The Percept-V Challenge: Can Multimodal LLMs Crack Simple Perception Problems?

Anonymous ACL submission

Abstract

Cognitive science research treats visual perception, the ability to understand and make sense of a visual input, as one of the early developmental signs of intelligence. Its TVPS-4 framework categorizes and tests human perception into seven skills such as visual discrimination, and form constancy. Do Multimodal Large Language Models (MLLMs) match up to humans in basic perception? Even though there are many benchmarks that evaluate MLLMs on advanced reasoning and knowledge skills, there is limited research that focuses evaluation on simple perception. In response, we introduce Percept-V, a dataset containing 6000 program-generated uncontaminated images divided into 30 domains, where each domain tests one or more TVPS-4 skills. Our focus is on perception, so we make our domains quite simple and the reasoning and knowledge required for solving them are minimal. Since modern-day MLLMs can solve much more complex tasks, our a-priori expectation is that they will solve these domains very easily. Contrary to our belief, our experiments show a weak performance of SoTA proprietary and open-source MLLMs compared to very high human performance on Percept-V. We find that as number of objects in the image increases, performance goes down rather fast. Our experiments also identify the perception skills that are considerably harder for all models.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated high performance on complex tasks, leading some researchers to suggest that these models are approaching human-level intelligence (Zhou et al., 2023; King, 2023) or even achieving MENSA-level capabilities (Schregel, 2020). However, current visual benchmarks typically assess image understanding alongside reasoning and specialized domain knowledge in fields such as engineering, medicine, and art (Nguyen, 2023; Yue et al., 2024). This integrated approach makes it challenging to isolate MLLMs' pure image understanding capabilities from their reasoning abilities and domain-specific knowledge. To our knowledge, there is little work that focuses specifically on visual perception, the fundamental ability to understand and interpret visual images.

Visual perception represents a foundational cognitive ability that develops early in human cognition, preceding the acquisition of complex reasoning and knowledge-based skills (Rabindran and Madanagopal, 2020). Recent research has raised concerns about MLLMs' perceptual capabilities (Zhang et al., 2024b; Fu et al., 2024), highlighting the need for a dedicated perception benchmark. To address this gap, we introduce Percept-V, a perception-focused dataset comprising 6000 program-generated, uncontaminated image-based problems from 30 domains, which primarily evaluate visual understanding while requiring minimal problem-solving skills and no specialized domain knowledge. Sample questions are presented in Figure 1.

Our dataset construction draws from cognitive science literature, specifically utilizing the skill classification framework from TVPS-4 (Test of Visual Perceptual Skills) (Martin and Gardner, 2006), which is widely used for human perception assessment. TVPS-4 organizes visual perceptual abilities into seven distinct categories: visual discrimination, visual memory, visual sequential memory, visual-spatial relationships, figure-ground perception, form constancy, and visual closure (detailed in Section 3).

Percept-V comprises 30 domains, each designed to evaluate a combination of 1-3 TVPS-4 skill categories. Every domain features a single standardized question prompt that describes the specific task requirements. To systematically assess performance across varying levels of difficulty, each domain contains 200 images with varying problem size, which

