

---

# Decomposing Complex Visual Comprehension into Atomic Visual Skills for Vision Language Models

---

Hyunsik Chae<sup>†</sup>, Seungwoo Yoon<sup>†</sup>, Chloe Yewon Chun<sup>†</sup>,  
Gyehun Go<sup>†</sup>, Yongin Cho<sup>†</sup>, Gyeongmin Lee<sup>†</sup>, Ernest K. Ryu<sup>\*</sup>

<sup>†</sup>Seoul National University, <sup>\*</sup>UCLA, Department of Mathematics  
<https://github.com/Atomic-Visual-Skills/AVS>

## Abstract

Recent Vision Language Models (VLMs) have demonstrated impressive multimodal comprehension and reasoning capabilities, but they often struggle with trivially simple visual tasks. In this work, we introduce the Atomic Visual Skills Benchmark (AVSBench) to evaluate whether VLMs possess capabilities to understand basic geometric features, which we refer to as atomic visual skills. Specifically, we systematically categorize the atomic visual skills and handcraft a set of 5,073 diverse questions designed to assess each individual atomic visual skill. Using AVSBench, we evaluate the current leading VLMs and find that they struggle with most of these atomic visual skills that are obvious to humans.

## 1 Introduction

Recent Vision Language Models (VLMs), also referred to more generally as Multimodal Large Language Models (MLLM), integrate vision components into language models and demonstrate an impressive breadth of multimodal comprehension and reasoning capabilities [7]. At the same time, however, VLMs often struggle with trivially easy visual tasks as shown in Figure 1, a puzzling phenomenon that seems almost contradictory to their remarkable performance [11, 43]. We propose two hypotheses to explain the observed shortcomings of current vision-language models:

*Hypothesis 1: The comprehension of complex visual diagrams requires the composition of smaller atomic visual skills.*

*Hypothesis 2: Current vision language models are incapable of such atomic visual skills.*

In this work, we introduce the Atomic Visual Skills Benchmark (AVSBench) to test *Hypothesis 2*. AVSBench is designed to rigorously evaluate VLMs’ ability to comprehend fundamental geometric features, which we refer to as *atomic visual skills*. We systematically categorize 36 atomic visual skills that encompass diagrams arising in high school-level geometry and *handcraft* a set of 5,073 diverse questions designed to assess the understanding of the individual atomic visual skills.

We then evaluate the state-of-the-art VLMs on AVSBench, and the results clearly support *Hypothesis 2*. While our problems are designed to be trivial to humans, VLMs struggle; state-of-the-art models like Gemini-1.5-pro and GPT-4o score around 70%-75% on problems with the “easy” categorization, score around 60% on the “medium” problems, and 30% on “hard” problems. The confirmation of *Hypothesis 2* also lends support to *Hypothesis 1*, and suggests a promising direction of future work of training vision language models specifically on atomic visual skills to improve their performance in comprehending complex visual diagrams.

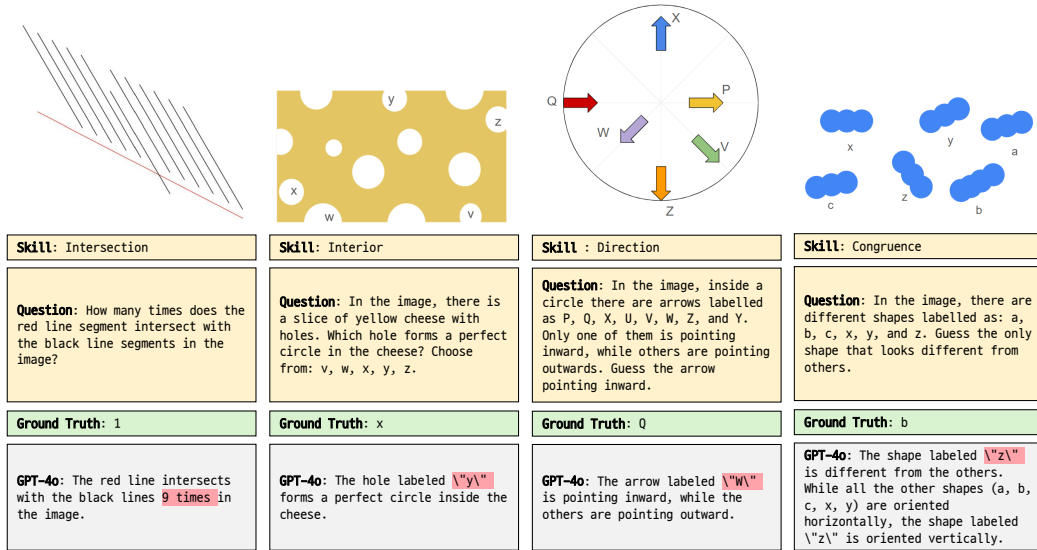


Figure 1: Examples of AVSBench problems and responses by GPT-4o. Other state-of-the-art models exhibit similar failures. These examples demonstrate a deficiency in the VLMs’ understanding of basic geometric concepts.

## 2 Atomic Visual Skills Benchmark (AVSBench)

Many visual reasoning tasks in existing benchmarks, such as the ones listed in Section A, are composite tasks that can be broken down into more elementary components. This observation leads us to define a set of *atomic visual skills* based on the following criteria: (i) each skill is intuitive and trivial for adult humans, (ii) each skill cannot be decomposed further, or doing so would be unnatural, and (iii) the list of atomic visual skills should comprehensively cover the abilities required for comprehending geometric diagrams arising in high-school level mathematics. While this definition is not a fully rigorous one, we found it to be sufficiently clear and substantive for our work.

Using these criteria, we identified 36 atomic visual skills, including the ability to understand concepts such as angle, boundary, orthogonality, curvature, and direction. The complete list and further illustrations are provided in D.

For adult humans, these skills are trivially simple and require little to no reasoning to perform. Therefore, we use the term *comprehension* instead of *reasoning* to emphasize our belief that these skills do not require much reasoning or thinking to perform, for both humans and VLMs. This belief is partially supported by Findings 3 of Section 3.1.

We then constructed the Atomic Visual Skills Benchmark (AVSBench) to evaluate VLMs’ ability to perform the 36 atomic visual skills. AVSBench, as summarized in Figure 2, comprises 5,073 new handcrafted image-question-answer triplets with the following characteristics:

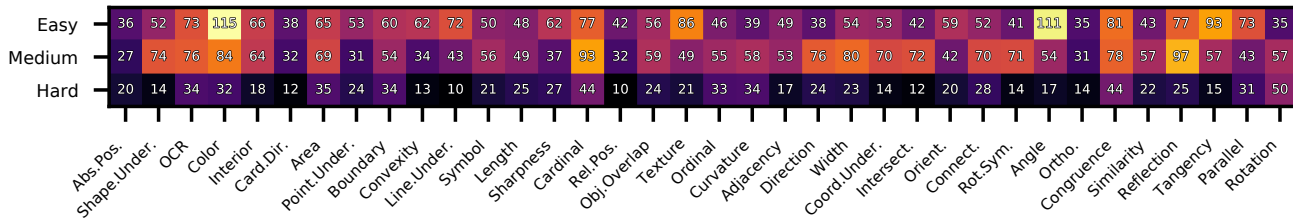


Figure 2: List of 36 atomic visual skills and the number of easy, medium, and hard problems for each skill. The difficulty is judged by the authors. We provide a total of 5,073 new handcrafted problems.

