrow$ "to go"); and (iii) synthesize these components into a coherent idiomatic phrase. This setting forces the model to move beyond simple perception and engage in visual–textual entanglement, distinguishing generic recognition from the ability to perform rigorous, multi-step symbolic reasoning.

Our empirical evaluation reveals a stark reality: state-of-the-art LVLMs struggle profoundly with this benchmark. Even the largest open-weight models, such as Qwen 2.5-32B and InternVL-30B, fail to surpass 10% in Exact Match (EM) accuracy. Crucially, standard mitigation strategies offer little relief; neither scaling model parameters nor increasing context through few-shot prompting yields significant improvements. Furthermore, even when evaluated by a lenient semantic judge (GPT-4o) to account for near-synonyms, performance plateaus below 20% across all settings. This resistance to scaling and In-Context Learning (ICL) underscores a fundamental limitation in current architectures: they lack the robust vision–text entanglement necessary for abstract reasoning, suggesting that raw compute and data are insufficient to bridge this specific cognitive gap.

2 Related Work

Scene Understanding and Image Captioning. Foundational datasets such as MS COCO Lin et al. (2014), Flickr30k Young et al. (2014), Visual Genome Krishna et al. (2017), and LVIS Gupta et al. (2019) emphasize literal mappings between visual regions and linguistic labels. While knowledge-augmented variants like OK-VQA Marino et al. (2019) incorporate external facts, they treat vision and text as parallel streams for retrieval rather than entangled components. These frameworks largely overlook scenarios where the visual rendering of text or the symbolic fusion of modalities necessitates non-literal interpretation and decoding of hidden semantic identities.

Spatial and Compositional Reasoning. Benchmarks including CLEVR Johnson et al. (2017), GQA Hudson & Manning (2019), VSR Liu et al. (2023a), SpatialVLM Chen et al. (2024a), Spatial4D-Bench Wang et al. (2026), iVISPAR Mayer et al. (2025), and SpatialBench Xu et al. (2025) evaluate a model's grasp of geometric coordinates and relative 3D positioning. However, these tasks typically view spatiality through a physical lens rather than a symbolic one. They fail to assess structural arrangement as a syntactic operator—where the placement of elements (e.g., vertical stacking or specific alignments) transforms individual components into a unified abstract concept rather than a mere physical description.

Visual Counting and Quantity Estimation. Tasks focused on enumeration, such as Count7W Chattopadhyay et al. (2017), TallyQA Acharya et al. (2019), HowManyQA Trott et al. (2020), and CAPTURe Pothiraj et al. (2025), measure precise object individuation. In these settings, cardinality is typically the terminal objective. Existing benchmarks lack requirements for multi-step synthesis,

where numerical results must serve as intermediate symbolic clues. Consequently, they do not test a model's ability to creatively integrate counts with other visual-textual cues to resolve open-ended or latent riddles.

Advanced Reasoning and Multi-discipline Benchmarks. Comprehensive evaluative sets like ScienceQA Lu et al. (2022), MMMU Yue et al. (2024), MathVista Lu et al. (2024), and SEED-Bench Li et al. (2023a) stress expert-level knowledge and formulaic deduction. While benchmarks in this category test general intelligence, they remain grounded in linear logic and factual recall. There remains a significant gap in evaluating lateral thinking and "out-of-the-box" creativity, particularly in resolving ill-posed visual problems that lack a direct, deductive path to a solution.

3 THE REBUSBENCH DATASET

To rigorously evaluate abstract visual reasoning, we introduce **RebusBench**, a dataset of **1,164 rebus puzzles** aggregated from diverse sources specializing in lateral thinking ESL Vault; Just Family Fun (2025). In contrast to synthetic VQA datasets that often rely on rigid, procedurally generated patterns, RebusBench features human-authored puzzles grounded in authentic idiomatic usage. We collected both the puzzle images and their intended solutions directly from mentioned sources. Subsequently, we applied strict manual verification to filter out ambiguities, ensuring that every retained instance possesses a unique, clear solution free from obscure cultural trivia or subjective interpretation.

A defining feature of RebusBench is that it eschews a fixed ratio of visual-to-textual reasoning in favor of a continuous cognitive spectrum. On one end, the dataset includes **visually dominant** instances driven by geometric deformation, where the physical manipulation of text constitutes the primary clue. For example, in the *"Slice of cake"* puzzle (Figure 1b), the word "CAKE" is visually rendered with a slice cut out of it. Solving this requires the model to treat the text as a malleable object rather than a fixed string, mapping the missing slice to the concept of "slicing." Conversely, the benchmark covers **symbolically dominant** puzzles that rely on abstract logical patterns and absence. In the *"No one to blame"* puzzle (Figure 1d), the image displays a numerical sequence "0, 2, 3..." alongside the word "BLAME." Here, the reasoning is driven by the specific absence of the number "1"; the model must detect this void, map it to the linguistic concept "No one," and concatenate it with the visible text. This diversity ensures that models are tested on their ability to dynamically weight visual attributes and symbolic sequences rather than relying on a single modality.