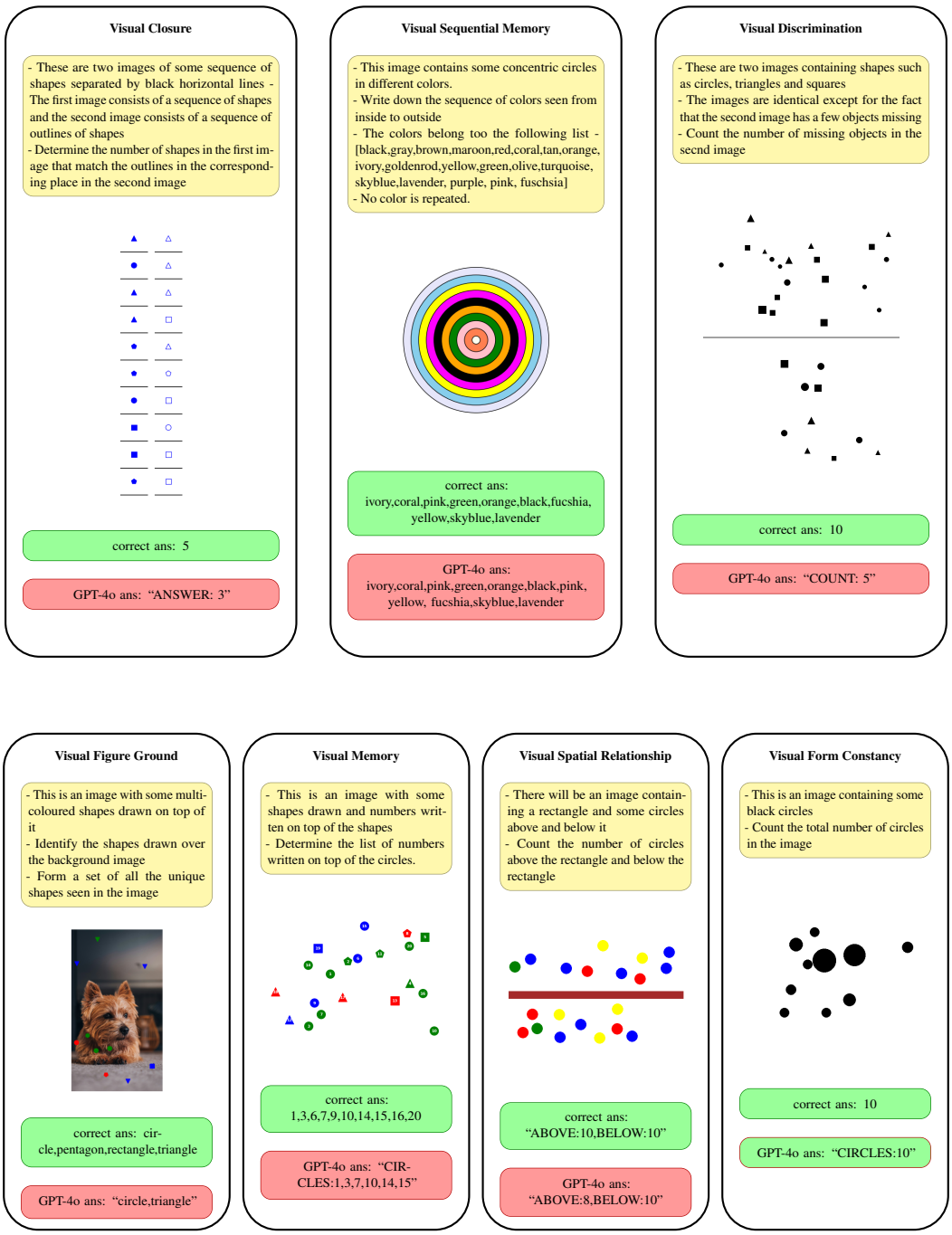

Figure 1: Sample questions from the Percept-V dataset illustrating tasks related to different visual perception skills as defined by the TVPS-4 framework. Each example shows the image prompt, the correct answer, and a typical incorrect response from an MLLM.

Skill	Domain
Visual Discrimination (D)	change colour, vanishing objects, comparing size, counting shapes, sort lines, sort circles, locate circles colour, locate circles shape, match shadow, match outline
Visual Memory (M)	list colours, list shapes, circle boxes, colours present, count coloured circles, numbered shapes, graph counting, sort circles, sort lines, counting shapes, identifying shapes
Visual Sequential Memory (SM)	grid path, layered colours, layered shapes, comparing size, match outline, match shadow
Visual Figure Ground (FG)	list colours, list shapes, locate circles colour, locate circles shape, layered shapes, layered colours
Visual Form Constancy (FC)	mirror image, water image, counting circles, counting shapes, numbered shapes
Visual Closure (C)	match outline, match shadow, layered shapes, layered colours
Visual Spatial Relationship (SR)	comparing size, circle location, circle right triangle, counting locations, cross and knots, inside circles, maze solving

Table 1: Domains using each skill. Note that a single domain can appear in multiple rows in the above table since multiple skills may be required to solve the its Problem Instances.

is operationalised through parameters such as the number of objects present, grid dimensions, or the number of processing steps required to complete the task. This structured approach enables us to evaluate MLLM performance as a function of visual complexity and provides insight into how they handle increasing perceptual demands.

We experiment on three state-of-the-art proprietary MLLMs – GPT-4o, O4-mini and Gemini 2.5 Flash and two open source MLLMs – Qwen 2.5 VL Instruct and DeepSeek VL2 Tiny, which include both language and reasoning models, and also perform a human study. Our findings reveal substantial gaps between MLLM and human performance on these simple visual perception tasks. Across all evaluated models, we observe consistent performance degradation as problem size increases, suggesting systematic limitations in handling visual complexity. While some models demonstrate relative strength in specific skill categories, overall performance on a skill remains remarkably similar across models, indicating broad deficits in visual understanding. Overall, our work is an important step towards understanding the visual perception limitations that must be addressed in the future to develop more robust MLLMs.

2 Related Work

Most exist benchmarks for evaluating the performance of MLLMs necessitate a combination of knowledge, reasoning and perception for good performance. The NTSEBench dataset (Pandya et al., 2024) comprises questions from the commonsense reasoning test from India. The MathVista dataset (Lu et al., 2023) examines different SoTA models on fine-grained, deep visual understanding and compositional reasoning. The MathVision benchmark (Wang et al., 2024) tests MLLMs on questions with a visual context with 16 distinct mathematical disciplines and 5 levels of difficulty.

OlympiadBench(He et al., 2024) collects advanced reasoning questions from the fields of mathematics, physics and chemistry to assess the complex reasoning abilities of MLLMs. Works like MMMU (Yue et al., 2024) evaluate MLLMs on many disciplines such as engineering, medicine, humanities, science, business and art. An extension of MMMU in Chinese called CMMMU (Zhang et al., 2024a) additionally tests the language understanding skills of MLLMs. SciBench (Wang et al., 2023) contains college-level science questions. Some datasets require interpretation of abstract figures, geometry shapes and scientific plots of ArXiv papers (Li et al., 2024), and experiments suggest that MLLMs could likely be used as scientific assistants in the future. EMMA (Hao et al., 2025) is a recent work requiring cross-modal reasoning for answering problems in domains such as mathematics, physics, chemistry and coding. All these datasets, along with perception, require complex

Type	Description	# Domains
Single	Boolean / Numeric	11
Fixed Length	A fixed length list	6
List	An ordered list	8
Set	An unordered set	5

Table 2: Question prompt types based on the expected answer format, see example 1

reasoning skills to answer questions, and several datasets additionally require specialized knowledge of a subject.

Some recent benchmarks specifically try to carefully disentangle the reasoning and perceptual abilities of LLMs. (Chung et al., 2025) propose MatHLENS which separately tests the perceptual, reasoning and integrating capabilites of MLLMs. In recent times, additional studies and benchmarks have been proposed that specifically highlight the lack of MLLMs' perceptual abilities, such as Lee et al. (2025). Zhang et al. (2024b) shows that MLLMs perform poorly in small object and low quality object recognition. Fu et al. (2024) develops a benchmark of simple perceptual questions that can be hard for MLLMs. Most of these works use complex image scenes for evaluation and do not study the impact of problem size on perception tasks.