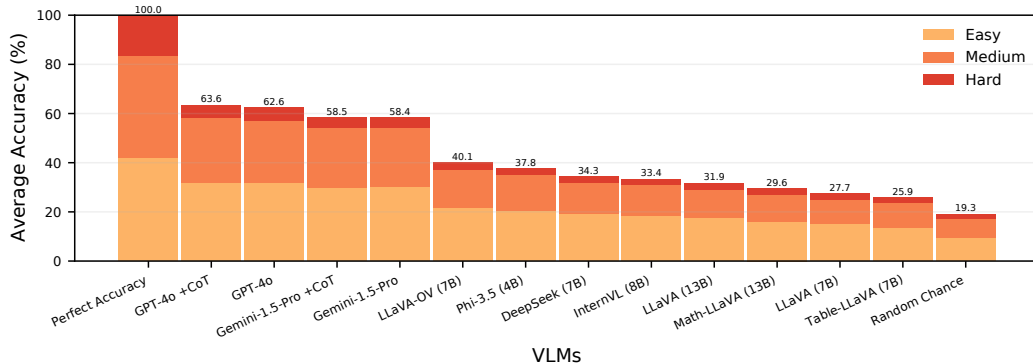


Figure 3: Evaluation results on AVSBench. +CoT implies the performance of the model on the right with chain-of-thought (CoT) prompting [22]. The area ratios of each colored section are aligned with the actual ratio of problem counts. Details about the models including their full name are in E. Full quantitative results are illustrated in Table 5.

- **Originality.** All images and questions are newly generated, ensuring that they are free from data contamination concerns.
- **Diversity.** Although we focus on the set of only 36 skills, the problems feature diverse expressions and formats, as illustrated by the sample problems in C.
- **Skill isolation.** Each question targets a specific atomic skill, minimizing the overlap with other skills. Recognizing the impossibility of achieving complete isolation, our method incorporates diverse tasks to mitigate the influence of any task or their relevant overlapping skills. For instance, to minimize the influence of other skills while evaluating the cardinal skill, we asked about cardinals of various concepts and objects, from colors to points, lines, and other figures.
- **Focus on high school geometry.** We focus on the visual skills required to solve high school-level geometry problems for the following reasons: (i) the scope of the high school mathematics curriculum is more or less clearly defined, (ii) (as our results of Section 3 show) these atomic visual skills are sufficiently challenging for current VLMs, (iii) the range of skills is broad enough to be applicable to other visual comprehension tasks, such as interpreting charts, tables, and scientific or mathematical figures.

### 3 Current vision language models struggle with atomic visual skills

We evaluate three types of VLMs on AVSBench: (i) state-of-the-art proprietary models: GPT-4o [40, 39] and Gemini-1.5-pro [49], (ii) popular mid-sized open-weight models: LLaVA-Next (7B, 13B) [30], LLaVA-OneVision (7B) [26], Phi-3.5-Vision (4B) [1], InternVL2 (8B) [10], Deepseek-VL (7B) [31], and (iii) VLMs specifically trained for geometry or other visual data: Math-LLaVA (13B) [47], Table-LLaVA (7B) [60]. Further details of model versions are provided in Section E.

The evaluation protocol consists of three steps. First, we provide the VLM with the image-question pair and solicit a response. As we further discuss later, we also explore the effect of the chain-of-thought (CoT) prompting [54, 22]. Second, we extract the answer from the VLM’s response using GPT-4o mini [40]. Third, we ask GPT-4o mini to score the answer by comparing the extracted answer with the answer key. We award 1 point for a correct answer and 0 points otherwise without any partial credit. More details on our evaluation protocol are provided in F.

#### 3.1 Experimental results and findings

Figure 3 presents the results comparing the selected VLMs and the baseline corresponding to random guessing. Details including exact values are provided in G. On “easy” problems, models with about 10B parameters score between 32.5% and 51.0% while closed-source models, including GPT-4o and Gemini-1.5-pro, achieve over 70%, far above random chance (22.4%). On “medium” problems, models with about 10B parameters score between 23.8% and 37.8%, slightly above random chance

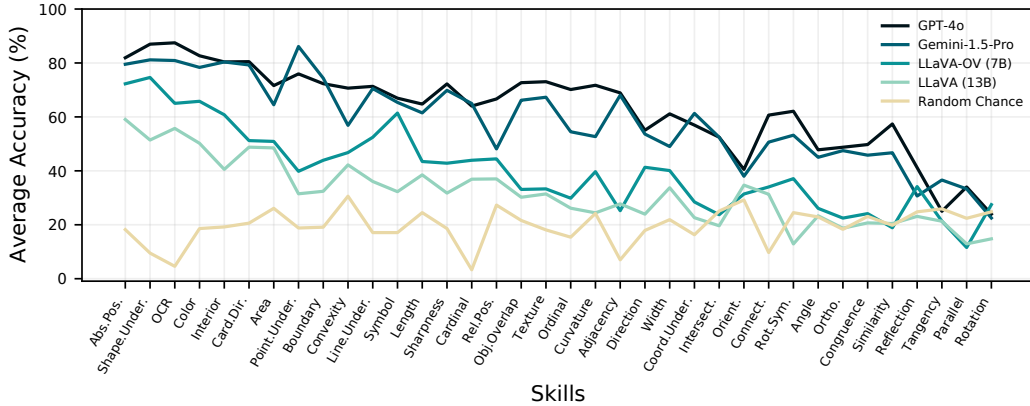


Figure 4: Accuracies of a leading model, 3 outstanding models, and random chance on each skill. The skills are ordered in descending order of accuracy, averaged over all models.

(19.1%). Closed-source models achieve between 58.6% and 64.6%. For “hard” problems, most open models score close to random chance (11.7%). The closed-source models GPT-4o (32.3%) and Gemini-1.5-pro (26.9%) scored significantly better than random chance but clearly struggled.

**Findings 1: Models share strengths and weaknesses.** Figure 4 presents the accuracies of selected models on each skill. The performances across skills varied significantly. For example, most VLMs performed well on OCR, Absolute Position, and Shapes, but performed poorly on tangency, parallel, and angle. Interestingly, the different models largely shared the same set of skills they did well on and the same set they found challenging.

**Findings 2: Domain-specific models are not better.** Surprisingly, Math-LLaVA [47] and Table-LLaVA [60], which are VLMs specifically trained for geometry or visual data, did not perform better than general VLMs of similar size, on almost any skills within AVSBench as the results of Table 5 show.

**Findings 3: Chain-of-thought is not helpful in enhancing atomic visual skills.** We also evaluated the best-performing models—GPT-4o and Gemini-1.5-pro—with chain-of-thought (CoT) prompting [22] on AVSBench. Surprisingly, CoT did not help for most skills, and for some skills, it even worsened the performance, as shown in Table 5. This contrasts with prior work, which found CoT to be beneficial for certain visual reasoning tasks [32, 53]. We attribute this difference to our hypothesis that the atomic visual skills of AVSBench require simple “comprehension” and, therefore, do not benefit from the additional “reasoning” steps afforded by CoT prompting. More concrete comparison, see G.

## 4 Conclusion

We present the Atomic Visual Skills Benchmark (AVSBench), a benchmark designed to rigorously evaluate VLMs’ ability to perform atomic visual skills in a decomposed manner. We then show that current state-of-the-art VLMs struggle with such atomic visual skills.

The failure of VLMs to carry out such simple atomic visual tasks raises the question: How is it that VLMs are sometimes successful at performing complex visual tasks? For this, we hypothesize that the existing impressive performance on complex tasks is due to overfitting or unimodal shortcuts. Indeed, recent studies such as [11, 18, 57, 50] report that VLMs tend to depend on language shortcuts, as we further reference and discuss in Section A of the appendix.

Recall that our *Hypothesis 1* posits that the atomic visual skills are necessary subcomponents for comprehending complex visual diagrams. In future work, we plan to train and fine-tune VLMs directly on the atomic visual skills and ascertain *Hypothesis 1*.

