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 many benchmarks 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' visual perception capabilities from their reasoning abilities and domain-specific knowledge.

Visual perception represents a foundational cognitive ability that precedes the acquisition of complex reasoning and knowledge-based skills in humans (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 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. See Figure 1 for sample questions.

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 such as visual discrimination, visual-spatial relationships, and form constancy (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. Each instance asks a perception question over an image that generally has simple objects such as circles and triangles, in different colors and sizes. The use of basic objects is similar to benchmarks in robotics (Li et al., 2024a), and mathematics (Wang et al., 2024).

To systematically assess performance across varying levels of difficulty, each domain contains 200 images with varying problem size, which is operationalised through the number of objects present,

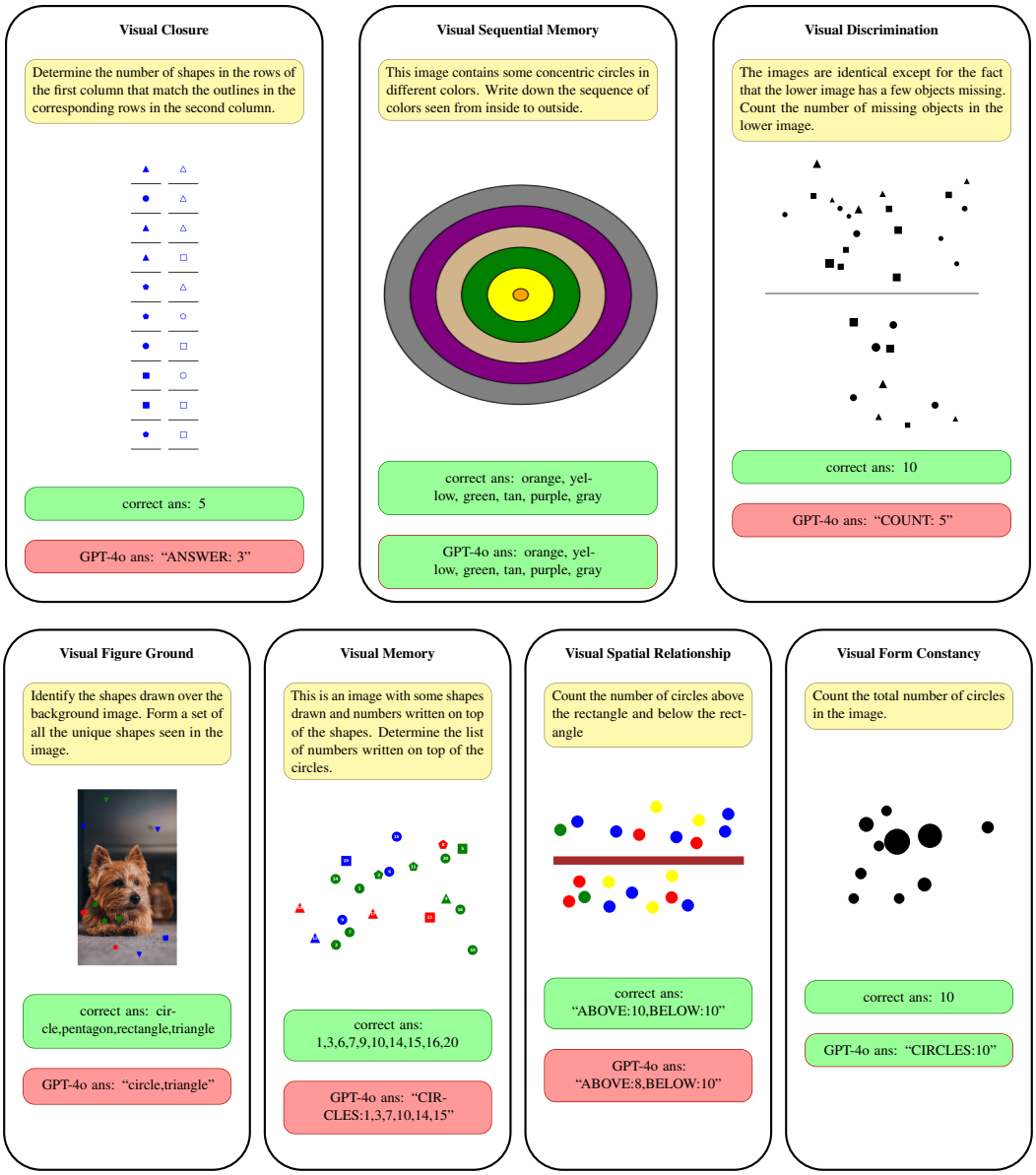

Figure 1: Sample questions from the Percept-V dataset illustrating tasks related to different visual perception skills as defined by the TVPS-4 framework. Each example shows the summarised image prompt, the correct answer, and a typical incorrect response from an MLLM.

grid dimensions, or the number of processing steps required to complete the task. This 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 four state-of-the-art proprietary MLLMs – GPT-5-mini, 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 existing benchmarks for evaluating the performance of MLLMs necessitate a combination of

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.

knowledge, reasoning and perception for good performance. Alongside perception, the NTSEBench dataset (Pandya et al., 2024) tests commonsense reasoning, the MathVista dataset (Lu et al., 2023) evaluates compositional reasoning, and the MathVision benchmark (Wang et al., 2024) includes questions from 16 mathematical disciplines.

In a similar spirit, OlympiadBench (He et al., 2024) collects advanced reasoning questions from mathematics, physics and chemistry. MMMU, CMMMU and EMMA (Yue et al., 2024; Zhang et al., 2024a; Hao et al., 2025) evaluate MLLMs on many disciplines such as engineering, medicine, humanities, science, business and art. 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., 2024b). All these datasets, along with perception, require complex reasoning skills, and several datasets additionally require specialized knowledge of a subject.

Some recent benchmarks specifically try to disentangle reasoning and perception. Chung et al. (2025) propose MatHLENS which separately tests the perceptual, reasoning and integrating capabilites of MLLMs. In recent times, some studies specifically highlight the lack of MLLMs' perceptual abilities, e.g., 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.

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

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. 43% of our domains require multiple skills, which allows us to assess MLLM's ability when applying skills in concert. Their work does not explicitly address this.

3 Dataset

We now describe the details of our dataset, which we name Percept-V. It has 30 different domains where each domain tests one or more *basic percep-*

tion skills from the TVPS-4 cognitive framework (Martin and Gardner, 2006). 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 between objects.
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 rotated.
7. *Visual Closure*: The ability to recognize a form, when a part of the picture is missing.

Each domain has multiple *problem instances* with varying perceptual complexity. Table 1 lists, for each skill, the domains that require that skill to solve the problem instances correctly. In total, 17 domains test one skill, 8 domains have two skills, and 5 domains test 3 skills in concert.

Skills associated with visual memory and visual sequential memory typically evaluate the ability to memorize and recall a detail once it has been presented and then removed from the field of view. We note that this kind of set-up is not possible for modern day MLLMs since anything once shown is always present in model's memory. Therefore, specifically, for these skills, our domains can be seen as testing the skill of scanning and transcribing (which is a subset of the original skill), and our mapping to the skills of visual (sequential) memory, is approximate. Nevertheless, even in this set-up, which is presumably easier than the original skills that additionally require memory, the performance of MLLMs is quite low (Table 3), pointing a fundamental limitation in their basic perception skills.