We construct Percept-V to fill this gap – the questions in our dataset require relatively simple reasoning and no specialized knowledge, thus better evaluating perception performance of MLLMs. Our dataset is automatically generated thus avoiding the issues of contamination, and allows for explicitly testing against the variation in problem sizes in a given domain.

Most related to our work is a very recent unpublished paper (Kanade and Ganu, 2025), which also develops a dataset in the same spirit as ours. While it is contemporaneous research, our dataset is much larger – 30 domains, compared to 12, and allows for a more fine-grained variation along problem size, making our evaluation more extensive. We ground each domain into the specific skills from TVPS-4 framework covering all 7 skills, whereas they work with only a subset of them. 43% of our domains require multiple TVPS-4 skills, which allows us to assess MLLM's ability when applying skills in concert. Their work does not explicitly address the issue of handling multiple skills in a single domain.

3 Dataset

We now describe the details of our dataset referred to as Percept-V. Our dataset is composed of 30 different domains where each domain tests one or more *basic perception* skills from the TVPS-4 cognitive framework. The skills listed in the TVPS-4 framework are as follows:

1. *Visual Discrimination*: The ability to determine differences or similarities in objects based on size, colour, shape, etc.

2. *Visual Memory*: The ability to recall visual traits of a form or object.

3. *Visual Sequential Memory*: The ability to recall a sequence of objects in the correct order.

4. *Visual Spatial Relationships*: - Understanding the relationships of objects within the environment.

5. *Visual Figure Ground*: The ability to locate something in a busy background.

6. *Visual Form Constancy*: The ability to know that a form or shape is the same, even if it has been made smaller/larger or has been turned around.

7. *Visual Closure*: The ability to recognize a form or object when part of the picture is missing.

Each domain in our dataset consists of multiple *problem instances* which require a subset of TVPS-4 skills (Martin and Gardner, 2006) for them to be solved correctly. Table 1 lists the domains for each skill, which require the knowledge of that skill in order for its problem instances to be solved. In total, 17 domains test one skill, 8 domains have two skills, and 5 domains test 3 skills in concert.

Each domain is associated with a unique question prompt. Further, each problem instance in a domain is composed of a pair (Q, I), where Q represents the question prompt associated with the domain, and I is an image generated automatically through Python programmes that used libraries like Matplotlib, Pillow and OpenCV.[1]. In other words, we apply the same generic question prompt to different images resulting in different

[1] image generation for each domain is controlled by a set of hyperparameters described in A

problem instances. The problem instances are categorized across multiple sizes, where size of a problem instance captures its inherent complexity, such as number of objects or number of distractors, or number of sequence of steps required to solve the problem instance. For example, in the $counting_circles$ domain shown under Visual Form Constancy in 1, the problem size is controlled by controlling the number of circles to be counted. On the other hand, in the $numbered_shapes$ domain under Visual Memory in 1, with increasing problem size, the number of circles as well as distractors i.e., triangles and squares increase. Further, each question prompt, Q is composed of three types of prompts (in, r, op), where in provided the description of the input image I, r provided details of the task to be performed and op specified the answer format of the model output.

Each domain has its problem instances divided across the sizes varying from 1 to 20, with 10 instances generated for each size in each domain in the dataset. This results in 200 problem instances for each domain and a total of 6000 problem instances divided across 30 domains. Examples of problem instances from our dataset are provided in Figure 1.

There are four types of question prompts in our dataset based on the type of the answer they expect. Details are listed in Table 2. The answer can be a single value (Boolean, numeric), a fixed length answer (a tuple or a list of a fixed length for all problem sizes), an ordered list (a list that has to maintain a specific sequence with its length dependent on the problem size) or an unordered set (a set with only unique values). For each answer type, the question prompt also lists a specific format in which the answer is to be outputted, so that it can be automatically evaluated. All MLLMs mostly honor the format consistently – the errors are from incorrect perception and not format mismatches.

We note all the images in our dataset are generated using automated programs. This has multiple advantages: (1) it helps control the complexity of the problem instances created, and (2) it helps avoid any potential contamination issues, which has been pointed out as a serious concern in some of the existing benchmarks (Shojaee et al., 2025). Further details of the images in each domain along with specific examples, as well as our automated generative process, is described in the Appendix.

4 Experiments

Through our experiments, we try to answer the following questions:

1. How well do the SOTA open as well as proprietary LLMs perform on Percept-V?
2. Do some LLMs perform better than others especially for a subset of skills?
3. How does the LLM performance vary across variation in skills: are some skills harder than others?
4. How does the LLM performance vary with increasing size of the problem instances in different skills?
5. Finally, how does the LLM performance compare with human performance on Percept-V?

In order to answer these questions, we perform our experimental comparison using three different state-of-the-art proprietary LLMs: (1) GPT-4o (2) Gemini 2.5 Flash (3) o4-mini, and two open source MLLMs: (1) Qwen-2.5-7B-Instruct (2) Deepseek-VL2-Tiny. We note that among these, o4-mini and Gemini 2.5 Flash are reasoning/thinking models. For open source models, all our experiments were run on an NVidia A100 GPU. More details about model parameters and GPU runtime are presented in A. The hyperparameters like temperature and top_p are both kept as 0.0 to produce more deterministic reproducible answers from models like Gemini, GPT-4o and Qwen (for the remaining models, the temperature could not be made 0.0 due to model restrictions). We run all experiments in a zero-shot mode to keep the costs under control; further, our preliminary experiments on a subset of models suggest that few shot prompting does not yield any significant gains for our domains.