## References

- [1] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, A. Benhaim, M. Bilenko, J. Bjorck, S. Bubeck, M. Cai, Q. Cai, V. Chaudhary, D. Chen, D. Chen, W. Chen, Y.-C. Chen, Y.-L. Chen, H. Cheng, P. Chopra, X. Dai, M. Dixon, R. Eldan, V. Fragoso, J. Gao, M. Gao, M. Gao, A. Garg, A. Del Giorno, A. Goswami, S. Gunasekar, E. Haider, J. Hao, R. J. Hewett, W. Hu, J. Huynh, D. Iter, S. A. Jacobs, M. Javaheripi, X. Jin, N. Karampatziakis, P. Kauffmann, M. Khademi, D. Kim, Y. J. Kim, L. Kurilenko, J. R. Lee, Y. T. Lee, Y. Li, Y. Li, C. Liang, L. Liden, X. Lin, Z. Lin, C. Liu, L. Liu, M. Liu, W. Liu, X. Liu, C. Liu, P. Madan, A. Mahmoudzadeh, D. Majercak, M. Mazzola, C. C. T. Mendes, A. Mitra, H. Modi, A. Nguyen, B. Norick, B. Patra, D. Perez-Becker, T. Portet, R. Pryzant, H. Qin, M. Radmilac, L. Ren, G. de Rosa, C. Rosset, S. Roy, O. Ruwase, O. Saarikivi, A. Saied, A. Salim, M. Santacroce, S. Shah, N. Shang, H. Sharma, Y. Shen, S. Shukla, X. Song, M. Tanaka, A. Tupini, P. Vaddamanu, C. Wang, G. Wang, L. Wang, S. Wang, X. Wang, Y. Wang, R. Ward, W. Wen, P. Witte, H. Wu, X. Wu, M. Wyatt, B. Xiao, C. Xu, J. Xu, W. Xu, J. Xue, S. Yadav, F. Yang, J. Yang, Y. Yang, Z. Yang, D. Yu, L. Yuan, C. Zhang, C. Zhang, J. Zhang, L. L. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, and X. Zhou. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*, 2024.
- [2] Z. Allen-Zhu and Y. Li. Physics of language models: Part 3.2, knowledge manipulation. *arXiv:2309.14402*, 2023.
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. *International Conference on Computer Vision*, 2015.
- [4] S. Arora and A. Goyal. A theory for emergence of complex skills in language models. *arXiv:2307.15936*, 2023.
- [5] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, and C. Sutton. Program synthesis with large language models. *arXiv:2108.07732*, 2021.
- [6] L. Berglund, M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". *International Conference on Learning Representations*, 2024.
- [7] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, M. Ibrahim, M. Hall, Y. Xiong, J. Lebensold, C. Ross, S. Jayakumar, C. Guo, D. Bouchacourt, H. Al-Tahan, K. Padthe, V. Sharma, H. Xu, X. E. Tan, M. Richards, S. Lavoie, P. Astolfi, R. A. Hemmat, J. Chen, K. Tirumala, R. Assouel, M. Moayeri, A. Talattof, K. Chaudhuri, Z. Liu, X. Chen, Q. Garrido, K. Ullrich, A. Agrawal, K. Saenko, A. Celikyilmaz, and V. Chandra. An introduction to vision-language modeling. *arXiv:2405.17247*, 2024.
- [8] J. Cao and J. Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. *International Conference on Computational Linguistics*, 2022.
- [9] J. Chen, T. Li, J. Qin, P. Lu, L. Lin, C. Chen, and X. Liang. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. *Computer Vision and Pattern Recognition*, 2022.
- [10] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to GPT-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv:2404.16821*, 2024.
- [11] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. BLINK: Multimodal large language models can see but not perceive. *arXiv:2404.12390*, 2024.
- [12] O. Golovneva, Z. Allen-Zhu, J. Weston, and S. Sukhbaatar. Reverse training to nurse the reversal curse. *arXiv:2403.13799*, 2024.
- [13] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Computer Vision and Pattern Recognition*, 2017.

- [14] D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. K. Li, F. Luo, Y. Xiong, and W. Liang. DeepSeek-Coder: When the large language model meets programming – the rise of code intelligence. *arXiv:2401.14196*, 2024.
- [15] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. *Computer Vision and Pattern Recognition*, 2018.
- [16] M. Hanna, O. Liu, and A. Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *arXiv:2305.00586*, 2023.
- [17] T. He, D. Doshi, A. Das, and A. Gromov. Learning to Grok: Emergence of in-context learning and skill composition in modular arithmetic tasks. *arXiv:2406.02550*, 2024.
- [18] C.-Y. Hsieh, J. Zhang, Z. Ma, A. Kembhavi, and R. Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv:2306.14610*, 2023.
- [19] K. Kafle, B. Price, S. Cohen, and C. Kanan. DVQA: Understanding data visualizations via question answering. *Computer Vision and Pattern Recognition*, 2018.
- [20] M. Kazemi, H. Alvani, A. Anand, J. Wu, X. Chen, and R. Soricut. GeomVerse: A systematic evaluation of large models for geometric reasoning. *arXiv:2312.12241*, 2023.
- [21] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. *arXiv:1603.07396*, 2016.
- [22] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv:2205.11916*, 2022.
- [23] B. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International Conference on Machine Learning*, 2018.
- [24] N. Lee, K. Sreenivasan, J. D. Lee, K. Lee, and D. Papailiopoulos. Teaching arithmetic to small transformers. *International Conference on Learning Representations*, 2024.
- [25] M. Lewis, N. V. Nayak, P. Yu, Q. Yu, J. Merullo, S. H. Bach, and E. Pavlick. Does CLIP bind concepts? probing compositionality in large image models. *arXiv:2212.10537*, 2022.
- [26] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li. LLaVA-OneVision: Easy visual task transfer. *arXiv:2408.03326*, 2024.
- [27] Z. Lin, X. Chen, D. Pathak, P. Zhang, and D. Ramanan. Revisiting the role of language priors in vision-language models. *arXiv:2306.01879*, 2024.
- [28] Z. Lin and K. Lee. Dual operating modes of in-context learning. *arXiv:2402.18819*, 2024.
- [29] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024.
- [30] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Neural Information Processing Systems*, 2023.
- [31] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, Y. Sun, C. Deng, H. Xu, Z. Xie, and C. Ruan. DeepSeek-VL: Towards real-world vision-language understanding. *arXiv:2403.05525*, 2024.
- [32] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *International Conference on Learning Representations*, 2024.
- [33] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna. CREPE: Can vision-language foundation models reason compositionally? *Computer Vision and Pattern Recognition*, 2023.

- [34] A. Masry, X. L. Do, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *Findings of the Association for Computational Linguistics*, 2022.
- [35] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. PlotQA: Reasoning over scientific plots. *Conference on Applications of Computer Vision*, 2020.
- [36] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv:2202.12837*, 2022.
- [37] M. Okawa, E. S. Lubana, R. Dick, and H. Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Neural Information Processing Systems*, 2023.
- [38] S. Ontanón, J. Ainslie, V. Cvicek, and Z. Fisher. Making transformers solve compositional tasks. *arXiv:2108.04378*, 2021.
- [39] OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2024.
- [40] OpenAI. GPT-4o system card. <https://openai.com/research/gpt-4o-system-card>, Aug 2024.
- [41] R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, and T. Dekel. Teaching CLIP to count to ten. *International Conference on Computer Vision*, 2023.
- [42] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv:2210.03350*, 2022.
- [43] P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen. Vision language models are blind. *arXiv:2407.06581*, 2024.
- [44] R. Ramesh, E. S. Lubana, M. Khona, R. P. Dick, and H. Tanaka. Compositional capabilities of autoregressive transformers: A study on synthetic, interpretable tasks. *International Conference on Machine Learning*, 2024.
- [45] B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Remez, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bitton, M. Bhatt, C. C. Ferrer, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve. Code Llama: Open foundation models for code. *arXiv:2308.12950*, 2023.
- [46] J. Shen, Y. Yuan, S. Mirzoyan, M. Zhang, and C. Wang. Measuring vision-language stem skills of neural models. *International Conference on Learning Representations*, 2024.
- [47] W. Shi, Z. Hu, Y. Bin, J. Liu, Y. Yang, S.-K. Ng, L. Bing, and R. K.-W. Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv:2406.17294*, 2024.
- [48] J. Song, Z. Xu, and Y. Zhong. Out-of-distribution generalization via composition: a lens through induction heads in transformers. *arXiv:2408.09503*, 2024.
- [49] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*, 2024.
- [50] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *Computer Vision and Pattern Recognition*, 2022.
- [51] S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, A. Wang, R. Fergus, Y. LeCun, and S. Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *arXiv:2406.16860*, 2024.
- [52] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. *Computer Vision and Pattern Recognition*, 2024.