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 programs.[1] In other words, we apply the same generic question prompt to different images resulting in different problem instances. The instances vary in 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 instance. Each domain has its 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 instances divided across 30 domains. Examples of instances from our dataset are provided in Figure 1. For example, in the $counting\_circles$ domain, the problem size is controlled by controlling the number of circles to be counted. Similarly, in the $numbered\_shapes$ domain, the number of circles, and the number of distractor objects, i.e., triangles and squares, increase with increasing problem size.

Each question prompt, Q is composed of three types of prompts (in, r, op), where in provides the description of the input image I, r provides details of the task to be performed and op specifies the answer format of the model output. 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 (see Section 4.5).

All the images in our dataset are generated using automated scripts. Every domain has a separate script to generate images specifically for that domain. The script takes a list of problem sizes and number of instances of each problem size to be generated as its arguments. The script also randomizes the position of objects in the image to ensure the generation of a new instance. Each domain uses only simple, basic shapes like circles, triangles, squares, and pentagons, and structures like grids, rows, and rectangular mazes, which are produced using standard Python libraries like Pillow, Matplotlib, and OpenCV. The answers for each problem instance are also computed during the data

[1] image generation for each domain is controlled by a set of hyper-parameters described in Appendix A

generation process, and hence, there is no scope for error or ambiguity in the gold answer. We note that our methodology of generating the dataset using automated scripts as detailed above has multiple advantages: (1) it helps control the complexity of the created problem instances and (2) it helps avoid any potential contamination issues, pointed out as a serious concern in some of the existing benchmarks (Shojaee et al., 2025).

4 Experiments

Through our experiments, we try to answer the following questions: (1) How well do the SoTA open as well as proprietary MLLMs perform on Percept-V? (2) Do some MLLMs perform better than others especially for a subset of skills? (3) How does an MLLM's performance vary across variation in skills: are some skills harder than others? (4) How does the MLLM performance vary with increasing size of the problem instances in different skills? (5) Finally, how does the MLLM performance compare with human performance on Percept-V?

In order to answer these questions, we perform our experimental comparison using four different state-of-the-art proprietary LLMs: (1) GPT-4o (2) Gemini 2.5 Flash (3) o4-mini, (4) GPT-5-mini and two open source MLLMs: (1) Qwen-2.5-7B-Instruct (2) Deepseek-VL2-Tiny. We note that among these, GPT-5-mini, o4-mini and Gemini 2.5 Flash are reasoning/thinking models, and all, except Qwen 2.5 VL Instruct, are equipped with 'think with image' capabilities – i.e., these models leverage visual information as intermediate steps in their thought process (Su et al., 2025).

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 Appendix A. The hyperparameters like temperature and top_p are both kept as 0.0 to produce more deterministic reproducible answers from models like GPT-5-mini, 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 primarily in a zero-shot mode to keep the costs under control. We further experiment with one-shot prompting, but that does not yield any significant gains for our domains. 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 research questions (Q1, Q2 and Q3). Table 3 presents the performance of various LLMs across different skills (aggregated over domains that test that skill), as well as over each individual domain.

Domain-wise Analysis: For a significant majority of the domains, across MLLMs 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 30 domains is 20% or less, and on 24 domains 50% or less. Thinking models do somewhat better, with these numbers being 2 and 17, respectively, for o4-mini, which is closely followed by Gemini and GPT-5-mini. Open source models are in general below par with 20 domains obtaining 20% or less accuracy, 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, GPT-5-mini is the best performing model, but with its accuracy only at 55.22%.

These results clearly point to the inability of MLLMs, both closed source, and open, as well thinking and non-thinking, in achieving any meaningful performance on most of the domains in our Percept-V dataset. This is in contrast with humans, who perform significantly better on these tasks (see Section 4.4). We also see that MLLMs are relatively consistent in their performance across domains, for example, 'inside circles' has best overall performance, and this is also the domain, where most MLLMs achieve their (near) best. Similarly, 'colors present' has overall the worst performance, and this is also the domain, where individual MLLMs 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 (even though a legend clarifies all colors at the top of each image). Overall, this suggests that some domains are inherently harder than others for all current-day MLLMs.

Skill-wise Analysis: A skill-wise analysis of performance presents no different story. GPT-5-mini performs the best, with average accuracy across

Skill / Domain	GPT-5-mini	GPT-4o	o4-mini	Gemini	Qwen	DeepSeek	Avg. Acc.
— Performance by Skill —							
Visual Discrimination	55.0	23.7	55.3	51.25	11.5	4.1	33.47
Visual Memory	49.95	33.55	47.55	46.55	19.59	1.32	33.08
Visual Sequential Memory	43.5	21.5	47.5	41.92	4.67	5.25	27.39
Visual Figure Ground	64.75	25.42	63.25	62.83	7.58	0.08	37.32
Visual Form Constancy	58.7	27.3	48.9	50.3	25.6	3.0	35.63
Visual Closure	44.88	23.0	50.25	43.25	5.75	7.88	29.16
Visual Spatial Relationship	66.64	50.36	62.64	69.79	27.79	17.64	49.14
Average of skills	54.77	29.26	53.63	52.27	14.64	5.61	35.03
— Performance by Domain —							
change_colour	32.0	18.5	32.0	23.5	8.0	4.5	19.75
circle_boxes	30.0	33.0	34.0	28.0	12.5	10.5	24.67
circle_location	56.5	41.5	46.0	57.5	14.5	5.5	36.92
circle_right_triangle	100.0	82.0	99.5	99.0	25.0	50.0	75.92
colours_present	3.5	1.5	4.5	4.5	0.0	0.0	2.33
comparing_size	27.0	36.5	31.0	55.0	5.0	0.0	25.75
count_coloured_circles	51.0	44.0	37.5	44.5	39.0	2.5	36.42
counting_circles	68.0	57.5	55.0	55.5	42.5	5.5	47.33
counting_locations	44.5	20.5	36.5	33.5	15.5	0.0	25.08
counting_shapes	72.5	44.0	58.5	61.0	41.0	0.0	46.17
cross_and_knots	99.5	57.5	100.0	96.0	23.5	0.0	62.75
graph_counting	39.0	31.0	33.5	36.0	26.0	0.0	27.58
grid_path	54.5	0.5	53.0	23.5	0.0	0.0	21.92
identifying_shapes	79.5	69.0	92.5	47.0	41.5	0.0	54.92
inside_circles	96.0	87.5	95.5	93.0	96.5	68.0	89.42
layered_colours	35.0	25.0	33.5	34.0	0.5	0.0	21.33
layered_shapes	24.0	11.5	20.0	18.5	9.5	0.0	13.92
list_colours	41.0	18.0	35.5	58.0	1.5	0.5	25.75
list_shapes	88.5	70.5	90.5	92.0	21.0	0.0	60.42
locate_circles_colour	100.0	11.5	100.0	82.0	6.0	0.0	49.92
locate_circles_shape	100.0	16.0	100.0	92.5	7.0	0.0	52.58
match_outline	66.0	23.0	84.5	72.5	7.5	17.0	45.08
match_shadow	54.5	32.5	63.0	48.0	5.5	14.5	36.33
maze_solving	43.0	27.0	30.0	54.5	14.5	0.0	28.17
mirror_image	33.5	12.5	29.0	24.0	21.5	5.0	20.92
numbered_shapes	81.5	12.0	75.5	88.5	8.0	1.0	44.42
sort_circles	36.5	24.0	34.0	30.0	20.0	0.0	24.08
sort_lines	26.5	22.0	27.0	22.5	5.0	0.0	17.17
vanishing_objects	35.0	9.0	23.0	25.5	10.0	5.0	17.92
water_image	38.0	10.5	26.5	22.5	15.0	3.5	19.33
Full Percept-V Dataset	55.22	31.65	52.7	50.75	18.1	6.43	35.81

