VisNumBench: Evaluating Number Sense of Multimodal Large Language Models

Tengjin Weng[1,2], Jingyi Wang[3], Wenhao Jiang[2]*, Zhong Ming[1,2,4]*
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[3]Shenzhen International Graduate School, Tsinghua University
[4]Shenzhen Technology University
wtjdsb@gmail.com, jingyi-w24@mails.tsinghua.edu.cn, cswhjiang@gmail.com, mingz@szu.edu.cn

Abstract

Can Multimodal Large Language Models (MLLMs) develop an intuitive number sense similar to humans? Targeting this problem, we introduce Visual Number Benchmark (VisNumBench) to evaluate the number sense abilities of MLLMs across a wide range of visual numerical tasks. VisNumBench consists of about $1,900$ *multiple-choice question-answer pairs derived from both synthetic and real-world visual data, covering seven visual numerical attributes and four types of visual numerical estimation tasks. Our experiments on VisNumBench led to the following key findings: (i) The 17 MLLMs we tested—including open-source models such as Qwen2.5-VL and InternVL2.5, as well as proprietary models like GPT-4o and Gemini 2.0 Flash—perform significantly below human levels in number sense-related tasks. (ii) Multimodal mathematical models and multimodal chain-of-thought (CoT) models did not exhibit significant improvements in number sense abilities. (iii) Stronger MLLMs with larger parameter sizes and broader general abilities demonstrate modest gains in number sense abilities. We believe VisNumBench will serve as a valuable resource for the research community, encouraging further advancements in enhancing MLLMs' number sense abilities. Code and dataset are available at https://wwwtttjjj.github.io/VisNumBench/.*

1. Introduction

Number sense is an innate cognitive ability shared by both humans and animals through the approximate number system [14]. It enables individuals to perceive, process, and manipulate numerical information intuitively. By fostering a deeper understanding of abstract number concepts, it facilitates the grasp of complex mathematical theories and their practical application in real-world scenarios. Figure 1 illustrates the human ability to perceive and estimate numerical quantities. For instance, a person can quickly recognize a group of five bananas and intuitively estimate that there are about two more groups of the same size. By leveraging this innate number sense and grouping strategy, one can infer that the total number of bananas is approximately fifteen.

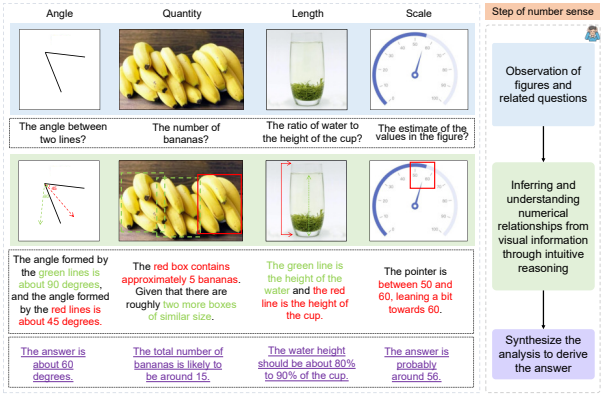

Figure 1. Explanations of number sense: how humans intuitively perceive and estimate values of angle, quantity, length, and scale.

Multimodal Large Language Models (MLLMs) have made remarkable strides in tackling complex multimodal tasks [2, 7, 24, 27]. Recent research has focused on enhancing their mathematical and scientific reasoning capabilities by incorporating external tools [31, 50]. To assess these abilities, numerous benchmarks [10, 18, 20, 29, 30, 32] have been developed to evaluate the performance of MLLMs on mathematical reasoning and numerical interpretation tasks. While existing benchmarks effectively assess structured numerical reasoning problems, they primarily emphasize abstract symbolic computation, mathematical problem-solving, or interpreting numerical data in textual contexts. However, these evaluations overlook a critical aspect of human-like numerical cognition: intuitive number sense.

*Corresponding author.