- [53] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. *International Conference on Machine Learning*, 2024.
- [54] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and Denny Zhou. Chain-of-Thought prompting elicits reasoning in large language models. *Neural Information Processing Systems*, 2024.
- [55] Z. Xu, Z. Shi, and Y. Liang. Do large language models have compositional ability? an investigation into limitations and scalability. *ICLR Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.
- [56] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *International Conference on Learning Representations*, 2023.
- [57] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K. Chang, P. Gao, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? *arXiv:2403.14624*, 2024.
- [58] H. Zhao, S. Kaur, D. Yu, A. Goyal, and S. Arora. Can models learn skill composition from examples? *ICML Workshop on LLMs and Cognition*, 2024.
- [59] T. Zhao, T. Zhang, M. Zhu, H. Shen, K. Lee, X. Lu, and J. Yin. VL-Checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv:2207.00221*, 2022.
- [60] M. Zheng, X. Feng, Q. Si, Q. She, Z. Lin, W. Jiang, and W. Wang. Multimodal table understanding. *arXiv:2406.08100*, 2024.



## A Prior works

**VLM benchmarks and language shortcuts.** Existing VLM benchmarks evaluate models on their ability to solve diverse vision-language problems from general real-world tasks [3, 13, 15], tasks that require specific skills such as high-school geometry [32, 20, 9, 8], analyzing charts and tables [34, 60, 35], and other scientific visual data [19, 21]. However, most VLM benchmarks do not contain a mechanism for verifying whether a correct solution is based on correctly comprehending the visual information, allowing the models to sometimes rely on linguistic biases to find a solution [7]. Lin et al. [27] revealed that by simply avoiding implausible or less fluent sentences, blind language models can distinguish the correct description of an image from wrong ones on CREPE [33], VL-Checklist [59], and ARO[56]. Mathverse [57] observed that, when solving geometry problems, VLMs rely mostly on textual inputs without correctly interpreting diagrams.

Some recent work has started to seek unbiased ways to measure visual capabilities. Winoground [50] prevents choosing image captions based on the plausibility of the sentence structure, by providing two images with same objects or concepts but with different relationships. Blink [11] and CV-Bench [51] present novel vision-oriented tasks with minimized effects of linguistic biases.

**Compositional reasoning.** There has been intensive recent research on the compositional capabilities of Language Models [4, 55, 17, 48, 44, 58, 23, 38, 42]. VLMs have additionally shown compositional capabilities in visual tasks [25, 37, 33, 59, 56]. However, such studies left the visual portion with less attention, thus vulnerable to linguistic shortcuts such as removing grammatically wrong sentences or choosing more realistic sentences as answers. To mitigate this issue, Sugar-Crepe [18] generated sentences with ChatGPT to provide incorrect captions of given images, with different compositional structures while as realistic as the ground truths.

**Research on atomic skills of LLMs.** To understand the capabilities of LLMs, there has been prior work on studying LLMs in simple idealized experiments. This includes research on in-context learning [28, 36], arithmetic (addition and multiplication) [38, 16, 24], fact search and reverse fact search [2, 6, 12], and programming [5, 45, 14].

However, there have been far fewer studies of this kind for vision language models. Paiss et al. [41] focused on counting objects in image and suggested CountBench. Shen et al. [46] suggest a skill-based approach to evaluating VLMs, but their list of skills is not atomic. CV-Bench [51] evaluates 4 vision-centric skills: spatial relationship, object count, depth order, and relative distance. MMVP [52] challenges VLMs to understand 9 visual patterns.

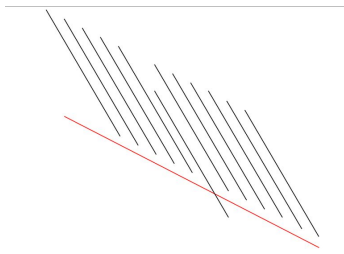
BlindTest [43] observed failures of VLMs with 7 simple tasks focusing on fundamental geometric features, some of which share similar approaches with AVSBench. While these tasks are novel and effective, they lack diversity in color, shape, or word choice. Their Task 1 for instance, uses only one red and one blue line, each with exactly one sharp turn. For Task 3, they adopted only three types of strings and a red circle to generate visual context. Moreover, these 7 tasks are insufficient to collectively evaluate the full spectrum of visual capabilities, leading to limited scope in evaluation objectives. In contrast, our AVSBench offers a systematic, comprehensive framework for evaluating a comprehensive set of visual skills of VLMs with an extensive dataset that is rich in color, shape, and other variations.

## B Problem difficulty categorization

The problems are categorized into three difficulty levels: easy, medium, and hard. Problems categorized as easy or medium should be quickly solvable by humans, whereas hard questions, although more time-consuming, are designed to be clear and easily verifiable. We clarify that the difficulty levels were determined by the authors, so there is a degree of subjectivity to the categorization.

## C Sample problems

We present 99 sample problems to provide the readers with the general characteristics of our AVS-Bench dataset, along with the responses of Gemini-1.5-pro[49], GPT-4o[40], and LLaVA-Next-13B[29]. For readability, we report a summary rather than the full text of the model response. See Figure 5 for examples of full model responses.



<b>Question</b> How many times does the red line segment intersect with the black line segments in the image?
<b>Ground Truth</b> 1
<b>GPT-4o</b> <b>Response</b> : The red line intersects with the black lines 9 times in the image. <b>Extracted Answer</b> : 9 <b>Score</b> : 0
<b>Gemini-1.5-pro</b> <b>Response</b> : The red line intersects the black lines <b>**9**</b> times. \n <b>Extracted Answer</b> : 9 <b>Score</b> : 0
<b>LLaVA-Next-13B</b> <b>Response</b> : The red line intersects with the black lines at <b>three</b> points. <b>Extracted Answer</b> : 3 <b>Score</b> : 0

Figure 5: Full responses of VLMs and scoring by GPT-4o mini.

**Question**  
Choose the most appropriate color to fill in the box marked with '?' in the image. The answer is one of 'a', 'b', 'c', or 'd'.

<b>Ground Truth</b> a	<b>Gemini</b> a
<b>GPT</b> a	<b>LLaVA</b> b

**Question**  
In the image, all but one shape is congruent to the others, meaning they can be perfectly overlapped by moving, rotating, and flipping. Choose the one shape that looks different. Answer from: a, b, c, d, e, f.

<b>Ground Truth</b> c	<b>Gemini</b> d
<b>GPT</b> d	<b>LLaVA</b> e

**Question**  
In the image, there are different shapes labeled as a, b, c, x, y, and z. Identify the one shape that looks different from the others.

<b>Ground Truth</b> y	<b>Gemini</b> z
<b>GPT</b> y	<b>LLaVA</b> b

**Question**  
In the image, there are different shapes labeled as a, b, c, x, y, and z. Guess which shape is different from the others.

<b>Ground Truth</b> b	<b>Gemini</b> z
<b>GPT</b> z	<b>LLaVA</b> x

**Question**  
In the image, there are triangles labeled with numbers or letters. Which numbered triangle is congruent to triangle X?