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

skills at 54.77%, closely followed by o4-mini and Gemini.[2] GPT-5-mini achieves more than 50% accuracy only on 4 out of 7 skills, with its higher accuracy on any skill being 66.64%, demonstrating that it struggles to get a reasonable performance on all skills. GPT-4o 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 GPT-4o across skills, followed by open source models, which perform quite poorly. Similar to domain-wise analysis, there is consistent behavior of MLLM performance skill-wise, with some skills being harder (easier) than others for all MLLMs.

For instance, most LLMs perform their best on Visual Spatial Relationship, and (near) worst on Visual Closure. Visual spatial relationships 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. 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.

4.2 Performance vs Size

We now 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

[2] A domain is included in all skills it exhibits

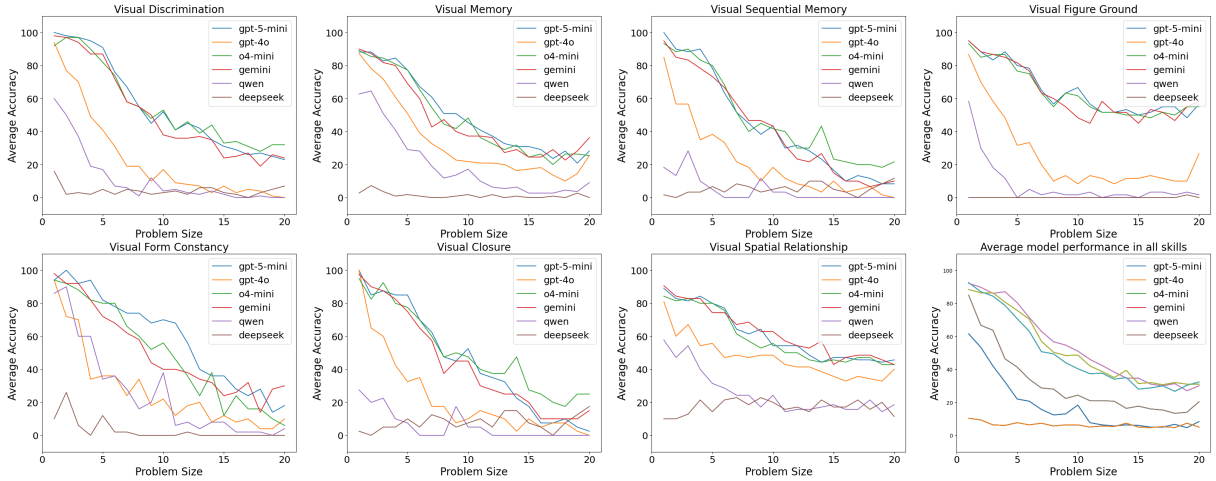

Figure 2: The overall accuracy of all models in different skills.

graph in the bottom row presents performance of each MLLM 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 less 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, GPT-5-mini, 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 sizes very well.

4.3 Alternate Skill Analysis

Separate from TVPS-4 classification, we observe that our dataset uses three other frequent skills – counting, optical character recognition (OCR) and handling grid-like images. They appear in 13, 9, and 7 domains, respectively. We note that counting is a basic reasoning skill, employed primarily so as to ease automatic evaluation of answers, nevertheless, it may distract from the main results. An analysis of counting (C) vs non-counting (NC) domains shows the following results: GPT-4o (C:29.03, NC:32.13), GPT5-mini (C: 47.73, NC: 58.54), o4-mini (C:42.99, NC:58.44), Gemini (C:40.92, NC:55.22), Qwen (C:19.88, NC:15.93), and DeepSeek (C:5.65, NC:6.88). This suggests that while performance is not very dissimilar, for some thinking MLLMs, inclusion of counting skill may make the problem somewhat harder. Still, even for non-counting domains, the performance is not particularly stronger, so it is safe to conclude counting ability may be one of the factors, but does not explain all of the weak results on our dataset.

Figure 3 shows performance for the three alternate skills vs. problem size. We observe that models have a similar sharp drop in case of counting and OCR. However, the case for grid understanding is different as models, particularly, reasoning models like GPT-5-mini, o4-mini and Gemini have a more-or-less plateau-like curve with no significant accuracy drops. In fact, GPT-5-mini shows 100% accuracy in 3 out of 7 grid-based domains! We hypothesize that the training data distribution (especially in RLVR tasks) may frequently include grid-like structures, and that might have resulted in their better understanding of grids.

4.4 Human Study

We now answer Q5 – how do MLLMs compare with human performance. We recruit 16 college students, 6 female and 10 male. They were presented randomly selected questions of any problem size from the 30 domains. Results of this study are presented in Figure 4. For fair comparison, the models are also evaluated on the exactly same samples, and exactly the same protocol, as 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 73.08% as compared to about 96.42% accuracy shown by humans in the same skill. Overall (see Table 5 in appendix) – human performance is more than 40% points higher than the closest MLLM. Interestingly, when test-

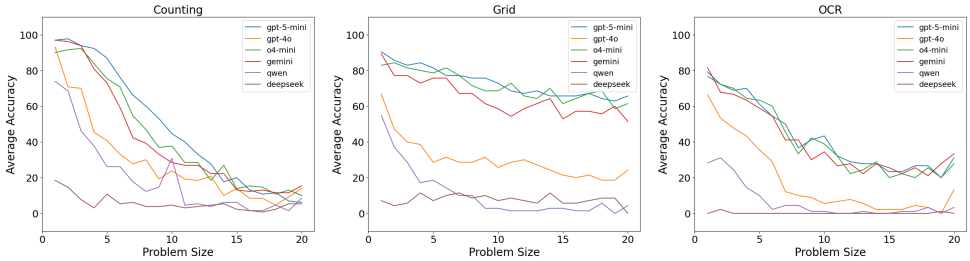

Figure 3: The overall accuracy of all models in counting, grid understanding and ocr skills.

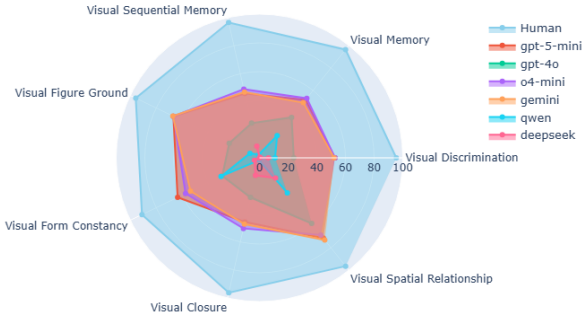

Figure 4: Comparison of MLLM average accuracy versus human accuracy across the seven TVPS-4 skills.

ing on 'colours_present', the domain on which all MLLMs have abysmal performance, we find the gap to be humongous – humans achieve 90% performance whereas best MLLMs on the same questions achieve only 6.67%.