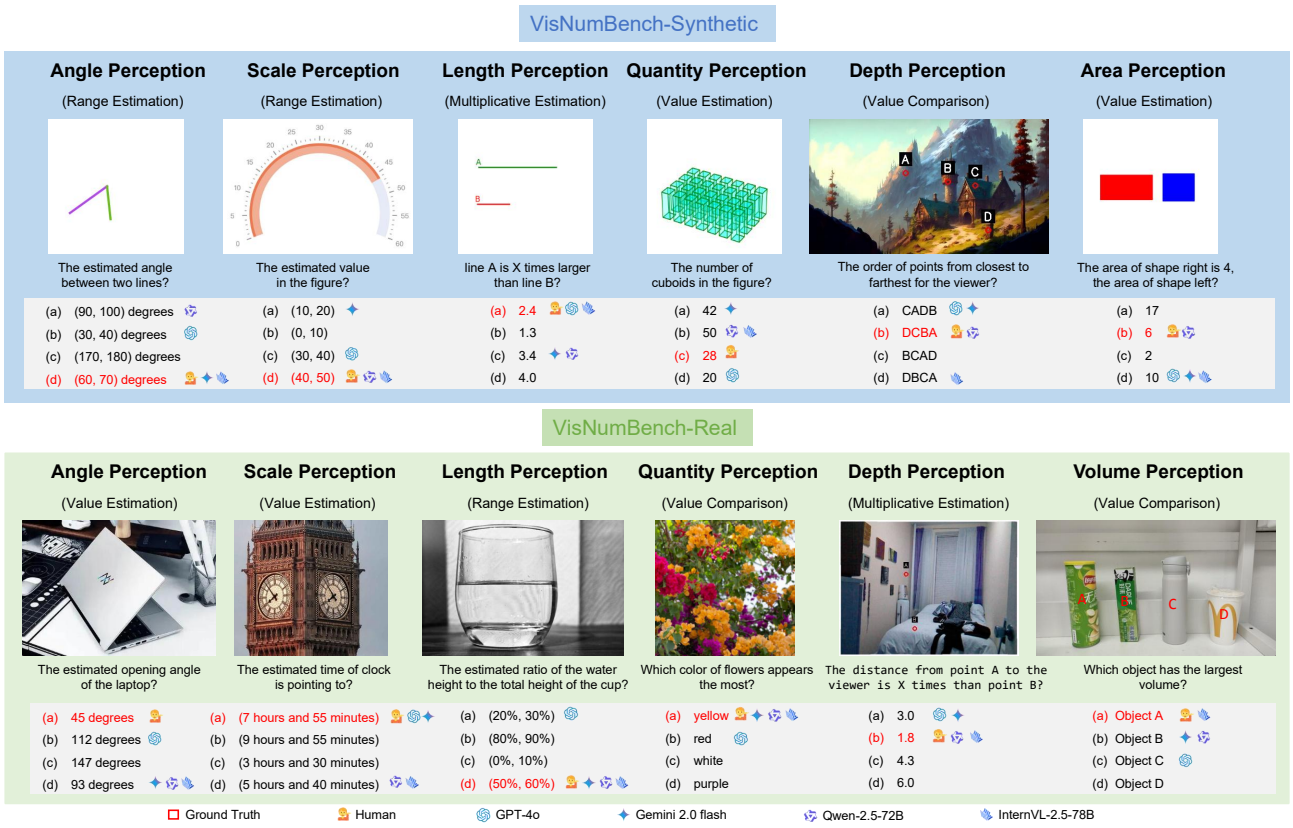

Figure 2. Examples from VisNumBench and responses from MLLMs. VisNumBench is divided into two subsets: VisNumBench-Synthetic and VisNumBench-Real. It focuses on seven key visual numerical attributes: angle, scale, length, quantity, depth, area, and volume. Even state-of-the-art MLLMs often struggle to answer the questions in VisNumBench accurately.

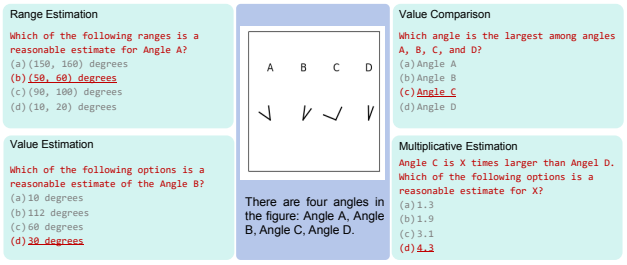

Figure 3. Illustrations of four distinct visual numerical estimation tasks: range estimation, value comparison, value estimation, and multiplicative estimation.

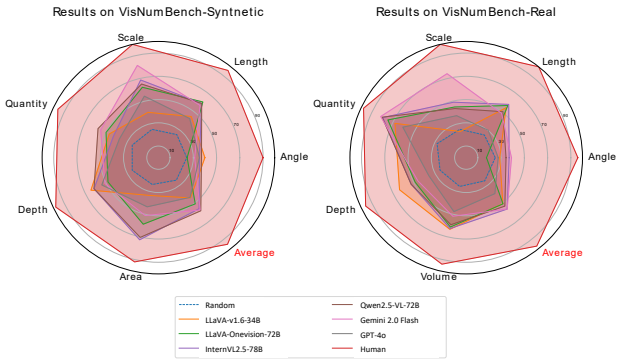

Figure 4. Evaluation results of MLLMs on the VisNumBench. The performance of MLLMs on VisNumBench is significantly poor in terms of accuracy, whereas human performance is nearly perfect.

Unlike humans, who effortlessly estimate quantities, perceive proportions, and grasp numerical relationships at a glance, MLLMs often depend on explicit reasoning steps rather than perceptual intuition. This limitation raises fundamental questions about whether current models genuinely comprehend numerical concepts or merely manipulate them based on learned patterns in text and images.

In this work, we introduce the Visual Number Benchmark (VisNumBench), inspired by human number sense abilities. As illustrated in Figure 2, VisNumBench is structured into two components based on different visual scenarios: VisNumBench-Synthetic and VisNumBench-Real. VisNumBench-Synthetic comprises controlled, synthetic images in which numerical relationships are explicitly defined. VisNumBench-Real contains real-world images, providing a more complex and less controlled environment. VisNumBench targets seven key dimensions of visual nu-

Table 1. Dataset statistics of VisNumBench based on various visual numerical attributes.