For each problem size of each domain, all 10 candidate problems are evaluated on all models and the average performance on these 10 problems is considered as the model accuracy for that problem size. For each LLM for a given domain, we use exactly the same prompt. This prompt is constructed by appending the question prompt followed by details of output format. Detailed examples of prompts are given in the Appendix.

4.1 Main Results

This section answers our main experimental question (Q1, Q2 and Q3). Table 3 presents the performance of various LLMs across different skills

Skill / Domain	GPT-4o	o4-mini	Gemini	Qwen	DeepSeek	Avg. Acc.
— Performance by Skill —						
Visual Discrimination	23.70	55.30	51.25	11.50	4.25	29.20
Visual Memory	33.55	47.55	46.55	19.59	1.50	29.75
Visual Sequential Memory	21.50	47.50	41.92	4.67	5.42	24.20
Visual Figure Ground	25.42	63.25	62.83	7.58	0.08	31.83
Visual Form Constancy	27.30	48.90	50.30	25.60	2.50	30.92
Visual Closure	18.40	40.20	34.60	4.60	6.50	20.86
Visual Spatial Relationship	50.36	62.64	69.79	27.79	25.21	47.16
Average of skills	28.60	52.19	51.03	14.48	6.49	37.04
— Performance by Domain —						
change colour	18.50	32.00	23.50	8.00	4.50	17.30
circle boxes	33.00	34.00	28.00	12.50	10.50	23.60
circle location	41.50	46.00	57.50	14.50	6.00	33.10
circle right triangle	82.00	99.50	99.00	25.00	50.00	71.10
colours present	1.50	4.50	4.50	0.00	0.00	2.10
comparing size	36.50	31.00	55.00	5.00	0.00	25.50
count coloured circles	44.00	37.50	44.50	39.00	5.00	34.00
counting circles	57.50	55.00	55.50	42.50	4.00	42.90
counting locations	20.50	36.50	33.50	15.50	0.00	21.20
counting shapes	44.00	58.50	61.00	41.00	0.00	40.90
cross and knots	57.50	100.00	96.00	23.50	47.50	64.90
graph counting	31.00	33.50	36.00	26.00	0.00	25.30
grid path	0.50	53.00	23.50	0.00	0.00	15.40
identifying shapes	69.00	92.50	47.00	41.50	0.00	50.00
inside circles	87.50	95.50	93.00	96.50	73.00	89.10
layered colours	25.00	33.50	34.00	0.50	0.00	18.60
layered shapes	11.50	20.00	18.50	9.50	0.00	11.90
list colours	18.00	35.50	58.00	1.50	0.50	22.70
list shapes	70.50	90.50	92.00	21.00	0.00	54.80
locate circles colour	11.50	100.00	82.00	6.00	0.00	39.90
locate circles shape	16.00	100.00	92.50	7.00	0.00	43.10
match outline	23.00	84.50	72.50	7.50	15.00	40.50
match shadow	32.50	63.00	48.00	5.50	17.50	33.30
maze solving	27.00	30.00	54.50	14.50	0.00	25.20
mirror image	12.50	29.00	24.00	21.50	5.00	18.40
numbered shapes	12.00	75.50	88.50	8.00	0.00	36.80
sort circles	24.00	34.00	30.00	20.00	0.50	21.70
sort lines	22.00	27.00	22.50	5.00	0.00	15.30
vanishing objects	9.00	23.00	25.50	10.00	5.00	14.50
water image	10.50	26.50	22.50	15.00	3.50	15.60
Full Percept-V dataset	31.65	52.70	50.75	18.10	8.25	32.29

Table 3: Comparison of LLM performance across different skills and domains. All values are percentages.

(aggregated over domains that test that skill), as well as over each individual domain. There are several observations to make.

Domain-wise Analysis: For a significant majority of the domains, across LLMs the performance is below par. In general, as expected, closed source models do better than open source. Among closed source models, GPT-4o which is a non-thinking model, does the worst, with its performance on more than 12 (out of a total of 30) domains 20% or less, and on 24 domains 50% or less. Thinking models do somewhat better, with these numbers being 2 (out of 30), and 17, respectively, for o4-mini, which is closely followed by Gemini. Open source models are in general below par with 20 domains performing 20% or less, and only a single domain performing better than 50%, for Qwen. DeepSeek has similar performance. In terms of overall average performance across domains as well, o4-mini is the best performing model with its accuracy only at 52.7%.

These results clearly point to the inability of LLMs, both closed source, and open, as well thinking and non-thinking, to be able to achieve any meaningful performance on most of the domains in our Percept-V dataset. This is contrast with humans, which perform significantly better on these tasks (see Section 4.3). We also see that LLMs are relatively consistent in their performance across domains, for example, 'inside circles' has best overall performance, and this is also the domain, where most LLMs achieve their (near) best. Similarly, 'colors present' has overall the worst performance, and this is also the domain, where individual LLMs