4.5 Other Results

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 4.

Answer Type	GPT-5-mini	GPT-4o	o4-mini	Gemini	Qwen	DeepSeek
single	0.14	0.00	0.00	0.00	0.00	3.68
fixed	0.00	0.00	0.17	0.83	5.18	61.58
list	2.19	0.00	0.00	0.00	0.00	22.44
set	0.00	0.00	0.10	0.20	19.10	60.00

Table 4: Format Error (%) Analysis of models

One-shot Results: We additionally perform one-shot experiments on proprietary models to assess how much exemplars can help MLLMs in understanding the task better. Contrary to conventional wisdom, we find that one-hot prompting *degrades* the overall performance, with average of skills accuracy reducing to 51.25% for GPT-5-mini, 17.62% for GPT-4o, 52.26% for o4-mini and 44.42% for Gemini (refer to Table 6 in Appendix). Moreover, the skill-wise graphs also show degrading trends across sizes, similar to zero-shot (see Fig. 35 in Appendix). This hints that MLLMs' weak performance is not likely because of task understanding, but due to inherent ability limitations.

5 Conclusion

Our paper takes an important step forward in assessing the perceptual abilities of multimodal LLMs. We present Percept-V, an extensive dataset of 6000 automatically generated images from 30 domains where each domain is mapped with a subset of perceptual skills from the TVPS-4 Cognitive Science framework. To isolate perceptual ability, questions in our dataset makes use very few basic reasoning skills (such as counting), and similarly very basic knowledge (of colors and basic shapes). A variety of open and closed source MLLMs perform rather weakly on almost all domains, with best average performance of any model being 55.2%, demonstrating significant gap in their ability to solve basic perceptual tasks. Our human study experiments show that the gap between the best MLLM and humans is quite high suggesting that contrary to LLMs' high performance against humans in existing tasks (e.g., LLMs winning Math olympiad), MLLMs are rather behind, especially on basic perception skills. We will release Percept-V upon publication, and hope that it will become a de-facto standard for benchmarking perceptual abilities of any new and existing 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.

Jinming Li, Yichen Zhu, Zhiyuan Xu, Jindong Gu, Minjie Zhu, Xin Liu, Ning Liu, Yaxin Peng, Feifei Feng, and Jian Tang. 2024a. Mmro: Are multimodal llms eligible as the brain for in-home robotics? *Preprint*, arXiv:2406.19693.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. 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*.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *Preprint*, arXiv:2506.23918.

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

Input 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

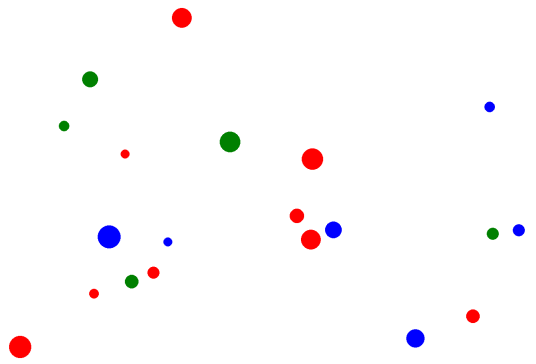

Figure 5: Change Color

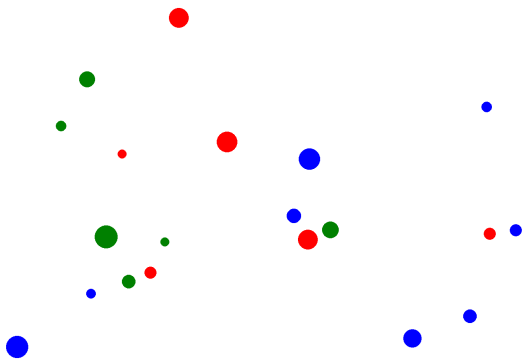

Figure 6: Change Color

A.1.2 Circle Boxes

Input 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

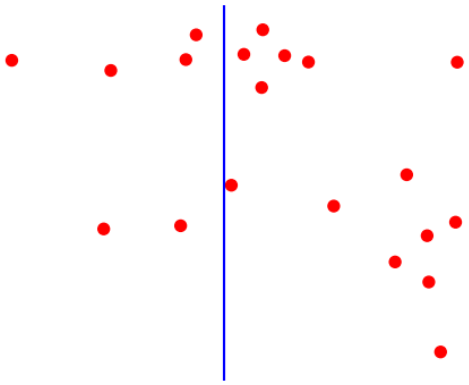

Figure 7: Circle Boxes

A.1.3 Circle Location

Input 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.

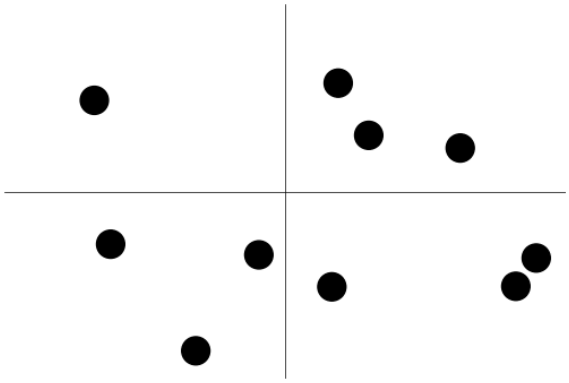

Figure 8: Circle Location

A.1.4 Circle Right Triangle

Input 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.

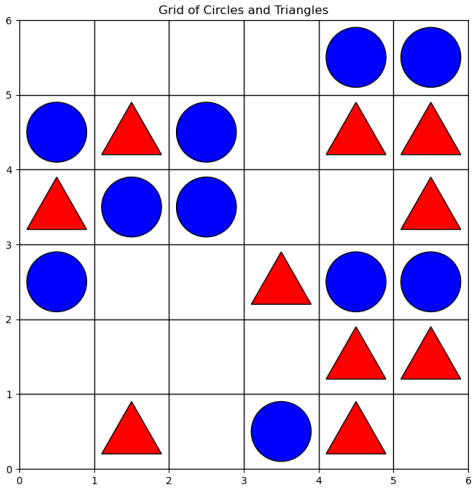

Figure 9: Circle Right Triangle

A.1.5 Colours Present

Input 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.