VisNumBench	Angle	Length	Scale	Quantity	Depth	Area	Volume	Total
VisNumBench-Synthetic	170	181	140	196	135	189	-	1011
VisNumBench-Real	149	162	143	147	154	-	147	902
Answer Format	4/5 options	3/4 options	4 options	3/4 options	4 options	4/5 options	3/4 options	3/4/5 options

merical attributes through four distinct types of visual numerical estimation tasks, as depicted in Figure 3.

We evaluated 17 MLLMs on VisNumBench and found that even state-of-the-art models perform poorly on our proposed benchmark, which is shown in Figure 4. Furthermore, our experiments reveal that adopting multimodal mathematical models and multimodal Chain-of-Thought (CoT) models did not lead to substantial performance improvements. However, the performance of the latest models is better than the previous models from the same family. For example, Qwen2.5VL [42] performs better than Qwen2VL [45]. It seems that optimization on data, training techniques, and model architecture will help models improve their number sense ability. In this work, we aim to advance MLLMs toward higher levels of intelligence by developing models that enhance visual number sense abilities. The main contributions of this paper are listed as follows:

1. We introduce VisNumBench, a comprehensive benchmark integrating diverse data sources and an automated evaluation framework to assess the numerical sense abilities of MLLMs across various visual numerical tasks.
2. We conduct a comprehensive evaluation of various MLLMs on VisNumBench, finding that even the most advanced models still exhibit limited numerical sense.
3. Further experiments on historical models from the same family show that their numerical sense abilities have improved over time. To enhance this ability within a short period, more specialized optimizations in data, training techniques, and model architecture may be required.

2. Related Work

2.1. Multimodal Large Language Models

Recent advancements in MLLMs have demonstrated exceptional capabilities across a wide range of tasks. Leveraging multimodal pre-training, MLLMs have achieved outstanding performance in both open-source models [1, 4, 5, 46, 52] and proprietary models [3, 35, 37]. Consequently, these models have been widely adopted in various domains, including mathematical reasoning [56], chart understanding [33], medical image analysis [23], and text-rich image comprehension [58]. Their growing success has spurred the development of an increasing number of benchmarks to assess performance across diverse visual and linguistic tasks.

2.2. Benchmarks for MLLMs

Advancements in MLLMs have led to the development of numerous benchmarks aimed at evaluating model performance across a broad spectrum of general multimodal tasks. Several recent studies [15, 16, 20–22, 28, 53, 54] have introduced more comprehensive reasoning and perception multimodal benchmarks that provide extensive and holistic assessments. Beyond general multimodal tasks, specialized benchmarks have been created to evaluate the mathematical reasoning capabilities of MLLMs. Tasks such as visual reasoning with numbers, arithmetic problem-solving, and algebraic manipulation play a crucial role in assessing both the numerical proficiency and higher-order cognitive abilities of MLLMs. Prominent benchmarks, including MATHVISTA [30], Math-Vision [44], MathOdyssey [13], and SMART-840 [9], are designed to test models on a diverse range of mathematical challenges, such as word problems, equation-solving, and complex multi-step reasoning.

These benchmarks aim to assess the ability of models to understand and process mathematical content in both images and text, as well as their ability to apply mathematical operations in reasoning contexts. However, evaluating MLLMs is crucial not only for measuring their proficiency in traditional mathematical reasoning but also for understanding their ability to handle more tangible, real-world dependent mathematical perception tasks. Humans typically acquire basic mathematical knowledge through intuitive number sense, which they then apply to real-world scenarios. This intuitive understanding of numbers and their relationships is a fundamental aspect of human cognition, enabling the seamless application of mathematical concepts in everyday life. While MLLMs can process complex mathematical problems, they may struggle with tasks that require this intuitive number sense. Therefore, existing benchmarks should not only evaluate the proficiency of models in traditional mathematical reasoning but also assess their ability to apply mathematical concepts in real-world contexts. This would help bridge the gap between abstract problem-solving and practical application.

2.3. Number Sense of MLLMs

In the context of MLLMs, previous research [11, 25, 47] has focused on tasks that rely on ordinal regression to evaluate number sense. Examples of such tasks include Age

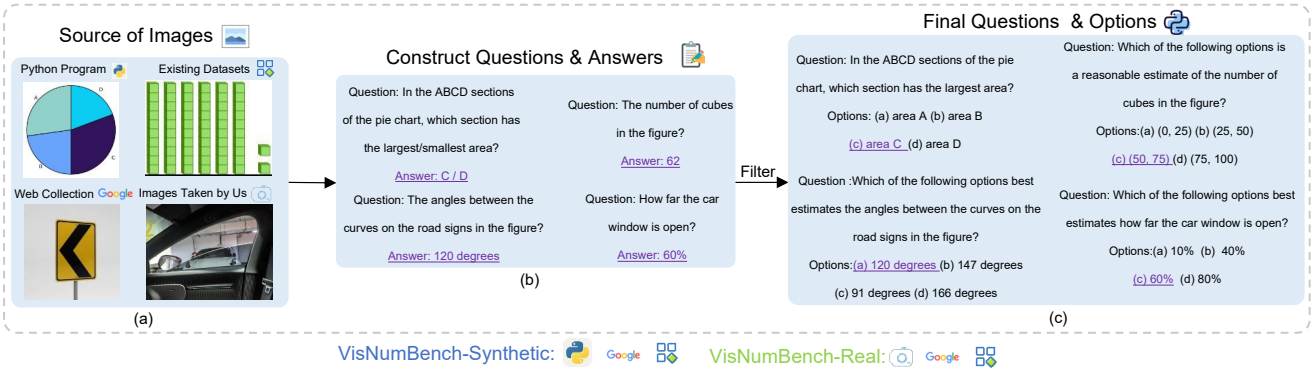

Figure 5. An illustration of the construction steps for images and questions in VisNumBench-Synthetic and VisNumBench-Real.