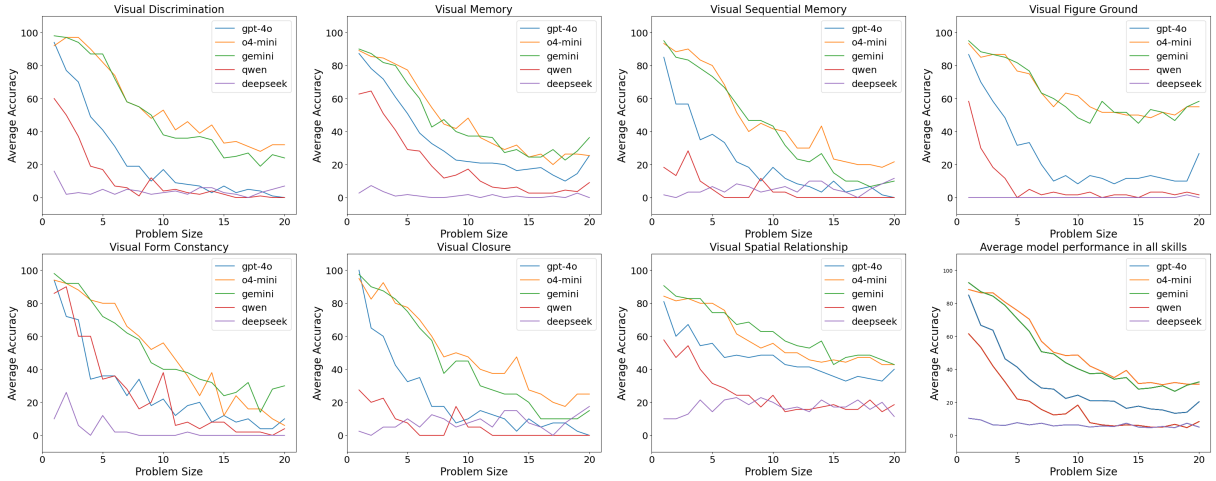

Figure 2: The overall accuracy of all models in different skills.

also perform poorly compared to other domains. This domain requires models to count objects of a given color, and we hypothesize that MLLMs get confused in related colors (such as turqouise and blue). Overall, this suggests that some domains are inherently harder than others for all LLMs.

Skill-wise Analysis: A skill-wise analysis of performance presents no different story. o4-mini is best performing with average accuracy across skills at 52.19% closely followed by Gemini.[2] The best performing o4-mini can achieve more than 50% accuracy only on 3 out of a total of 7 skills, with its higher accuracy on any skill being 63%, demonstrating that it struggles on majority of the skills to get a reasonable performance. GPT4o sees a sharp decline in performance with average accuracy at 28.6%, and among open source models, Qwen is only at 14.48%. We see that o4-mini and Gemini are consistently better than GPT4o across skills, followed by open source models, which perform quite poorly. Similar to the case of domains, there is consistent behavior of LLM performance skill-wise, with some skills being harder (easier) than others for all LLMs.

For instance, most LLMs perform their best on Visual Spatial Relationship, and (near) worst on Visual Closure. Visual spatial relationships generally test relations such as an object being to the left of or above another object. Since image captions generally contain such relationships, it is likely that MLLMs got trained well on this skill and hence perform well. On the other hand, visual closure questions test on inside-out relationships, layered objects, or matching a solid object with just its outline. Such phenomena are likely absent from multimodal training data, making models weaker on this skill.

[2]We note that average accuracy across skills may not match average across across domains, since the same domain may be repeated across multiple skills.

4.2 Performance vs Size

In this section, we set out to answer Q4: what is the variance in LLM performance across varying size? Results comparing the skill-wise model performance across different problem sizes are presented in Figure 2. There are 7 graphs, one for each skill, and the last graph in the bottom row presents performance of each LLM averaged over skills. We observe that performance of models drop significantly with the increase in problem size for all skills. We observe that for most of the skills, thinking models (o4-mini and Gemini) solve smaller problems with lesser number of objects sufficiently well, with near-perfect accuracy, for several skills. However, this performance drops dramatically with increasing size. As seen in Table 3, o4-mini and Gemini are among the best performing models, and are relatively robust to size change compared to other LLMs, despite a performance drop. These are followed by GPT-4o, and open source models perform the worst. Clearly, these results indicate that LLMs are not able to generalize across size across different skills. We hypothesize that the key reason for this performance drop is their inability to 'discern' the underlying pattern of the task to be solved in a given domain, resulting in sharp decline in performance as the size increases. It is also possible that for thinking models, they are able to 'discern' this pattern, but are unable to

operationalize this for larger sizes.

4.3 Human Study

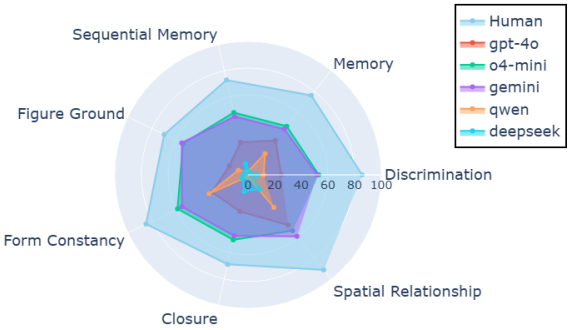

Figure 3: Comparison of MLLM average accuracy versus human accuracy across the seven TVPS-4 skills.

This section answers Q5, i.e., how does LLM performance vary in contrast with human subjects. A study on 16 college students was done with the Percept-V. It had 6 female and 10 male subjects. The subjects were presented randomly selected questions of any problem size from the 30 domains. Results of Human Study are presented in Figure 3. Here, for fair comparison, the models are also evaluated on the exactly same samples as were given to the humans. It can be seen that models lag severely as compared to humans. Even the best performance of the models exhibited by Gemini in Visual Spatial Relationship is around 60% as compared to about 90% accuracy shown by humans in the same skill. Overall see Table 4 – human performance is about 30 percent points higher than than closest MLLM. When testing on 'colours present', the domain on which all MLLMs have abysmal performance, we find the gap to be humongous – humans achieve 81.25% performance whereas best MLLMs on the same questions achieve only 6.25%.