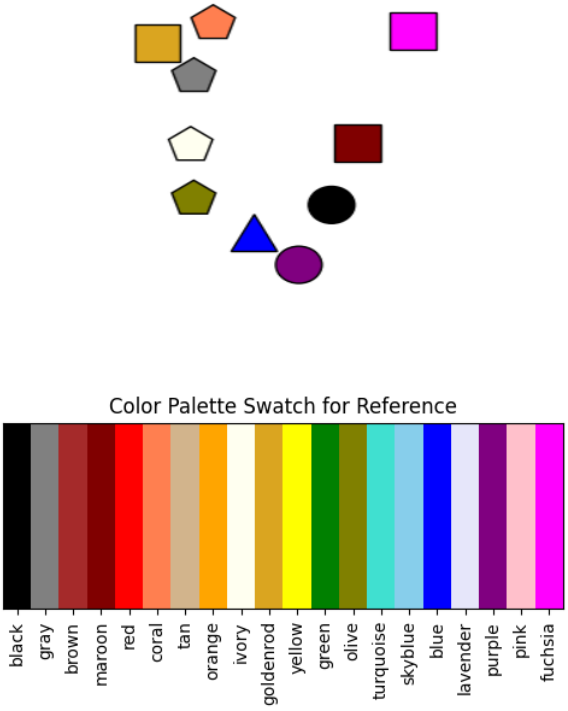

Figure 10: Colours Present

A.1.6 Comparing Size

Input Prompt:

- This is an image with several rows of circles where each row is above a horizontal line. - The image may also contain a single row with a blue and a green circle. - 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 - Count the total number of red circles in the image - Last line of the output must be of the form COUNT:x

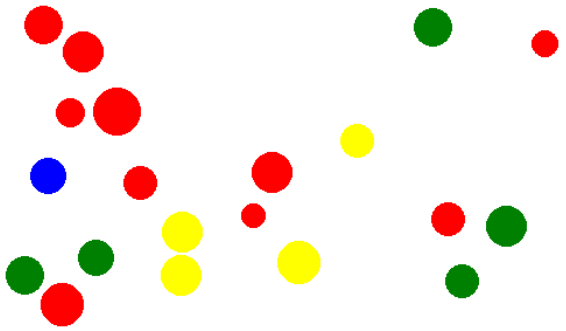

Figure 11: Count Colored Circles

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.

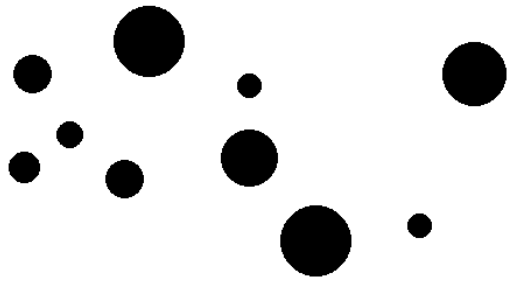

Figure 12: Counting Circles

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

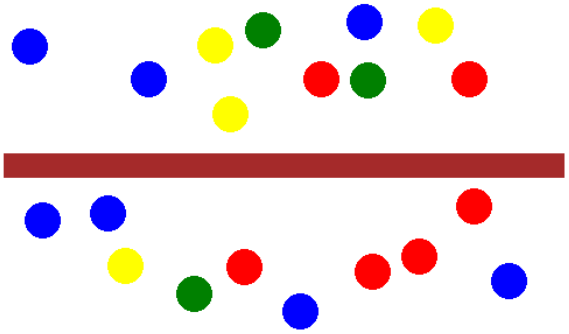

Figure 13: Counting Locations

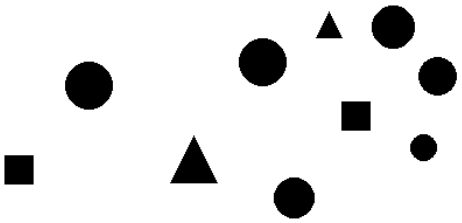

Figure 14: Counting Shapes

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. - The graph may contain a single blue node with no edges. - The graph may be disconnected. - Count the number of nodes and edges in the graph. - Last line of the output must be of the form NODES:x EDGES:y.

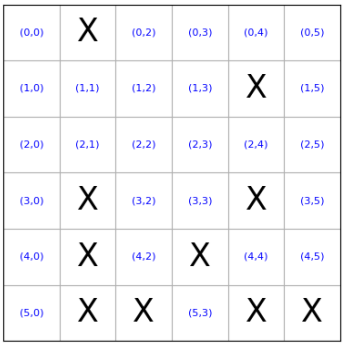

Figure 15: Cross and Knots

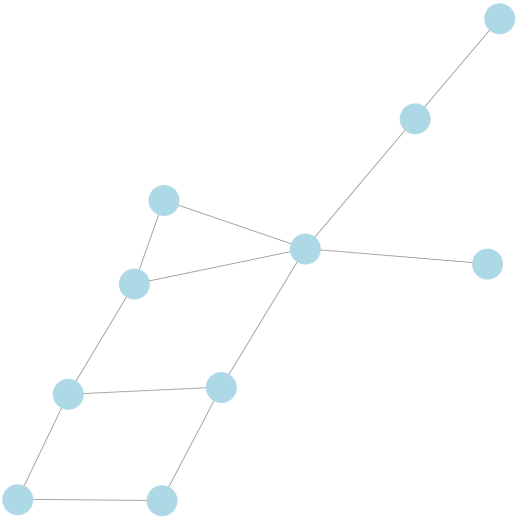

Figure 16: Graph Counting

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 shapes present in the S and E cells should also be included in the sequence. - 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. - The image may also contain a single shape.

- Identify the shape present in each row. - In case there is a single shape, treat it as a single 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

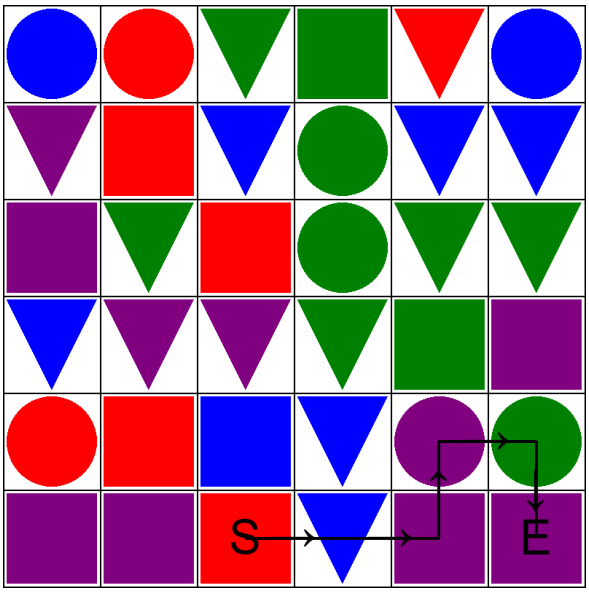

Figure 17: Grid Path

A.1.15 Inside Circles

Input Prompt:

- This is an image containing some black circles and a single red dot - Determine if the red dot is contained inside any circle in the image - 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 ANSWER: x where x is Yes if the dot is contained in any circle and No otherwise.