<b>Ground Truth</b> 4	<b>Gemini</b> 3
<b>GPT</b> 1	<b>LLaVA</b> 1

**Question**  
I gave you a graph as an image. How many connected components are there? Your answer should be a single number. For example, if there are 8 connected components, your answer should be "8".

<b>Ground Truth</b> 3	<b>Gemini</b> 4
<b>GPT</b> 2	<b>LLaVA</b> 4

**Question**  
In the image, there are 9 nodes connected to each other. Guess which numbered node is not connected to any other nodes.

<b>Ground Truth</b> 4	<b>Gemini</b> 4
<b>GPT</b> 4	<b>LLaVA</b> 1

**Question**  
Which color's circle is connected to the red circle through black lines in the image? Choose from: pink, orange, yellow, green, blue, purple, black.

<b>Ground Truth</b> Green	<b>Gemini</b> Blue
<b>GPT</b> Green	<b>LLaVA</b> Yellow(2/3), Pink(1/3)

**Question**  
In the image, there are nodes labeled A through E and 1 through 3, connected to each other by straight lines. Among A, B, C, D, and E, which node is not directly connected to node 2?

<b>Ground Truth</b> A	<b>Gemini</b> E
<b>GPT</b> A	<b>LLaVA</b> A

Figure 6: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 1/11.

**Question**  
You can see disjoint green rectangles in the image. Which green rectangle has its exact boundary marked in black? Answer from a, b, c, d, e, f, g.

<b>Ground Truth</b> g	<b>Gemini</b> c
<b>GPT</b> f	<b>LLaVA</b> d

**Question**  
How many vertices are there in the polygon provided?

<b>Ground Truth</b> 9	<b>Gemini</b> 10
<b>GPT</b> 10	<b>LLaVA</b> 6

**Question**  
How many straight lines are in the image?

<b>Ground Truth</b> 5	<b>Gemini</b> 6
<b>GPT</b> 7	<b>LLaVA</b> 4

**Question**  
I divided the photo on the left into pieces, as shown on the right. How many pieces are there after the split?

<b>Ground Truth</b> 5	<b>Gemini</b> 5
<b>GPT</b> 5	<b>LLaVA</b> 4

**Question**  
In the image, there are red, blue, green, purple, and yellow shapes, with points A and B near each shape. For which color's shape is the shortest path from A to B counterclockwise?

<b>Ground Truth</b> yellow	<b>Gemini</b> red
<b>GPT</b> red	<b>LLaVA</b> purple

**Question**  
In the image, there are six crosses with arrows. Among them, one arrow is rotating in the opposite direction compared to the other five. Find the one that rotates in a different orientation from the others. Answer with the letter that denotes it.

<b>Ground Truth</b> E	<b>Gemini</b> F
<b>GPT</b> C	<b>LLaVA</b> Bottom Right

**Question**  
In the image, there are five arrows on a shape. Each arrow points either clockwise or counterclockwise. Which arrow is NOT pointing the same orientation with the arrow A? Answer with single letter.

<b>Ground Truth</b> E	<b>Gemini</b> B
<b>GPT</b> B	<b>LLaVA</b> c

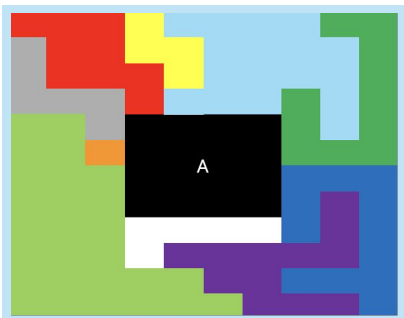
**Question**  
When you start from the black heart in the image, following which curve makes you run counterclockwise? Choose from: red, yellow, green, blue, purple.

<b>Ground Truth</b> Purple	<b>Gemini</b> Green
<b>GPT</b> Purple	<b>LLaVA</b> Red

**Question**  
There is a gradually changing colors in the shapes on the first row. What will be the color best fits in a shape labeled "D"? Choose one from "A", "B", "C", "D" and "E".

<b>Ground Truth</b> B	<b>Gemini</b> B
<b>GPT</b> D	<b>LLaVA</b> D

Figure 7: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 2/11.

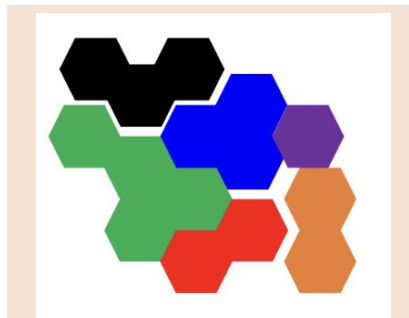


**Question**

In this problem, we say two areas are adjacent to each other if one is directly above, below, to the left, to the right, or diagonal to the other.

Answer with all the colors of the areas that are not adjacent to box A: red, orange, yellow, light green, green, light blue, blue, purple, gray, and white.

<b>Ground Truth</b> Yellow, Purple	<b>Gemini</b> Yellow, Light Blue
<b>GPT</b> Gray, Blue	<b>LLaVA</b> All of them

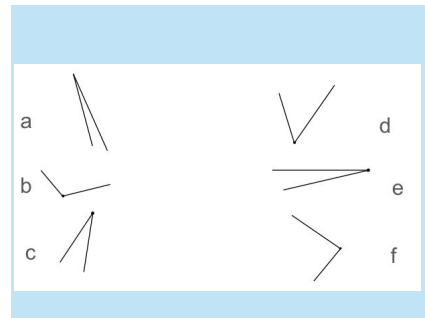


**Question**

In the image, there are black, red, orange, green, blue, and purple regions, with some regions sharing sides.

Choose all the regions that share at least one side with the red region and list their colors.

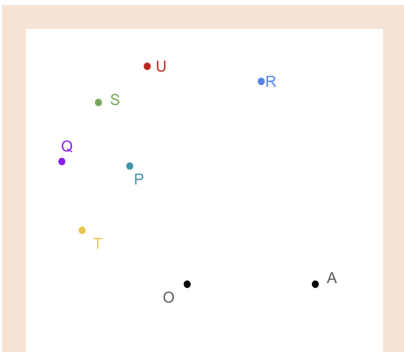
<b>Ground Truth</b> Green	<b>Gemini</b> Green, Blue, Orange
<b>GPT</b> Green, Orange	<b>LLaVA</b> Black, Green, Blue, Purple



**Question**

Among a, b, c, d, e, and f in the image, which pair of lines forms the only obtuse angle? Consider only the internal angle.

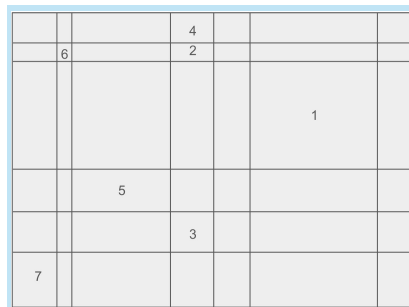
<b>Ground Truth</b> b	<b>Gemini</b> d
<b>GPT</b> e	<b>LLaVA</b> a and b



**Question**

Choose the only acute angle in the image from the following: AOP, AOQ, AOR, AOS, AOT, AOU.

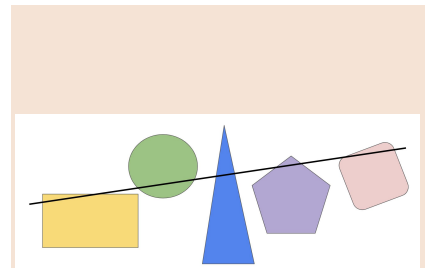
<b>Ground Truth</b> AOR	<b>Gemini</b> AOR
<b>GPT</b> AOP	<b>LLaVA</b> AOQ



**Question**

In the image, there are rectangles formed by a pair of horizontal lines and a pair of vertical lines, labeled by numbers from 1 to 7. Guess the smallest rectangle.