This experiment underscores a large gap in the overall perceptual ability of MLLMs compared to humans – an important direction for future work.

Model	Overall performance
GPT-4o	31.17%
o4-mini	51.46%
Gemini	49.27%
Qwen	20.40%
Deepseek	8.45%
Human	**81.67%**

Table 4: Overall performance in human study

4.4 Compliance with Output Format

We find that most proprietary models hardly falter in following the answer formats specified in the question, and most of their accuracy issues stem from lack of perceptual skills alone. However, this is not true for DeepSeek, which is the worst at following output formats in all categories. The best performing is GPT-4o (with o4-mini close second) that sticks to the output formats with 0% format errors. Single answer format is the easiest to follow for models, whereas fixed answer type is the hardest as can be seen from Table 5.

Answer Type	GPT-4o	o4-mini	Gemini	Qwen	DeepSeek
single	0.00	0.00	0.00	0.00	3.68
fixed	0.00	0.17	0.83	5.18	61.58
list	0.00	0.00	0.00	0.00	22.44
set	0.00	0.10	0.20	19.10	60.00

Table 5: Format Error (%) Analysis of models

5 Conclusion

Our paper takes an important step forward in assessing the perceptual abilities of multimodal LLMs. To isolate perceptual ability, our questions use very few and basic reasoning skills (such as counting), and similarly very basic knowledge (of colors and basic shapes). We present Percept-V, an extensive dataset of 6000 automatically generated images from 30 domains, which are uncontaminated, i.e., not seen by any LLM before. We also annotate each domain with their perceptual skills from the TVPS-4 Cognitive Science framework.

When tested on our dataset, all MLLMs perform rather weakly on almost all domains. The best average performance on our dataset is only 52.7% (o4-mini), and open models perform rather poorly (8-18%). As complexity of questions increases, all model performances degrade quite fast on almost all skills, with thinking models continuing to perform decently on a couple of skills.

We also perform human study to compare human performance with MLLMs. We find that the gap between the best MLLM and humans is quite high (over 30 percent points), suggesting that contrary to LLMs' high performance against humans in many tasks (e.g., LLMs winning Math olympiad), MLLMs are rather behind, especially on basic perception skills. We will release Percept-V upon publication. We hope that it will become a standard dataset for testing any new MLLM, and will help in improving the quality of future MLLMs.

Limitations

This paper evaluates MLLMs only against a limited pool of adult college students. Future work may be expanded into including annotators from different age groups and diverse educational background to improve the robustness of the results. Additionally, this paper only identifies the weakness of MLLMs in visual perceptual skills and does not investigate in detail, the rationale behind such weakness.

Ethical Considerations

All human participants were shown the instructions required to be followed during the experiment, and they volunteered to participate in the study. Additionally, all humans were asked for consent and financially compensated for their contribution – details can be found in Appendix A. Our work only tries to expose the perceptual limitations of SoTA MLLMs, so human study did not elicit any personally identifying information – the participants were simply answering simple perception questions.

References

Jiwan Chung, Neel Joshi, Pratyusha Sharma, Youngjae Yu, and Vibhav Vineet. 2025. What mllms learn about when they learn about multimodal reasoning: Perception, reasoning, or their integration? *Preprint*, arXiv:2510.01719.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *Preprint*, arXiv:2501.05444.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Aditya Kanade and Tanuja Ganu. 2025. Do you see me : A multidimensional benchmark for evaluating visual perception in multimodal llms. *Preprint*, arXiv:2506.02022.

Michael King. 2023. Administration of the text-based portions of a general iq test to five different large language models. *Authorea Preprints*.

Jonghyun Lee, Dojun Park, Jiwoo Lee, Hoekeon Choi, and Sung-Eun Lee. 2025. Exploring multimodal perception in large language models through perceptual strength ratings. *Preprint*, arXiv:2503.06980.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Nancy A Martin and Morrison F Gardner. 2006. *Test of visual perceptual skills*, volume 420. Academic Therapy Publications Novato, CA.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *arXiv preprint arXiv:2302.05729*.

Pranshu Pandya, Vatsal Gupta, Agney S Talwarr, Tushar Kataria, Dan Roth, and Vivek Gupta. 2024. Ntsebench: Cognitive reasoning benchmark for vision language models. *arXiv preprint arXiv:2407.10380*.

R. Rabindran and D. Madanagopal. 2020. Piaget's theory and stages of cognitive development—an overview. *Scholars Journal of Applied Medical Sciences*, 8(6):2152–2157.

Susanne Schregel. 2020. 'the intelligent and the rest': British mensa and the contested status of high intelligence. *History of the Human Sciences*, 33(5):12–36.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, and 1 others. 2024a. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*.

Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. 2024b. Exploring perceptual limitation of multimodal large language models. *Preprint*, arXiv:2402.07384.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and 1 others. 2023. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*.

A Appendix

A.1 Domain examples of MLLMs

A.1.1 change color

Prompt:
-These are two identical images with circles of different colours. - The colours of the circles at the same position in the two images may be different. - Count the number of differently coloured circles between the two images. - The output must be given in a single line in the form of COUNT:x

A.1.2 circle boxes

Prompt:
- This is an image with some red circles divided and a blue partition that divides the image into two sides. - Some circles are on the left side of the partition whereas some are on the right. Note that, a side of the image may be completely empty as well. - Determine the number of circles to be moved from the side with more number of circles to the side with less number of circles so that each side has equal number of circles. - In case of both the sides having equal number of circles, return zero. - The output must be given in a single line in the form of COUNT:x where x is the number of circles to be transferred

A.1.3 circle_location

Prompt:
- This is an image with some black circles drawn in the four quadrants - The quadrants are numbered in the default way i.e. quadrant 1 is the rightmost and topmost quadrant, quadrant 2 is the leftmost and topmost, quadrant 3 is the leftmost and lowest and quadrant 4 is the rightmost and lowest. - Determine the quadrant with the most number of black circles and the count of black circles in that quadrant. - In case of a tie, return the lesser quadrant number. - The output must be given in a single line in the form of QUADRANT:x COUNT:y where x is the quadrant number and y is the count of circles.