A.1.16 Layered Colors

Input Prompt:

- This image contains some concentric circles in different colours. - The image may also contain a single circle - Write down the sequence of colours seen from inside to outside. - In case of a single circle in the image, write down the colour of that circle. - 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,

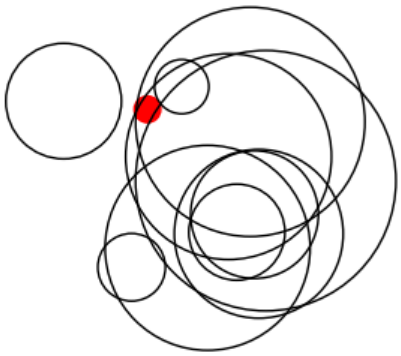

Figure 18: Inside Circles

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 shapes colud be of the following types - [circle, diamond, hexagon, octagon] - 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

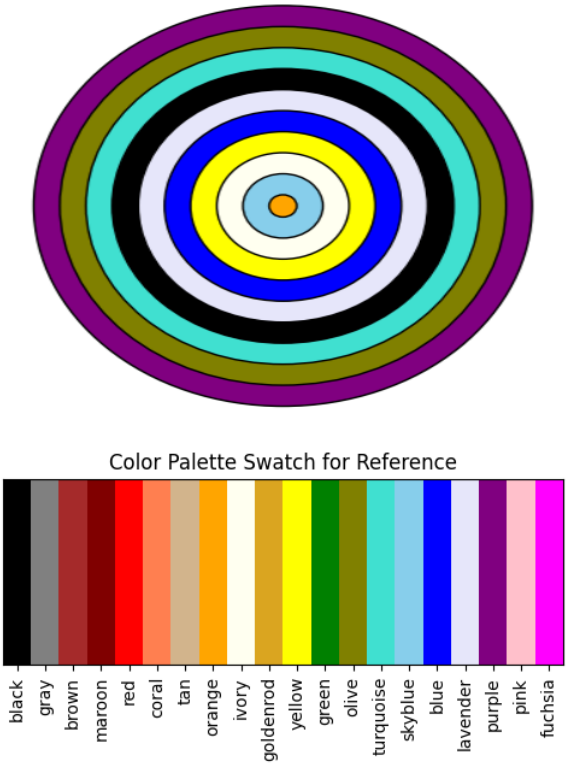

Figure 19: Layered Colors

A.1.19 List Shapes

Input Prompt:

- This is an image with some multi-coloured shapes drawn on top of it - The shapes belong to the following list - [circle, triangle, square, pentagon] - 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) with the comma and bracket

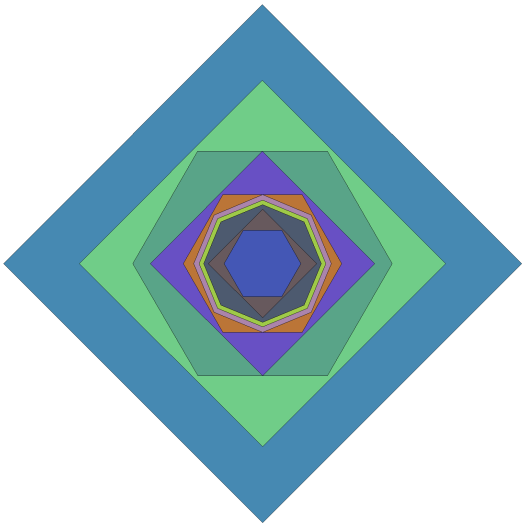

Figure 20: Layered Shapes

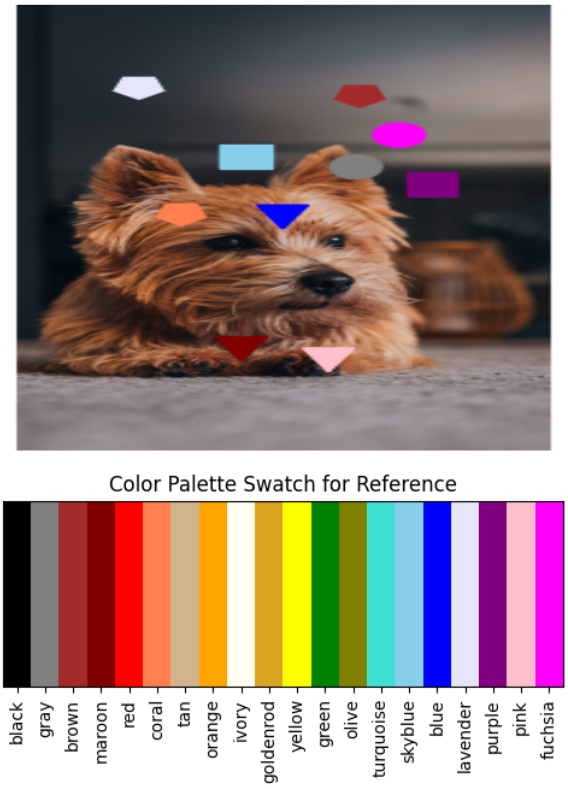

Figure 21: List Colors

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 two images may also contain a single shape each - The first image consists of rows of shapes and the second image consists of rows of outlines of shapes - Determine the number of shapes in the first image that match the outlines in the corresponding row in the second image. - In case each of the two images contain only a single shape, treat the single shape in each image as a row. - 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 two images may also contain a single shape each - The first image consists of rows of shapes and the second image consists of rows of shadows of those shapes. - Determine the number of shapes in the first image that match the shadows in the corresponding row in the second image. - In case each of the two images contain only a single shape, treat the single shape in each image as a row. - 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.

A.1.25 Mirror 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

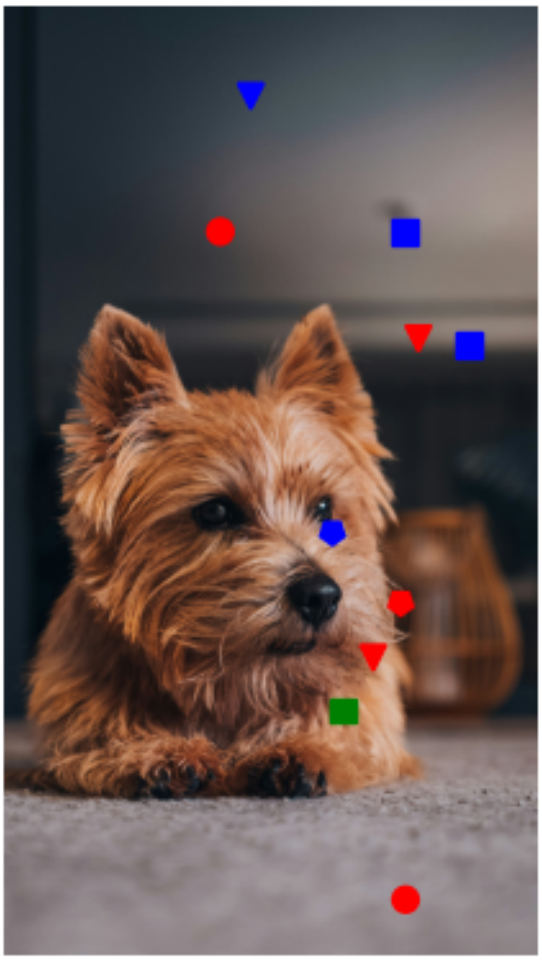

Figure 22: List Shapes

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.1.26 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.1.27 Sort Circles

Input Prompt:

- There are several black circles with different sizes. - The image may also have a single black circle. - Sort these circles by size, from smallest to largest.

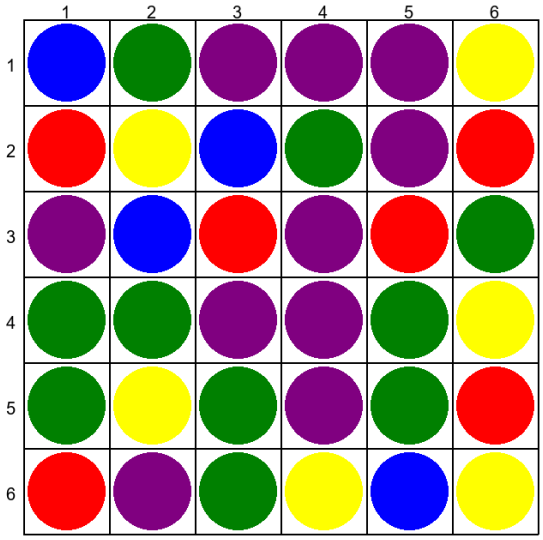

Figure 23: Locate Circles Color

- If there is only a single circles, give the number written on that circle. - Output must only contain the space separated labels of circles in their sorted order by size.