Estimation [38], Historical Image Dating [36], and Image Aesthetics Assessment [39]. These studies typically focus on estimating or ranking numerical attributes based on visual input, such as predicting the age of a person in an image or determining the historical context of a photograph. While these tasks assess the ability of the model to interpret numerical cues and sequences, they often focus on specific domains and may overlook the broader, more generalizable concept of number sense.

Assessing the number sense of MLLMs refers to evaluating their number sense abilities in a variety of scenarios. This includes tasks such as interpreting measurements, performing approximate calculations, comparing quantities, and identifying numerical relationships in different contexts. Such evaluations provide a better understanding of how models perform on tasks that require flexible and contextual reasoning about numbers while also enhancing their broad applicability and intelligence.

3. VisNumBench

3.1. Overview

We introduce VisNumBench, a benchmark specifically designed to directly evaluate the intuitive numerical abilities of MLLMs. Each instance in VisNumBench comprises a figure, a multiple-choice question, and a corresponding answer label. The dataset statistics of VisNumBench are presented in Table 1.

Compared with previous benchmarks, VisNumBench has the following novel features:

- **Comprehensive Scenario Integration:** VisNumBench incorporates both controlled synthetic figures and intricate real-world scenes, enabling a thorough evaluation of the number sense abilities of MLLMs.
- **Multidimensional Visual Numerical Attributes:** VisNumBench encompasses seven fundamental aspects of number sense—angle, length, scale, quantity, depth, area, and volume—ensuring a rigorous and comprehensive evaluation of the numerical capabilities of MLLMs.
- **Comprehensive Visual Numerical Estimation Tasks:** VisNumBench encompasses four distinct modes of visual numerical estimation—value comparison, value estimation, range estimation, and multiplicative estimation. These diverse tasks enable a thorough evaluation of MLLMs' ability to estimate numerical values across different visual numerical categories.

3.2. Data Collection Process

3.2.1. Source of Images

The development of VisNumBench required gathering figures from diverse sources, as depicted in part (a) of Figure 5.

- **Python Program.** We developed a series of Python scripts based on Matplotlib[1] to generate figures by randomly sampling parameters, which are also stored for future use. This approach allows precise control over various numerical properties, ensuring a well-balanced data distribution and minimizing potential biases.
- **Existing Datasets.** To enhance the diversity of the proposed benchmark and leverage existing high-quality data, we incorporated figures from multiple well-established datasets [16, 20, 30, 41, 48, 57], covering a broad range of numerical and spatial perception scenarios.
- **Web Collection.** To incorporate more natural and diverse visual data, we collected figures with numerical information from public web sources [12, 17, 43]. These figures were carefully curated and filtered to ensure relevance and clarity before designing the corresponding questions.
- **Images Taken by Us.** To better reflect real-world conditions, we manually constructed various number sense scenarios and captured images using a camera. These scenes encompassed real-world measurements, physical counting tasks, and estimation challenges under natural lighting and occlusion conditions.

3.2.2. QA Pair Construction and Quality Control

As shown in parts (b) and (c) in Figure 5, we involve QA pairs for images from different sources. For the figures gen-

[1] https://matplotlib.org/

Table 2. Accuracies of MLLMs on the VisNumBench-Synthetic (%) dataset. Dark gray and light gray indicate the best and the second best results among all models, respectively.

	Angle	Length	Scale	Quantity	Depth	Area	Average
Random	24.44	25.41	25.00	25.00	25.00	23.68	24.76
Open-source MLLMs							
Phi-3.5-vision	19.41	40.88	41.43	26.53	26.67	39.15	32.34
LLaVA-v1.5-7B	31.18	30.39	34.29	33.16	26.67	21.16	29.38
LLaVA-v1.5-13B	35.88	30.94	32.14	36.73	33.33	24.34	32.15
LLaVA-v1.6-34B	40.00	45.30	40.00	46.94	64.44	33.33	44.31
LLaVA-Onevision-7B	25.88	51.38	42.86	38.78	34.81	44.44	39.96
LLaVA-Onevision-72B	24.71	61.33	62.14	50.00	48.15	58.73	50.84
InternVL2.5-8B	26.47	41.99	49.29	34.69	41.48	46.03	39.66
InternVL2.5-38B	39.41	59.67	59.29	54.08	60.74	61.38	55.59
InternVL2.5-78B	35.29	59.67	68.57	42.86	61.48	72.49	56.18
Janus-Pro-7B	31.76	43.65	45.71	35.71	33.33	36.51	37.69
Qwen2.5-VL-3B	30.00	49.17	50.71	32.14	42.22	51.85	42.43
Qwen2.5-VL-7B	23.53	53.59	55.00	39.29	48.89	58.20	46.19
Qwen2.5-VL-72B	37.06	59.67	65.00	57.65	61.48	70.37	58.46
API-based models							
GPT-4o	35.29	43.09	54.29	37.24	54.07	43.39	43.72
Gemini 1.5 Flash	26.47	47.41	44.40	26.02	23.70	41.27	33.33
Gemini 2.0 Flash	31.18	57.46	81.43	55.10	51.11	70.90	57.57
Gemini 1.5 Pro	34.12	39.23	47.14	40.82	58.52	48.15	44.02
Human	90.00	96.00	100.00	96.00	98.00	92.00	95.33