A.1.4 circle_right_triangle

Prompt:
- This is a grid which contains some circles and triangles - Check if any triangle cell has a circle cell to its immediate right. - The last line of output should be "NO" if no such cell exists and "YES" otherwise.

A.1.5 colours_present

Prompt:
- This image contains some shapes in different colours. - The colours belong to the following list - [black, gray, brown, maroon, red, coral, tan, orange, ivory, goldenrod, yellow, green, olive, turquoise, skyblue, blue, lavender, purple, pink, fuchsia] - No colour is repeated - Produce a list of yes or no; write yes in the list if the colour at the corresponding index of the above list is present and no if it is absent.

- The output should be written in a single line as a comma-separated list of yes or no, written in order of the given list of colours. - For example, the output should be of the form ANSWER: yes , no, no, yes and so on.

A.1.6 Comparing Size

Input Prompt:
- This is an image with several rows of circles where each row is above a horizontal line.
- Each row contains two colored circles (one blue and one green) of different sizes.
- Analyse which circle is bigger between the two in each row.
- The last line of the output should be of form "ANSWER:" followed by the space separated color answers for all the rows in the image.
- For example ANSWER: Blue Green Blue.

A.1.7 Count Colored Circles

Input Prompt:
- This is an image containing some colored circles.
- This is an image containing some colored circles.
- Last line of the output must be of the form COUNT:x

A.1.8 Counting Circles

Input Prompt:
- This is an image containing some black circles.
- Count the total number of circles in the image.
- The last line of the output should be of form COUNT:x.

A.1.9 Counting Locations

Input Prompt:
- There will be an image containing a rectangle and some circles above and below it.
- Count the number of circles above the rectangle and below the rectangle.
- The output must be of form in a single line ABOVE:x BELOW:y

A.1.10 Counting Shapes

Input Prompt:
- This is an image containing shapes such as circles, triangles and squares.
- Count the number of circles, triangles and sqaures in the image.
- The last line of the output must be of the form CIRCLES:x TRIANGLES:y SQUARES:z

A.1.11 Cross and Knots

Input Prompt:
- This is a grid containing some black X.
- Each cell is given a coordinate (x,y) which is (row,column) with 0 based indexing.
- An adjacent cell to a cell A is any cell sharing an edge with A
- A safe cell is a cell that is not adjacent to any black X.
- Output the coordinates of any one safe cell as labelled in the grid.
- If no such cell exists output None.
- The last line of the output should just contain the cell coordinates (x,y) or None and nothing else.

A.1.12 Graph Counting

Input Prompt:
- This is a graph where the blue colored circles are nodes and the lines joining them are edges.
- Count the number of nodes and edges in the graph.
- Last line of the output must be of the form NODES:x EDGES:y.

A.1.13 Grid Path

Input Prompt:
- This is a grid which contains some circles, triangles and squares.
- Travel from cell labelled S to cell labelled E.
- Follow the path marked by the black line and arrows.
- Write down the sequence of shapes seen while moving along this path.
- The last line of the output should be a comma-separated list of shapes without colors, written in order of visiting of form SHAPES : Shape 1 , Shape 2 and so on.

A.1.14 Identifying Shapes

Input Prompt:
- This is an image with several rows of different shapes where each row is above a horizontal line. - Each row contains a coloured shape.
- Identify the shape present in each row.
- The ouput should have only a single line containing the answer.
- The output format should be "ANSWER:" followed by the space seperated shape names.
- For example, ANSWER: Circle Triangle Circle

A.1.15 Inside Circles

Input Prompt:
- This is an image containing some black circles and a single red dot.
- Count the total number of circles in the image which contain the red dot
- Note that a circle containing the red dot means the centre of the dot is present inside the circumference of the circle
- The last line of the output should be of form COUNT:x.

A.1.16 Layered Colors

Input Prompt:
- This image contains some concentric circles in different colours.
- Write down the sequence of colours seen from inside to outside.
- The colours belong to the following list - [black, gray, brown, maroon, red, coral, tan, orange, ivory, goldenrod, yellow, green, olive, turquoise, skyblue, blue, lavender, purple, pink, fuchsia]
- No colour is repeated
- The output should be written in a single line as a comma-separated list of colours, written in order from the innermost colour to the outermost colour
- For example, the output should be of the form COLOURS: Colour 1 , Colour 2 and so on.

A.1.17 Layered Shapes

Input Prompt:
- This image contains some shapes that are layered on top of each other in a white background.
- The image could also contain a single shape with no other shapes layered on top of it.
- Write down the sequence of shapes from the inside layer to the outside layer.
- If there is a single shape then identify it.
- The last line of the output should be a comma-separated list of shapes without colors, written in order from top to bottom of form SHAPES : Shape 1 , Shape 2 and so on.
- If there is a single shape then write SHAPES : shape-name where shape-name is the name of the shape seen

A.1.18 List Colors

Input Prompt:
- This is an image with some multi-coloured shapes drawn on top of it
- The colours belong to the following list - [black, gray, brown, maroon, red, coral, tan, orange, ivory, goldenrod, yellow, green, olive, turquoise, skyblue, blue, lavender, purple, pink, fuchsia]
- No colour is repeated
- Identify the colour of the shapes drawn over the background image
- Form a list of all the colours of the shapes
- The output must be given in a single line in the form of a list of all the colours found

A.1.19 List Shapes

Input Prompt:
- This is an image with some multi-coloured shapes drawn on top of it.
- Identify the shapes drawn over the background image
- Form a set of all the unique shapes seen in the image
- The output must be given in a single line in the form of a set of unique shapes found.