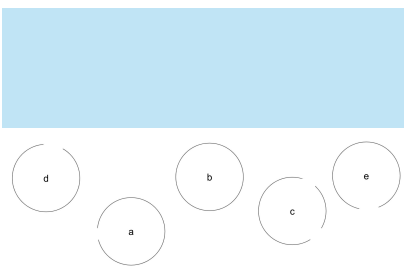
<b>Ground Truth</b> 6	<b>Gemini</b> 2
<b>GPT</b> 4	<b>LLaVA</b> 1



**Question**

In the image, there are five different shapes and a thick black line on them. Which color's shape possesses the largest area above the black line? Choose from: yellow, green, blue, purple, pink.

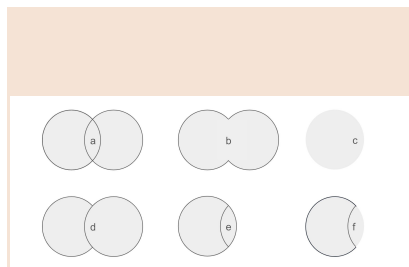
<b>Ground Truth</b> Green	<b>Gemini</b> Yellow
<b>GPT</b> Blue	<b>LLaVA</b> Yellow



**Question**

Let's say there are farms called a, b, c, d, and e, as shown in the image. Black curves represent their fences. From which farm are the sheep unable to escape?

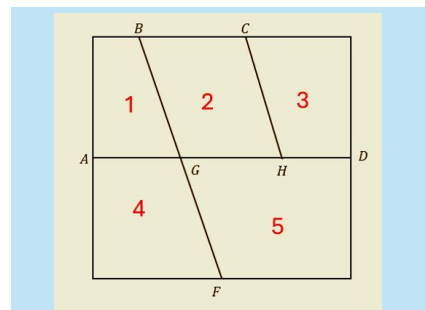
<b>Ground Truth</b> b	<b>Gemini</b> b
<b>GPT</b> b	<b>LLaVA</b> a



**Question**

There are six gray shapes labeled with alphabets in the image. Which shape has only its boundary marked as black? Answer from a, b, c, d, e, f."

<b>Ground Truth</b> b	<b>Gemini</b> f
<b>GPT</b> c	<b>LLaVA</b> d



**Question**

In the image, a square is divided into cells 1, 2, 3, 4, and 5. What is the name of the boundary between cells 4 and 5? Choose from BG, CH, FG, AG, and DG, and write your answer in alphabetical order. For example, the boundary between cells 1 and 2 is 'BG'.

<b>Ground Truth</b> FG	<b>Gemini</b> FH
<b>GPT</b> GH	<b>LLaVA</b> DH

Figure 8: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 3/11.

[Option 1] [Option 2] [Option 3]

**Question**  
Among the three options in the given picture, choose all the options where points A and B are connected by a black line.

<b>Ground Truth</b> 1	<b>Gemini</b> Option 1
<b>GPT</b> Option 1, 3	<b>LLaVA</b> Option 1, 3

**Question**  
In the image, there are shapes of different colors. Identify the color of the only convex shape. Choose from: red, purple, orange, green, or blue.

<b>Ground Truth</b> Orange	<b>Gemini</b> Orange
<b>GPT</b> Orange	<b>LLaVA</b> Blue

**Question**  
In the image, there is a 9 by 9 grid with some blocks colored. Among blue, purple, red, green, and yellow, whose area's shape is convex?

<b>Ground Truth</b> Purple	<b>Gemini</b> Purple
<b>GPT</b> Purple	<b>LLaVA</b> Blue

**Question**  
The picture describes six dots labeled as A, B, C, D, E and F on the x-y coordinate. Which dot has the same x-coordinate as dot A? <1> dot B <2> dot F <3> dot E <4> none of the above

<b>Ground Truth</b> 2	<b>Gemini</b> 4
<b>GPT</b> 1(2/3), 2(1/3)	<b>LLaVA</b> dot B

**Question**  
What are the x and y coordinates of point C? The answer should be in the form of C(x, y). The x and y coordinates of point C are integers.

<b>Ground Truth</b> C(-6, -2)	<b>Gemini</b> C(-6, -2)
<b>GPT</b> C(-6, -1)	<b>LLaVA</b> C(-7, -6)

**Question**  
In the image, there are two circles and a curve. Which color curve is most likely to be part of a circle? Choose from: red, orange, blue, purple, green, yellow, black.

<b>Ground Truth</b> Blue	<b>Gemini</b> Blue
<b>GPT</b> Blue	<b>LLaVA</b> Blue

**Question**  
The image shows a graph of a function with labeled points: A, B, C, and D. Choose the point where the function is most sharply bent. State the label of the point.

<b>Ground Truth</b> A	<b>Gemini</b> A
<b>GPT</b> D	<b>LLaVA</b> C

**Question**  
In the image, inside a circle, there are arrows labeled as P, Q, X, V, W, and Z. Only one of them is pointing inward, while the others are pointing outward. Guess the arrow pointing inward.

<b>Ground Truth</b> Q	<b>Gemini</b> Z
<b>GPT</b> W	<b>LLaVA</b> W

**Question**  
In the image, there are four paths. After following the arrows to the end of each path, one of the routes points in a different direction from the others. Identify the label of the route that points in the different direction. Choose from 'A', 'B', 'C', or 'D'.

<b>Ground Truth</b> B	<b>Gemini</b> D
<b>GPT</b> A	<b>LLaVA</b> B

Figure 9: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 4/11.

**Question**  
In the image, there are a cloud, a sun, a moon, a lightning, and a three-headed arrow. Identify the shape that is NOT pointed at by the arrow. Choose from cloud, sun, moon, and lightning.

<b>Ground Truth</b> Cloud	<b>Gemini</b> Cloud
<b>GPT</b> Moon	<b>LLaVA</b> Lightning

**Question**  
In the image, there are dots in various colors and an arrow. Guess the color of the dot the arrow is pointing at. Choose from: red, orange, blue, green, purple, yellow, brown, gray, black.

<b>Ground Truth</b> Yellow	<b>Gemini</b> No dot
<b>GPT</b> Purple	<b>LLaVA</b> Not Possible to Determine

**Question**  
In the picture, arrows A, B, C, and D are drawn. Among arrows A, B, C, and D, which one is pointing in a different direction from the other three arrows?

<b>Ground Truth</b> B	<b>Gemini</b> C
<b>GPT</b> C(2/3), D(1/3)	<b>LLaVA</b> D

**Question**  
In the image, there is a pentagon ABCDE and 5 points of different colors on line AD. Which colored point is most likely the intersection of line AD and line EC? Choose from: red, blue, green, yellow, or purple.

<b>Ground Truth</b> Yellow	<b>Gemini</b> Purple
<b>GPT</b> Green	<b>LLaVA</b> Red

**Question**  
How many times does the red line segment intersect with the black line segments in the image?

<b>Ground Truth</b> 1	<b>Gemini</b> 9
<b>GPT</b> 9	<b>LLaVA</b> 3

**Question**  
In the image, there are five vertices: 1, 2, 3, 4, and 5, connected by some edges. Among the following choices, which edge exists in the image? Choose from the following options: a) edge(1,2) b) edge(2,3) c) edge(3,4) d) edge(4,5) e) edge(5,1)

<b>Ground Truth</b> a	<b>Gemini</b> c, a, d
<b>GPT</b> b	<b>LLaVA</b> a, b, c

**Question**  
There are five points—A, B, C, D, and E—in the given picture. Some of these points are connected by line segments, while others are not. Choose all the pairs from the options where a line segment exists between the two points, and answer with the appropriate options.  
[Options] (1) A and B (2) A and C (3) A and D (4) A and E (5) B and C (6) B and D (7) B and E (8) C and D (9) C and E (10) D and E.