erated by the Python program, we manually design different questions based on the specific characteristics of each figure and generate corresponding annotations using the parameters saved. We can design the question such as: "*In the ABCD sections of the pie chart, which section has the largest/smallest area?*" Based on the parameters saved, the answer would be "*C / D*". For images from other sources, we manually designed questions and annotated them based on different numerical attributes of the images. In addition, we can generate different numerical estimation tasks based on the answer type. For example, for the question: "*Which of the following options is a reasonable estimate of the number of cubes in the figure?"*, when the answer is "*(50, 75)"*, it corresponds to a range estimation task. On the contrary, if the answer is "*62"*, it belongs to the value estimation task.

We employ a combination of automated and manual methods to design distractors. For numerical answers, we generate alternative options that are easily confusable with the correct answer. Some distractors are constructed based on their inherent properties (e.g., areas A, B, and D in part (c) of Figure 5). These distractors are designed to align with human perceptual biases, appearing plausible yet distinguishable from the correct answer.

To ensure the high quality of VisNumBench, we meticulously reviewed all collected data and filtered out any ambiguous or unclear entries. More details of the data construction process can be found in Appendix B.

4. Experiments

We evaluate 17 well-known MLLMs from 8 model families, including 13 open-source models: Phi-3.5-vision [1], LLaVA-v1.5 (7B, 13B), and LLaVA-v1.6-34B [26], LLaVA-Onevision (7B, 72B) [23], Qwen2.5-VL (3B, 7B, 72B) [42], InternVL2.5 (8B, 38B, 78B) [5], and Janus-Pro-7B [4]. Additionally, we assess 4 proprietary models: GPT-4o [19], Gemini 1.5 Flash, Gemini 2.0 Flash, and Gemini 1.5 Pro [37].

We randomly selected 600 samples (50 QA pairs from each numerical attribute), with 300 sourced from VisNumBench-Synthetic and 300 from VisNumBench-Real. Human evaluators independently answered each question and provided assessments. Accuracy (%) is reported for all experimental results, and all the results are provided in Tables 2 and 3.

Table 3. Accuracies of MLLMs on the VisNumBench-Real (%) dataset. Dark gray and Light gray indicate the best and second-best results among all models, respectively.

	Angle	Length	Scale	Quantity	Depth	Volume	Average
Random	25.00	25.00	25.00	27.83	25.00	25.40	25.54
Open-source MLLMs							
Phi-3.5-vision	30.20	37.65	27.97	48.30	48.70	29.93	37.25
LLaVA-v1.5-7B	22.82	32.72	25.87	36.73	25.32	27.21	28.49
LLaVA-v1.5-13B	28.86	43.21	29.37	46.94	49.35	41.50	40.02
LLaVA-v1.6-34B	28.86	54.94	23.08	68.03	63.64	63.27	50.55
LLaVA-Onevision-7B	18.12	44.44	20.28	64.63	44.81	50.34	40.58
LLaVA-Onevision-72B	17.45	57.41	44.76	74.83	48.70	61.22	50.78
InternVL2.5-8B	28.86	34.57	15.38	64.63	49.35	47.62	40.13
InternVL2.5-38B	30.20	51.85	26.57	83.67	61.04	58.50	52.11
InternVL2.5-78B	36.91	58.64	48.95	79.59	52.60	62.59	56.54
Janus-Pro-7B	22.82	32.10	35.66	48.98	35.71	30.61	34.26
Qwen2.5-VL-3B	30.20	44.44	35.66	51.70	43.51	49.66	42.57
Qwen2.5-VL-7B	24.16	38.89	32.17	59.18	48.70	42.86	41.02
Qwen2.5-VL-72B	34.23	50.62	43.36	80.27	52.60	59.18	53.33
API-based models							
GPT-4o	27.52	30.25	37.06	60.54	35.71	47.62	39.58
Gemini 1.5 Flash	14.77	35.80	26.57	57.14	24.68	43.54	33.70
Gemini 2.0 Flash	38.93	48.77	74.14	81.63	46.10	51.70	56.54
Gemini 1.5 Pro	30.20	45.68	27.97	68.03	64.29	55.10	48.67
Human	96.00	100.00	100.00	98.00	96.00	94.00	97.33

4.1. Evaluation Results and Analysis

From Tables 2 and 3, we observe that the performance of MLLMs is not comparable to that of humans. Among the seven types of questions, quantity-related tasks appear to be the easiest, while angle-related tasks are the most difficult. This is possibly because the amount of training data available for quantity-related tasks is significantly greater than that for angle-related tasks. By comparing the evaluations on synthetic and real images, we find that the performance of the same model does not exhibit significant variance. Thus, in terms of numerical reasoning ability, both synthetic and real images present similar challenges for existing MLLMs. Moreover, we observe that the best open-sourced model performs comparably to the best closed-source models.