A.1.20 Locate Circles Color

Input Prompt:
- This is a grid which contains some coloured circles.
- The rows and columns are numbered in the grid.
- The coordinate of a cell is given by the row number followed by the column number.
- Write down all the coordinates of the cells which contain a green circle.
- The last line of the output should be a space-separated list of the coordinates of the cells which contain a green circle.
- The coordinates should be in the format (row,column) withh the comma and bracket.

A.1.21 Locate Circles Shape

Input Prompt:
- This is a grid which contains some coloured circles , rectangles and triangles.
- The rows and columns are numbered in the grid.
- The coordinate of a cell is given by the row number followed by the column number.
- Write down all the coordinates of the cells which contain a green circle.
- The last line of the output should be a space-separated list of the coordinates of the cells which contain a green circle.
- The coordinates should be in the format (row,column) with the comma and bracket.

A.1.22 Match Outline

Input Prompt:
- These are two images of some sequence of shapes separated by black horizontal lines
- The first image consists of a sequence of shapes and the second image consists of a sequence of outlines of shapes
- Determine the number of shapes in the first image that match the outlines in the corresponding place in the second image.
- The output must be given in a single line in the form of ANSWER:x where x is the number of matching shapes

A.1.23 Match Shadow

Input Prompt:
- These are two images of some sequence of shapes separated by black horizontal lines.
- The first image consists of a sequence of shapes and the second image consists of a sequence of shadows of shapes
- Determine the number of shapes in the first image that match the shadows in the corresponding place in the second image.
- The output must be given in a single line in the form of ANSWER:x where x is the number of matching shadows.

A.1.24 Maze Solving

Input Prompt:
- This is a maze. - You start from the cell labelled S.
- Your goal is to reach cell labelled E through the shortest path possible.
- You can move up, down, left or right.
- You cannot move through the black walls.
- You can only move through white cells.
- Give the sequence of cells taken to go from S to E.
- Last line of output should be a comma separated line with cell numbers starting with S and ending at E and nothing else.

subsubsectionMirror Image

Input Prompt:
- These are two images containing different shapes in different colours
- The image consists of a vertical blue line at the centre.
- Some objects of the first image are mirrored along the given vertical line in the second image.
- Other objects in the second image maintain the same postions as in the first image.
- Determine the number of objects that are mirrored in the second image.
- The output must be given in a single line in the form of COUNT: x, where x is the number of being mirrored

A.1.25 Numbered Images

Input Prompt:
- This is an image with some shapes drawn and numbers written on top of the shapes.
- Determine the list of numbers written on top of the circles.
- The output must be given in a single line in the form of CIRCLES:x where x is the list of numbers seen on top of the circles

A.1.26 Sort Circles

Input Prompt:
- There are several black circles with different sizes.
- Sort these circles by size, from smallest to largest.
- Output must only contain the space separated labels of circles in their sorted order by size.

A.1.27 Sort Lines

Input Prompt:
- The image has sseveral parallel lines of varying lengths.
- Sort these lines by length, from shortest to longest.
- Last line of the output must only contain the space separated labels of thee lines in their sorted order by length and nothing else.

A.1.28 Vanishing Objects

Input Prompt:

- These are two images containing shapes such as circles, triangles and squares
- The images are identical except for the fact that the second image has a few objects missing
- Count the number of missing objects in the secnd image
- The output must be given in a single line in the form of COUNT:x

A.1.29 Water Image

Input Prompt:

- These are two images containing different shapes in different colours
- The image consists of a horizontal blue line at the centre.
- Some objects of the first image are mirrored along the given horizontal line in the second image.
- Other objects in the second image maintain the same postions as in the first image.
- Determine the number of objects that are mirrored along the horizontal line in the second image.
- The output must be given in a single line in the form of COUNT: x, where x is the number of being mirrored

A.2 Image Generation

Image is generated using hyperparameters like –num_image that give the number of images to be generated per problem size and –num_size that give the list of problem sizes

A.3 Financial Compensation for Participants

Each participant was paid financial remuneration based on the number of hours taken to finish the questionnaire which was typically around one hour for 30 questions.

A.4 Instructions for Participants

Thank you for participating in this experiment! This questionnaire will test your visual perception skills through some simple questions. Please follow the below instructions carefully - 1. Focus on accuracy and not on speed. There is no time bound , so you can take breaks while solving. However, you have to keep the website open on some tab/window otherwise, your progress will be lost. Also, take sufficient time to solve each question. 2. Please open this questionnaire on laptop/tablet . Some fonts are small, so, you may miss out if you open on smaller devices. Zoom in on images that appear tiny otherwise, this may affect your accuracy! 3. Each question follows a particular answer format. Follow the answer format strictly! Check for spelling mistakes while writing. 4. Try to attempt all the questions. If you do not understand any question, reach out to us! All the best!

A.5 GPU Hours

GPU Hours required were around 5 for Deepseek (1B parameters) and Qwen(7B parameters). The number of parameters for o4-mini is 8B, GPT-4o is 200B.