<b>Ground Truth</b> 1, 5, 7, 8	<b>Gemini</b> 1, 5, 6, 8, 9, 10
<b>GPT</b> 1, 4, 5, 6, 7, 8	<b>LLaVA</b> 1, 2, 3, 4

**Question**  
Read this rotated image.

<b>Ground Truth</b> Everland	<b>Gemini</b> Everyday
<b>GPT</b> EVERYONE	<b>LLaVA</b> PURPOSE

**Question**  
Read the six-letter word displayed in the picture.

<b>Ground Truth</b> Helium	<b>Gemini</b> STREAM
<b>GPT</b> HELLO!	<b>LLaVA</b> Not Recognizable

Figure 10: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 5/11.

**NEW YORK**

**Question**

What are the two characters marked with a blue circle in the image? Answer with the two characters.

<b>Ground Truth</b> EW	◆ <b>Gemini</b> E W
🌀 <b>GPT</b> E, W	<b>LLaVA</b> NY



**Question**

Which of the alphabets is not present in the image: "A", "U", "E", "F", "W", "N", "Y"?

<b>Ground Truth</b> N	◆ <b>Gemini</b> N
🌀 <b>GPT</b> N	<b>LLaVA</b> U



**Question**

Read the text in the given image.

<b>Ground Truth</b> Sejong	◆ <b>Gemini</b> salojas
🌀 <b>GPT</b> Buolas	<b>LLaVA</b> buoias



**Question**

There are shapes of different colors labeled with numbers. Counting from left to right, what is the color of the third circle? Choose from: pink, blue, yellow, green.

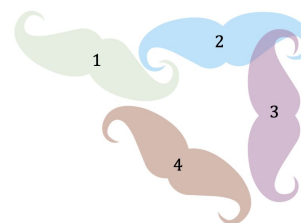
<b>Ground Truth</b> Pink	◆ <b>Gemini</b> Green
🌀 <b>GPT</b> Pink	<b>LLaVA</b> Green

A  
t  
P  
s  
q  
r  
M  
G  
l  
u

**Question**

There are several letters in the image. What is the letter that is fourth from the above? You should also consider its case. For example, if 'z' is fourth from the above, your answer should be "z".

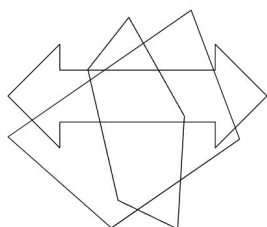
<b>Ground Truth</b> s	◆ <b>Gemini</b> s
🌀 <b>GPT</b> s	<b>LLaVA</b> q



**Question**

There are four shapes labeled 1, 2, 3, and 4 in the given picture. Each shape is semi-transparent and has a single color. Choose all the pairs where overlapping occurs, and answer with the label of the option: (a) 1 and 2 (b) 1 and 3 (c) 1 and 4 (d) 2 and 3 (e) 2 and 4 (f) 3 and 4.

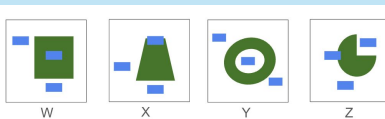
<b>Ground Truth</b> d	◆ <b>Gemini</b> d
🌀 <b>GPT</b> a, d, e, f	<b>LLaVA</b> d



**Question**

In the image, there are three overlapping shapes, two of which are a trapezoid and a two-headed arrow. Identify the last shape from the following options: a triangle, a rectangle, a pentagon, a hexagon, or a circle.

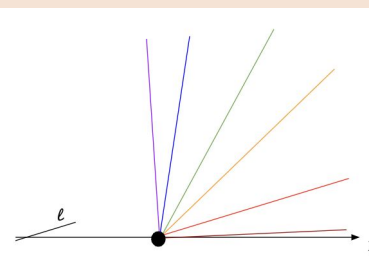
<b>Ground Truth</b> Pentagon	◆ <b>Gemini</b> Hexagon
🌀 <b>GPT</b> Pentagon	<b>LLaVA</b> Rectangle



**Question**

In the image, W, X, Y, and Z each represent a green shape and a group of blue shapes drawn together. Which image shows the only case where the green and blue shapes do not overlap?

<b>Ground Truth</b> Y	◆ <b>Gemini</b> X
🌀 <b>GPT</b> X	<b>LLaVA</b> Y



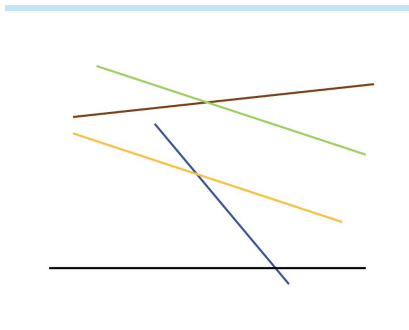
**Question**

Which color's line is parallel to line l? Choose from: blue, orange, green, purple, red, brown.

<b>Ground Truth</b> Red	◆ <b>Gemini</b> Brown
🌀 <b>GPT</b> Brown	<b>LLaVA</b> Purple

Figure 11: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 6/11.

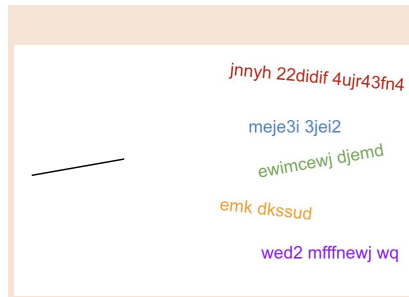




**Question**

In the image, there are several line segments with distinct colors. Given that there is a unique line segment that is parallel to the yellow line, what is the color of the line? Choose from brown, green, blue, and black.

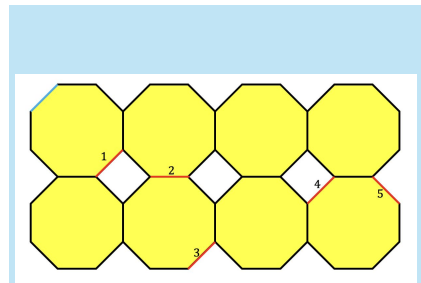
<b>Ground Truth</b> Green	<b>Gemini</b> Brown
<b>GPT</b> Green	<b>LLaVA</b> Green



**Question**

Which color's text is written parallel to the black line in the image? Choose from: red, blue, green, orange, or purple.

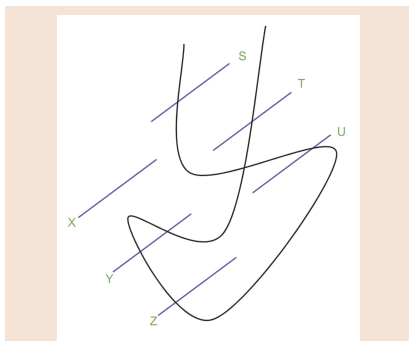
<b>Ground Truth</b> Green	<b>Gemini</b> None
<b>GPT</b> Green	<b>LLaVA</b> Blue



**Question**

Among the six red lines labeled 1, 2, 3, 4, and 5 respectively, choose all the red lines that are parallel to the blue line in the given image.

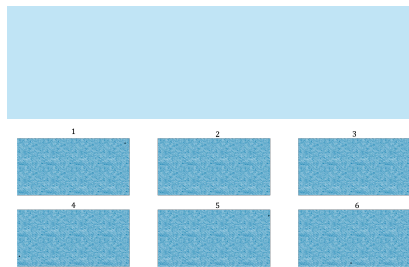
<b>Ground Truth</b> 1, 3, 4	<b>Gemini</b> 2, 5
<b>GPT</b> 1, 3, 4, 5	<b>LLaVA</b> 1, 5



**Question**

In the image, there are lines labeled S, T, U, X, Y, and Z, and a black curve. Which line intersects the curve twice?

<b>Ground Truth</b> Y	<b>Gemini</b> Z
<b>GPT</b> T	<b>LLaVA</b> U



**Question**

Among the six areas in the given picture, find the four areas that contain one or more black points and answer with their labels(numbers).