More detailed analyses and discussions are provided in the following subsections.

4.1.1. Performance on VisNumBench-Synthetic

Table 2 presents the results for various MLLMs on the VisNumBench-Synthetic dataset. Among the open-source models, Qwen2.5-VL-72B achieves the best performance, with an average accuracy of 58.46%. InternVL2.5-38B, InternVL2.5-78B, and LLaVA-v1.6-34B also demonstrate strong performance, each achieving either the best or the second-best accuracy in at least two tasks. LLaVA-v1.6-34B attains the highest accuracy in angle and depth estimation; however, its overall average accuracy is only 44.31%. LLaVA-Onevision-72B also performs well, achieving the highest accuracy in length estimation at 61.33%. In general, models with larger parameter sizes tend to exhibit superior performance, aligning with the intuition that larger models can better capture complex numerical relationships and fine-grained visual patterns.

In the API-based models, Gemini 2.0 Flash demonstrates the best performance, achieving an average accuracy of 57.57%. In contrast, GPT-4o and Gemini 1.5 Pro exhibit comparable performance, albeit with lower average accuracies. Gemini 1.5 Flash yields the weakest performance, with an average accuracy of 33.33%. Notably, certain open-source models perform on par with or even surpass proprietary models, suggesting that the disparity in numerical reasoning capabilities between open-source and closed-source models is minimal.

4.1.2. Performance on VisNumBench-Real

Accuracy on the VisNumBench-Real dataset shows similar trends. InternVL2.5-78B and Gemini 2.0 Flash stand

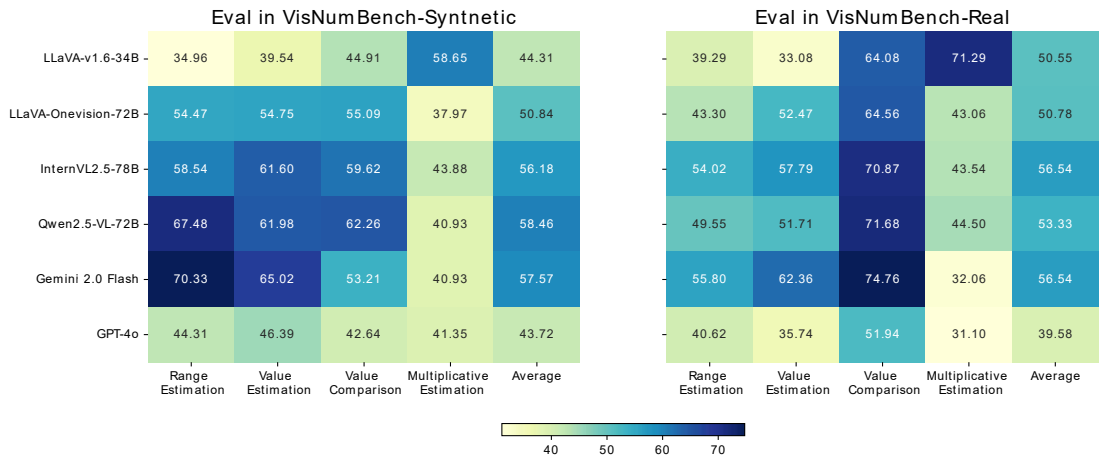

Figure 6. Confusion matrices (%) of MLLMs on the VisNumBench-Synthetic and VisNumBench-Real datasets across different visual numerical estimation tasks.

out with an average accuracy of 56.54%, achieving near-optimal results across multiple tasks, as shown in Table 3. InternVL2.5-38B attained an exceptionally high accuracy of 83.67% on the quantity task, while LLaVA-v1.6-34B excelled in the volume task, achieving the highest scores. Qwen2.5-VL-72B demonstrated relatively balanced performance, yielding a suboptimal average accuracy of 53.33%.

Surprisingly, Gemini 1.5 Pro achieved the highest accuracy in depth estimation, reaching 64.29%. However, its overall average accuracy remained unsatisfactory. Other proprietary models, such as GPT-4o and Gemini 1.5 Flash, exhibited relatively weaker performance. In general, the accuracy on VisNumBench-Real is lower than that on VisNumBench-Synthetic, likely due to the increased complexity and variability of real-world images.

4.1.3. Performance on Different Visual Numerical Estimation Tasks

As we analyze performance across different numerical estimation tasks, Figure 6 reveals that in the synthetic scenario, Gemini 2.0 Flash and Qwen2.5-VL-72B achieve the highest performance across all visual numerical estimation tasks, particularly in range estimation, value estimation, and value comparison, where their accuracies consistently exceed 60%. In contrast, GPT-4o exhibits the lowest performance in all tasks, especially in value comparison and multiplicative estimation. In the real-world scenario, most models achieved their best performance in value comparison tasks, which are also the easiest for humans. Although Gemini 2.0 Flash and InternVL2.5-78B continue to perform well in most tasks, their performance in multiplicative estimation has declined compared to the synthetic scenario. Additionally, GPT-4o continues to perform poorly across all tasks, particularly in multiplicative estimation and value comparison, where it falls significantly behind other models.