A.1.28 Sort Lines

Input Prompt:

- The image has several parallel lines of varying lengths. - The image may also contain only a single line. - Sort these lines by length, from shortest to longest. - If there is only a single line, give the number of that line. - 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.1.29 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 second image - The output must be given in a single line in the form of COUNT:x

A.1.30 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

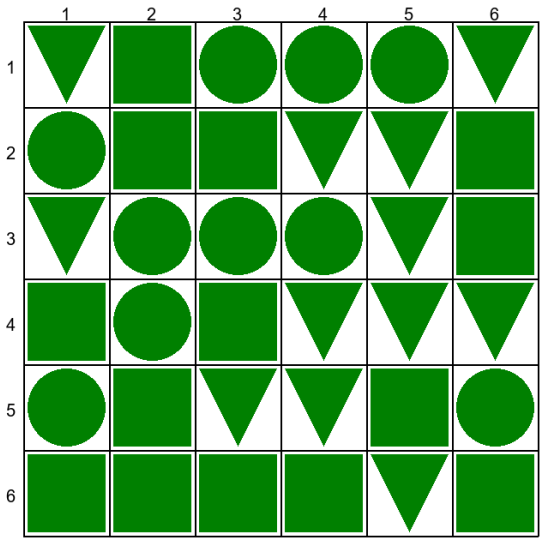

Figure 24: Locate Circles Shape

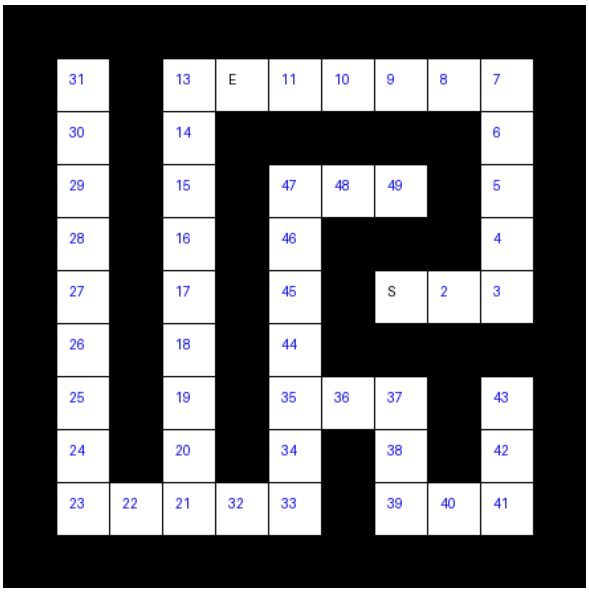

Figure 25: Maze Solving

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 Overall Performance in Human Study

Model	Overall performance
GPT-5-mini	56.81%
GPT-4o	31.94%
o4-mini	56.83%
Gemini	55.58%
Qwen	14.69%
Deepseek	7.08%
Human	**95.43%**

Table 5: Overall performance in human study

A.3 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.4 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.5 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.6 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.

A.7 One-shot Results

Skill / Domain	GPT-5-mini	GPT-4o	o4-mini	Gemini	Avg. Acc.
— Performance by Skill —					
Visual Discrimination	51.46	11.17	52.51	43.57	39.68
Visual Memory	46.55	20.93	45.96	39.38	38.2
Visual Sequential Memory	38.61	12.33	45.48	33.42	32.46
Visual Figure Ground	63.9	20.0	63.9	48.83	49.16
Visual Form Constancy	53.06	16.64	45.63	40.7	39.01
Visual Closure	40.58	17.75	50.76	40.96	37.51
Visual Spatial Relationship	64.61	24.5	61.59	64.11	53.7
Average of skills	51.25	17.62	52.26	44.42	41.39
— Performance by Domain —					
change_colour	30.15	11.5	24.62	21.11	21.84
circle_boxes	26.63	21.5	28.64	24.12	25.22
circle_location	55.78	13.0	48.24	54.27	42.82
circle_right_triangle	100.0	59.0	100.0	96.48	88.87
colours_present	4.02	1.5	5.03	1.01	2.89
comparing_size	17.59	2.5	23.62	36.18	19.97
count_coloured_circles	41.71	16.5	35.68	46.73	35.16
counting_circles	57.79	23.5	48.74	53.27	45.82
counting_locations	35.68	9.0	35.18	28.64	27.12
counting_shapes	60.8	20.2	50.75	56.28	47.01
cross_and_knots	100.0	21.0	100.0	91.46	78.11
graph_counting	33.17	3.0	27.14	31.16	23.62
grid_path	51.76	0.5	46.23	0.5	24.75
identifying_shapes	67.84	41.5	91.46	51.76	63.14
inside_circles	94.97	52.0	92.96	95.48	83.85
layered_colours	31.16	25.0	34.17	24.62	28.74
layered_shapes	21.11	12.5	19.1	19.1	17.95
list_colours	40.7	12.0	40.7	36.18	32.4
list_shapes	91.96	68.5	91.96	82.91	83.83
locate_circles_colour	99.5	1.0	98.49	63.32	65.58
match_outline	58.79	17.5	79.4	71.86	56.89
match_shadow	51.26	16.0	70.35	48.24	46.46
maze_solving	48.24	15.0	31.16	46.23	35.16
mirror_image	30.15	17.5	27.14	20.1	23.72
numbered_shapes	81.91	9.5	77.39	53.77	55.64
sort_circles	34.67	16.5	33.17	25.63	27.49
sort_lines	28.64	19.5	23.62	23.62	23.85
vanishing_objects	34.17	6.0	22.11	22.61	21.22
water_image	34.67	12.5	24.12	20.1	22.85
Full Percept-V Dataset	52.13	18.21	51.01	43.79	41.28

Table 6: Comparison of LLM performance across different skills and domains in one-shot setting. All values are percentages.