We formalize this task as open-ended generative reasoning. Given an image I, the model must generate the target idiom T. Theoretically, this requires the model to approximate a function $f(I) \rightarrow T$ that implicitly models a reasoning chain R, entangling visual attributes V (position, color, size) with linguistic priors L (phonology, idioms). Crucially, this is a *suppressive* task: to generate the correct target T (e.g., "Ready to go"), the model must suppress the high-probability literal caption (e.g., "A red letter E next to two Gos")—a hallmark of System 2 cognitive control.

Table 1: **Quantitative Performance on RebusBench.** We report Exact Match (EM) and GPT-4o Semantic Judge scores across One-shot and Three-shot settings. The results highlight a universal struggle across all architectures, where neither model scaling nor few-shot prompting yields significant improvements, confirming a fundamental gap in abstract visual reasoning.

MODEL	1-SHOT		3-SHOT	
	EM (% ↑)	GPT-4O (↑)	EM (% ↑)	GPT-4O (↑)
LLaVA-1.5 7B	1.20	0.1491	1.03	0.1381
InternVL 3.5 4B	4.73	0.1383	4.04	0.1272
InternVL 3.5 8B	4.81	0.1397	4.64	0.1295
InternVL 3.5 30B	5.15	0.1430	6.36	0.1462
Qwen 2.5 3B	3.52	0.1351	4.47	0.1340
Qwen 2.5 7B	5.15	0.1413	6.10	0.1496
Qwen 2.5 32B	7.39	0.1478	8.08	0.1508
Qwen 3 4B	4.12	0.1650	5.84	0.1638
Qwen 3 8B	6.36	0.1661	6.36	0.1704

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate a representative suite of open-weight LVLMs, varying in model families and parameter scales to benchmark current capabilities. Our evaluation includes LLaVA-1.5 7B Liu et al. (2023b), alongside the more recent InternVL 3.5 series (4B, 8B, 30B) Chen et al. (2024b), Qwen 2.5 series (3B, 7B, 32B) Bai et al. (2025b), and the Qwen 3 series (4B, 8B) Bai et al. (2025a).

Prompting Strategies. We assess performance across One-shot and Three-shot In-Context Learning (ICL) settings. In these few-shot configurations, we provide solved examples explained in the textual input prompt to guide the reasoning process. The full prompts for each configuration are provided in Appendix A.1.

Evaluation Metrics. We employ two complementary metrics. We report **Exact Match (EM)**, a strict metric where predictions are normalized (lowercased, stripped of whitespace/special characters) before comparison. To account for linguistic variations, we also use a **GPT-4o Judge (Semantic Score)** OpenAI (2024), which rates semantic similarity on a continuous scale from 0.0 to 1.0. The full evaluation prompt is detailed in Appendix A.2.

4.2 RESULTS

Overall Performance and Model Scaling. As summarized in Table 1, the evaluation reveals a distinct performance ceiling across all model families. We observe a universal failure to reliably solve rebus puzzles, with Exact Match (EM) scores saturating below 10% regardless of the architecture. Notably, increasing model parameters offers negligible returns. Transitions from smaller to significantly larger variants within the same family yield only marginal gains. For instance, scaling the InternVL 3.5 architecture from 4B to 30B only improves 1-shot performance from 4.73% to 5.15%. Similarly, the massive Qwen 2.5 32B model (7.39%) shows limited improvement over its 7B counterpart (5.15%) in the 1-shot setting. Even with 3-shot prompting, performance remains stagnant, with the best model achieving only 8.08%. This plateau suggests that simply adding capacity does not enable the models to spontaneously emerge the ability to bridge visual perception and idiomatic abstraction.

Impact of In-Context Learning. Furthermore, we find that In-Context Learning (ICL) remains insufficient for bridging this reasoning gap. Surprisingly, increasing the number of solved examples often yields negligible gains or even degrades performance. For instance, the InternVL 3.5 4B model drops from 4.73% (1-shot) to 4.04% (3-shot), while the Qwen 3 8B model remains exactly stagnant at 6.36% across both settings. Even the largest model, Qwen 2.5 32B, sees only a marginal increase from 7.39% to 8.08%. When evaluated by the more lenient GPT-4o semantic judge, performance remains universally low, peaking at only 0.1704 (Qwen 3 8B, 3-shot). This resistance to few-shot prompting demonstrates that current models cannot easily induce the necessary "System 2" logic from demonstrations alone.

5 CONCLUSION AND FUTURE WORK

Conclusion. We introduced **RebusBench** to evaluate the neurosymbolic reasoning of LVLMs. Our experiments reveal a critical performance ceiling: regardless of model scale or few-shot prompting, architectures consistently fail to surpass 10% Exact Match and 20% semantic accuracy. These results highlight a fundamental deficiency in current models, which struggle to entangle visual perception with linguistic abstraction—a necessary step for "System 2" reasoning.