Notably, in the multiplicative estimation task, LLaVA-v1.6-34B outperforms all other models by a significant margin. This suggests that certain models may be more specialized for specific types of tasks, and further fine-tuning could enhance performance across different tasks.

4.2. Further Analysis

How do math-special models perform on the VisNumBench? To investigate the number sense abilities of math-special models, we introduce two multimodal mathematical models: (1) InternVL2-8B-MPO [49], initialized from InternVL2-8B [8] and fine-tuned on the large-scale multimodal reasoning preference dataset MMPR [49], achieving an accuracy of 65.65% on MathVista; (2) Math-LLaVA-13B [40], initialized from LLaVA-v1.5-13B and fine-tuned on the MathV360K [40] dataset. As shown in Figure 7, InternVL2-8B-MPO achieved a 1.0% improvement in synthetic scenarios and a 0.3% increase in real-world scenarios. Its enhancements are task-specific rather than universally effective across different number sense challenges. In contrast, Math-LLaVA-13B exhibited a polarized performance trend: while it improved by 3.7% on the synthetic dataset, its accuracy declined by 6.9% in real-world scenarios. This suggests that although the model benefits from training on synthetic data, it struggles to generalize to the complexity and variability of real-world number sense tasks. Relying solely on synthetic data may be insufficient to enhance number sense capabilities in real-world applications. Additional strategies, such as incorporating more diverse real-world training data or refining model architectures, may be necessary to achieve meaningful improvements.

How do the multimodal reasoning models perform? To examine whether reasoning techniques can enhance

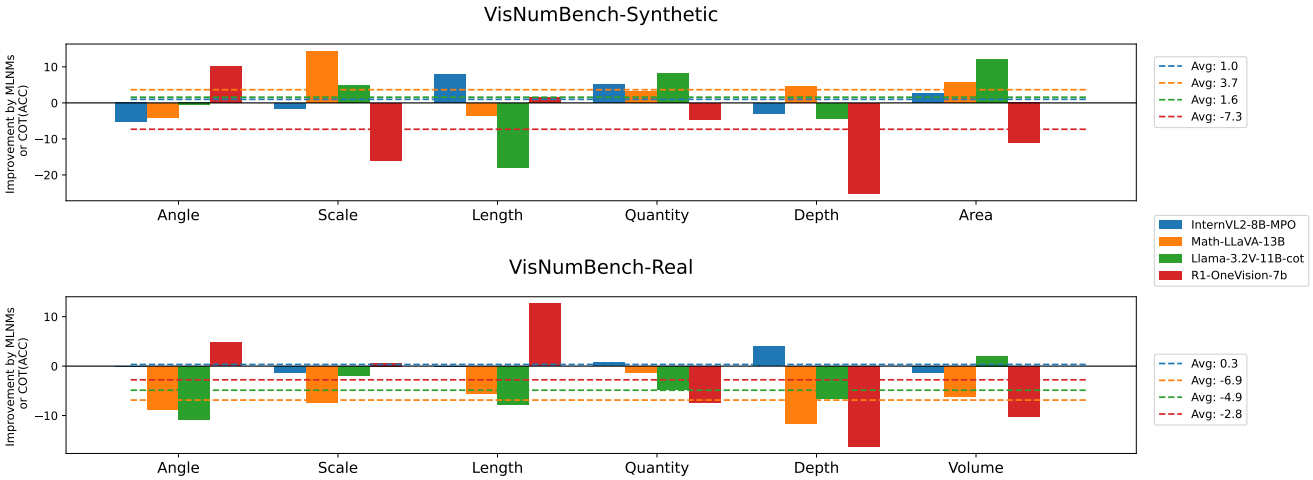

Figure 7. Improvements brought by multimodal mathematical models (InternVL2-8B-MPO and Math-LLaVA-13B) and multimodal CoT models (Llama-3.2V-11B-cot and R1-OneVision-7B). Table 9 and Table 10 in the appendix provide detailed results.

Table 4. Comparisons of the performance of models from the Qwen-VL family and the InternVL family in synthetic and real-world scenes. Table 11 in the appendix provides detailed results.

	Average (Synthetic)	Average (Real)
Qwen2-VL-2B	31.85	24.94
Qwen2.5-VL-3B	42.24(↑ +10.39)	42.57(↑ +17.63)
Qwen2-VL-7B	41.25	41.91
Qwen2.5-VL-7B	46.19(↑ +4.94)	41.02(↓ −0.89)
Qwen2-VL-72B	54.20	46.56
Qwen2.5-VL-72B	58.46(↑ +4.26)	53.33(↑ +6.77)
InternVL2-8B	39.56	39.58
InternVL2.5-8B	39.66(↑ +0.10)	40.13(↑ +0.55)
InternVL2-40B	45.50	45.12
InternVL2.5-38B	55.59(↑ +10.09)	52.11(↑ +6.99)

the number sense abilities of MLLMs, we evaluate two multimodal reasoning models: Llama-3.2V-11B-cot[2] [51] and R1-OneVision-7B[3]. Llama-3.2V-11B-cot is trained using LLaVA-o1-100k [51], achieving a 6.2% performance improvement on MathVista compared to Llama-3.2-11B-Vision-Instruct [34]. R1-OneVision-7B, trained with a rule-based reinforcement learning technique, attains an accuracy of 44.06% on Mathverse [55]. Accordingly, we assess these models on our benchmark. The results, presented in Figure 7, indicate that neither Llama-3.2V-11B-cot nor R1-OneVision-7B achieved the expected performance gains. On the contrary, their accuracy dropped significantly—except for a modest 1.6% improvement by Llama-3.2V-11B-cot in synthetic scenarios—especially in real-world settings. These findings suggest that developing reasoning techniques specifically tailored for number sense abilities may be necessary.