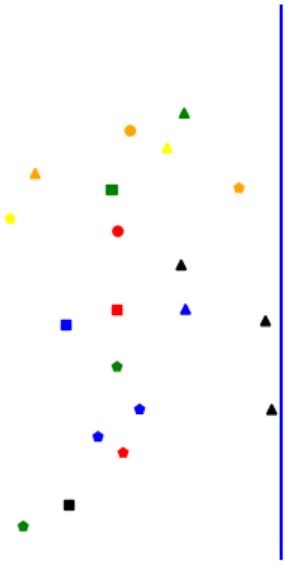

Figure 26: Mirror Image

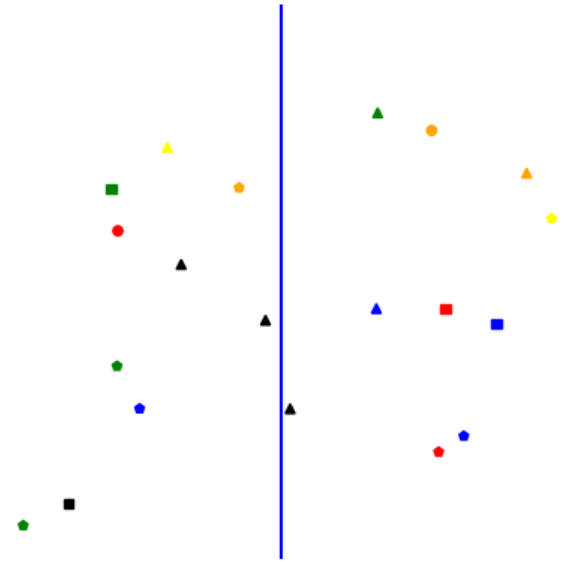

Figure 27: Mirror Image

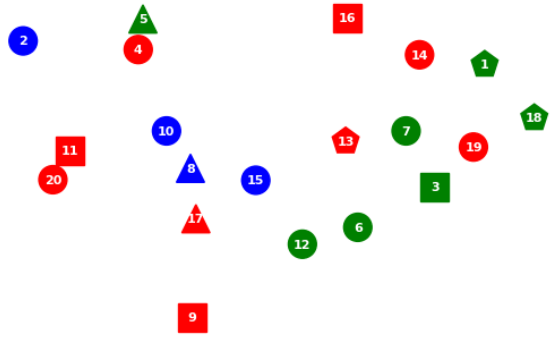

Figure 28: Numbered Images

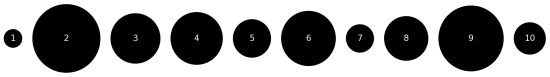

Figure 29: Sort Circles

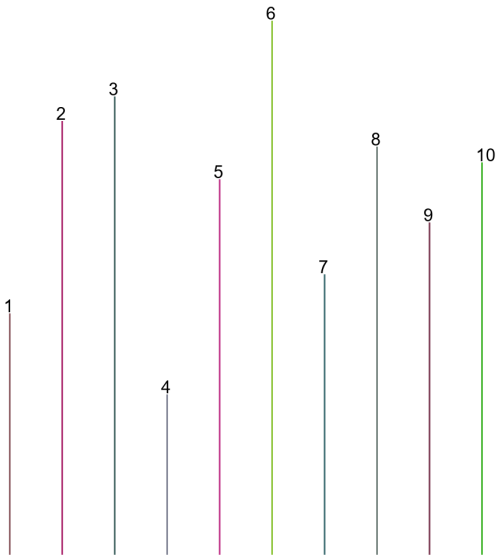

Figure 30: Sort Lines

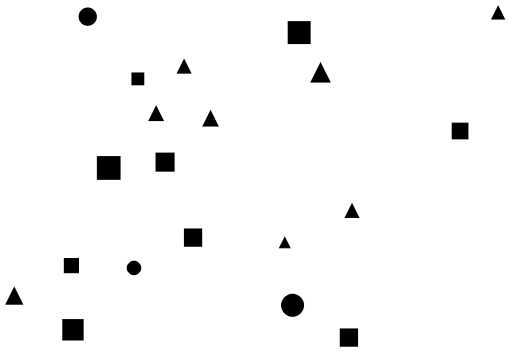

Figure 31: Vanishing Objects

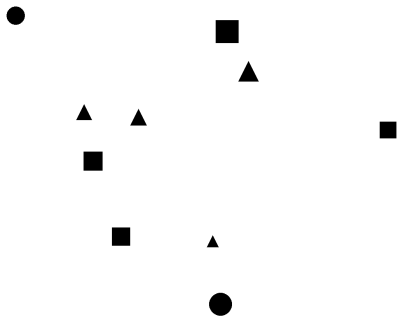

Figure 32: Vanishing Objects

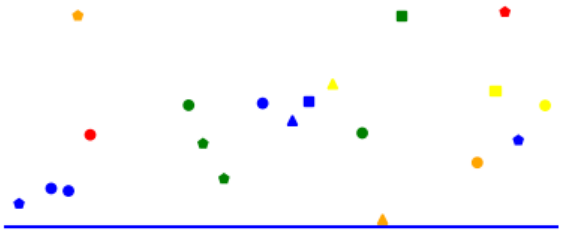

Figure 33: Water Image

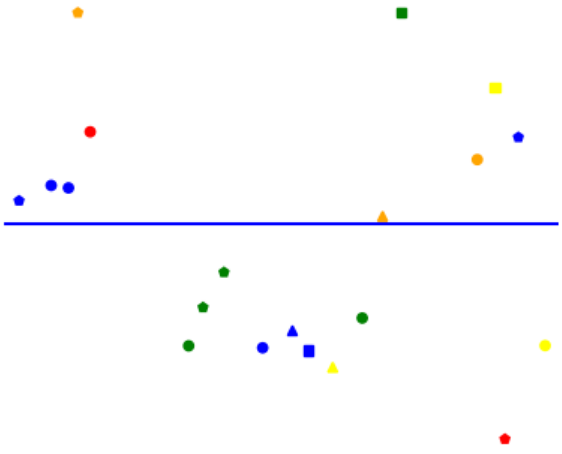

Figure 34: Water Image

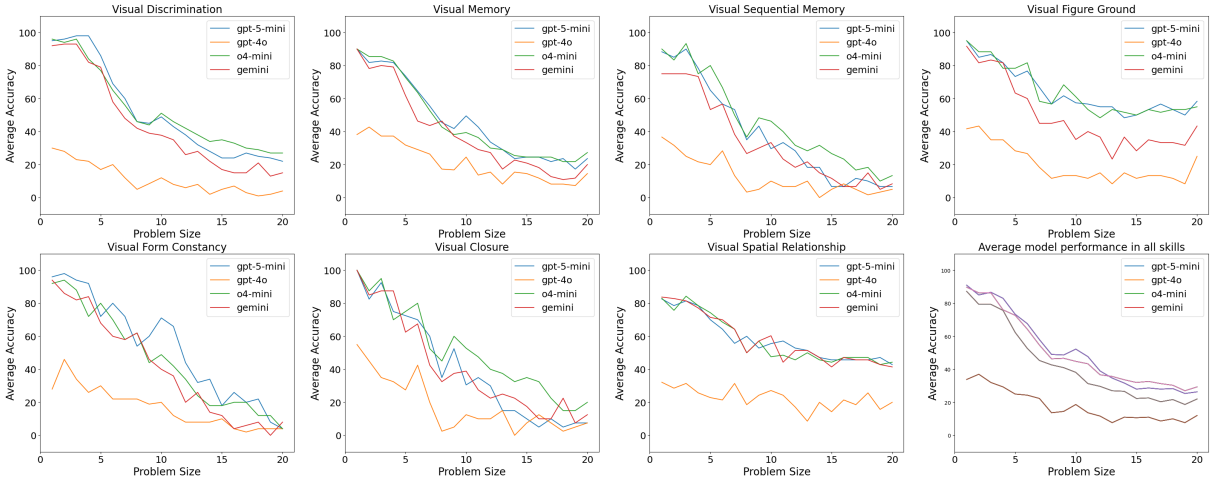

Figure 35: The overall accuracy of all models in different skills in one-shot setting.