Future Work. Future efforts will focus on expanding the dataset's diversity and adding fine-grained metadata to classify puzzles along the visual–textual spectrum. This will enable precise diagnosis of modality biases. Additionally, we plan to release a standardized evaluation pipeline and extend our benchmarking to include proprietary closed-source models, establishing a comprehensive baseline for the field.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pp. 8076–8084, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1135–1144, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

ESL Vault. Free printable rebus puzzles. https://eslvault.com/free-printable-rebus-puzzles/.

Google DeepMind. Gemini 3 flash: A state-of-the-art multimodal model, 2025. URL https://deepmind.google/technologies/gemini/. Accessed: 2026-02-06.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, 2017.

Just Family Fun. 100+ printable rebus puzzles with answers (2025 pdf). https://justfamilyfun.com/100-printable-rebus-puzzles-with-answers-2025/, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hane, Olga Russakovsky, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123:32–73, 2017.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Øyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

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

J. Mayer, M. Ballout, S. Jassim, F. N. Nezami, and E. Bruni. ivispar – an interactive visual-spatial reasoning benchmark for vlms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 26757–26781, 2025. URL https://arxiv.org/abs/2502.03214.

OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, May 2024. Accessed 2026-02-05.

OpenAI. Gpt-5.2: Advancements in multimodal reasoning and agentic intelligence, 2025. URL https://openai.com/index/gpt-5/. Accessed: 2026-02-06.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. *arXiv preprint arXiv:2504.15485*, 2025.

Alexander Trott, Caiming Xiong, and Richard Socher. Howmany-qa: Identifying and fixing bias in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Peng Wang, Jing Li, Yun Zhao, Hong Zhang, and Ming Liu. Spatial4d-bench: A versatile 4d spatial intelligence benchmark. *arXiv preprint arXiv:2601.00092*, 2026.

Peng Xu, Shuo Wang, Yuke Zhu, Jiacheng Li, and Yue Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over case frames. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

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

Appendix

A Prompting Strategy

A.1 Assessing the Model (LVLM Inference)

To benchmark the visual reasoning capabilities of the LVLM, we employ prompt variations aimed at solving the rebus puzzles. Specifically, we utilize **One-Shot & Three-Shot** settings where we provide one or three visual-text pairs as context. This approach evaluates the model's capacity for in-context learning, testing whether it can generalize the pattern of extracting semantic meaning from visual layout based on the provided examples.

LVLM Input: One-Shot Rebus Solver

```
"You are given an image that represents a rebus puzzle (a visual
word riddle).
A rebus puzzle encodes a common English word or phrase using
visual layout, repetition, color, position, or size of text and
symbols.
Do NOT read the image literally.
Instead, infer the hidden word or idiomatic expression suggested
by the visual arrangement.

Example:
- A red letter 'E' followed by 'GO GO' means 'ready to go'.

Question: What English word or phrase is represented?
Return ONLY the final answer in 1-5 words.
Do not explain."
```

LVLM Input: Three-Shot Rebus Solver

```
"You are given an image that represents a rebus puzzle (a visual
word riddle).
A rebus puzzle encodes a common English word or phrase using
visual layout, repetition, color, position, or size of text and
symbols.
Do NOT read the image literally.
Instead, infer the hidden word or idiomatic expression suggested
by the visual arrangement.

Examples:
- The word 'MAN' written three times means 'three men'.
- The word 'READ' placed inside a box means 'read between the
lines'.
- A red letter 'E' followed by 'GO GO' means 'ready to go'.

Question: What English word or phrase is represented?
Return ONLY the final answer in 1-5 words.
Do not explain."
```

A.2 Evaluating Model Output (LLM Judge)

Since rebus puzzles often have synonymous answers, strict string matching is insufficient. We utilize a text-only LLM as a semantic judge. The evaluation prompt inputs the `Ground Truth` and the LVLM's `Predicted` answer. The LLM is instructed to output a scalar score ranging from 0.0 to 1.0, penalizing unrelated answers while rewarding semantically correct interpretations.

LLM Evaluation: Rebus Judge

```
You are an expert evaluator for Rebus puzzles.
Your task is to compare a 'Ground Truth' answer with a 'Predicted'
answer.

Ground Truth:  "{ground_truth}"
Predicted:  "{prediction}"

Scoring Criteria:
- Score 1.0:  Perfect match or semantically identical (e.g.,
"Middle-aged" vs "middle aged", "Apple" vs "Apples").
- Score 0.0:  Completely unrelated.
- Otherwise:  Based on the level of capturing the core concept in
ground truth and partially correctness.

Return ONLY a single numerical float between 0.0 and 1.0.
No explanations, no text.
```