What helps improve the performance? To determine the factors contributing to the improvement of number sense ability in MLLMs, we evaluate historical models from the same family over time, specifically the Qwen-VL family and the InternVL family. The results are presented in Table 4. As observed, the performance of the latest models generally surpasses that of their predecessors. By comparing Qwen2-VL [45] with Qwen2.5-VL [42], as well as InternVL2 with InternVL2.5 [6], we observe improvements in several aspects: (1) data scale and quality, (2) a more powerful encoder, (3) model architecture, and (4) training strategy. These findings suggest that further exploration in these directions is essential for enhancing the number sense abilities of MLLMs.

5. Conclusion

In this work, we introduce VisNumBench, a novel benchmark designed to evaluate MLLMs on core number sense abilities that are inadequately addressed by existing evaluation benchmarks. Our assessment of 17 MLLMs uncovers substantial deficiencies in their capacity to demonstrate human-like number sense. Even the most advanced models still demonstrate limited numerical sense abilities. Further experiments on historical models from the same family show that to enhance this ability within a short period, more specialized optimizations in data, training techniques, and model architecture may be required.

[2] https://huggingface.co/Xkev/Llama-3.2V-11B-cot
[3] https://github.com/Fancy-MLLM/R1-Onevision

6. Acknowledge

This research was supported by the National Natural Science Foundation of China No.62272315.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 3, 5

[2] 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. 1

[3] AI Anthropic. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3, 2024. 3

[4] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. 3, 5

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3, 5

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 8

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 1

[8] 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*, pages 24185–24198, 2024. 7

[9] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Joanna Matthiesen, Kevin Smith, and Joshua B Tenenbaum. Evaluating large vision-and-language models on children's mathematical olympiads. *arXiv preprint arXiv:2406.15736*, 2024. 3

[10] Adam Dahlgren Lindström and Savitha Sam Abraham. CLEVR-Math: A dataset for compositional language, visual and mathematical reasoning. In *16th International Workshop on Neural-Symbolic Learning and Reasoning, NeSy 2022, Windsor, UK, september 28-30, 2022*. CEUR-WS, 2022. 1

[11] Yao Du, Qiang Zhai, Weihang Dai, and Xiaomeng Li. Teach clip to develop a number sense for ordinal regression. *arXiv preprint arXiv:2408.03574*, 2024. 3

[12] echarts. echarts. https://echarts.apache.org/zh/index.html. Accessed: 2025-01-26. 4, 1

[13] Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *arXiv preprint arXiv:2406.18321*, 2024. 3

[14] Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. Core systems of number. *Trends in cognitive sciences*, 8(7): 307–314, 2004. 1

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3

[16] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2025. 3, 4, 1, 2

[17] Google. Google images. https://images.google.com/. Accessed: 2025-01-26. 4, 1

[18] He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025. 1

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[20] Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. Visonlyqa: Large vision language models still struggle with visual perception of geometric information. *arXiv preprint arXiv:2412.00947*, 2024. 1, 3, 4

[21] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023.

[22] 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*, 2023. 3

[23] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 5

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[25] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. *Advances*

in Neural Information Processing Systems, 35:35313–35325, 2022. 3

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 5

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[28] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 3

[29] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. 1

[30] 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. *arXiv preprint arXiv:2310.02255*, 2023. 1, 3, 4

[31] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *The 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1

[32] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022. 1

[33] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. 3

[34] Meta AI. Llama 3.2 vision instruct (11b). https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct, 2024. Accessed: 2025-07. 8

[35] OpenAI. GPT-4V(ision) system card, 2023. 3

[36] Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 499–512. Springer, 2012. 4

[37] Sundar Pichai and Demis Hassabis. Our next-generation model: Gemini 1.5. ai, 2024. 3, 5

[38] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 341–345. IEEE, 2006. 4

[39] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *Proceedings of the international AAAI conference on web and social media*, pages 397–406, 2015. 4

[40] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 7

[41] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 4, 1

[42] Qwen Team. Qwen2.5-vl, 2025. 3, 5, 8

[43] WallpapersCraft. Desktop wallpapers hd, free desktop backgrounds. https://wallpaperscraft.com/. Accessed: 2025-01-26. 4, 1

[44] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024. 3

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 8

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[47] Rui Wang, Peipei Li, Huaibo Huang, Chunshui Cao, Ran He, and Zhaofeng He. Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. *Advances in Neural Information Processing Systems*, 36, 2023. 3

[48] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 4, 1

[49] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 7

[50] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023. 1

[51] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 8

[52] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 3

[53] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024. 3

[54] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 3

[55] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 8

[56] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *CoRR*, 2024. 3

[57] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 4, 1

[58] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 3