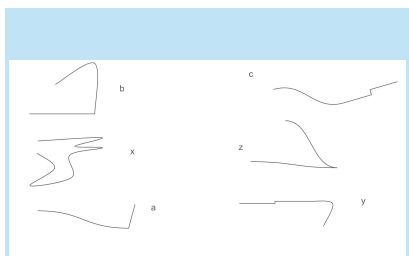
<b>Ground Truth</b> 1, 4, 5, 6	<b>Gemini</b> 1, 4, 5, 6
<b>GPT</b> 1, 2, 5, 6	<b>LLaVA</b> 1, 3, 4, 6



**Question**

In the image, there are curves labeled as: a, b, c, x, y, and z. Guess the only curve with a sharp point.

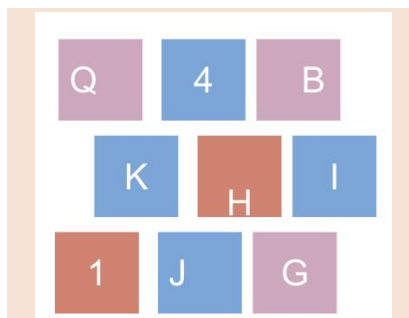
<b>Ground Truth</b> z	<b>Gemini</b> y
<b>GPT</b> y	<b>LLaVA</b> z



**Question**

In the image, there are curves labeled as: a, b, c, x, y, and z. Guess the only curve that is smooth everywhere.

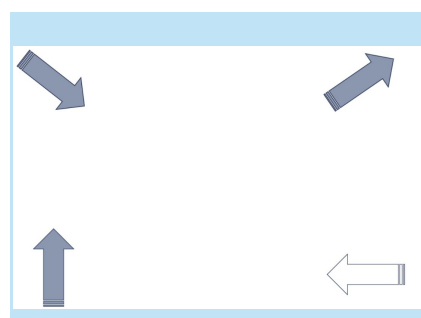
<b>Ground Truth</b> x	<b>Gemini</b> z
<b>GPT</b> c	<b>LLaVA</b> a



**Question**

In the image, there are multiple squares with a number or a letter labeled inside. Name the only label that is written at the bottom of its corresponding square.

<b>Ground Truth</b> 6	<b>Gemini</b> H
<b>GPT</b> 1	<b>LLaVA</b> G

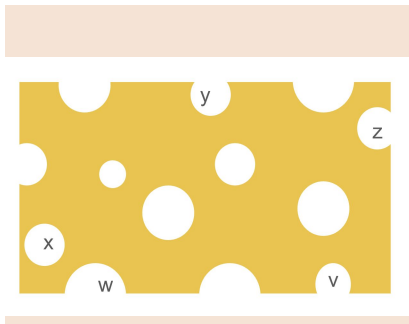


**Question**

In the image, there are four arrows located in each corner. Where is the one with a different color? Choose from [upper right, upper left, lower right, lower left].

<b>Ground Truth</b> Lower Right	<b>Gemini</b> Lower Right
<b>GPT</b> Lower Left	<b>LLaVA</b> Upper Right

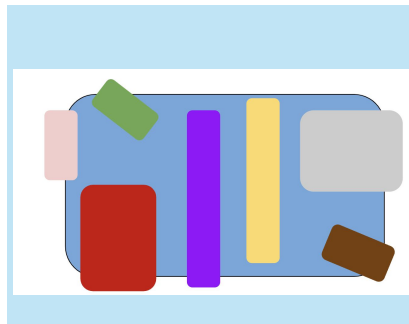
Figure 12: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 7/11.



**Question**

In the image, there is a slice of yellow cheese with holes. Which hole forms a perfect circle in the cheese? Choose from: v, w, x, y, z.

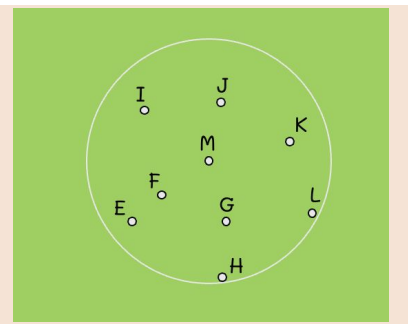
<b>Ground Truth</b> x	<b>Gemini</b> w
<b>GPT</b> y	<b>LLaVA</b> v



**Question**

Which color's area is totally contained inside the blue area? Choose just one from: red, green, pink, gray, purple, yellow, brown.

<b>Ground Truth</b> Yellow	<b>Gemini</b> Gray
<b>GPT</b> Purple	<b>LLaVA</b> Green



**Question**

In this picture, there is a circle and 9 points. The center of the circle refers to the point from which all points on the circle are equidistant. Which point among E, F, G, H, I, J, K, L, M is most likely to be the center of the circle in the picture?

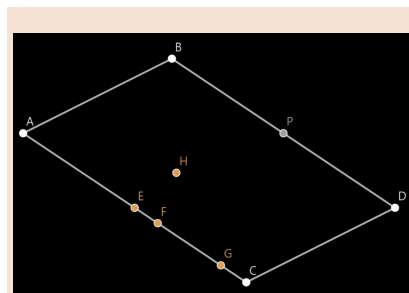
<b>Ground Truth</b> M	<b>Gemini</b> M
<b>GPT</b> M	<b>LLaVA</b> I



**Question**

In the image, there are squares of the same size, colored red, yellow, black, green, blue, orange, pink, and purple. Identify the color of the only square at a different height.

<b>Ground Truth</b> Orange	<b>Gemini</b> Orange
<b>GPT</b> Orange	<b>LLaVA</b> Orange



**Question**

In the image, point P is the midpoint of the line segment BD. Similarly, choose the point that is most likely to be the midpoint of the line segment AC. The answer should be one of E, F, G, or H.

<b>Ground Truth</b> E	<b>Gemini</b> F
<b>GPT</b> F	<b>LLaVA</b> E



**Question**

In the image, there are squares of the same size, colored red, yellow, black, green, blue, orange, pink, and purple. Identify the color of the square closest to the top.

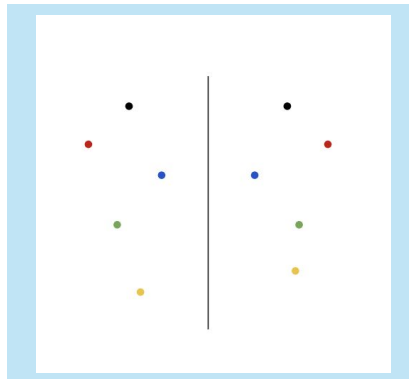
<b>Ground Truth</b> Blue	<b>Gemini</b> Blue
<b>GPT</b> Black	<b>LLaVA</b> Blue



**Question**

There are points labeled 1, 2, 3, 4, and 5 in the picture. Choose the phrase that describes the image correctly. Point 2 is on the (left/right/upper/lower/upper left/upper right/lower left/lower right) side of point 4.

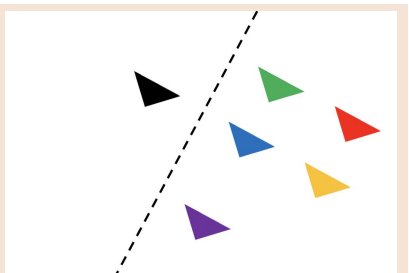
<b>Ground Truth</b> Upper	<b>Gemini</b> Upper Left
<b>GPT</b> Upper Left	<b>LLaVA</b> Upper Left



**Question**

Among black, blue, red, green, and yellow, choose the color of the pair of points that is not line-symmetric about the black line.

<b>Ground Truth</b> Yellow	<b>Gemini</b> Yellow
<b>GPT</b> Yellow	<b>LLaVA</b> Red and Yellow



**Question**

In the image, there are a black triangle on the left of the line and five triangles on the right of the line. Choose the triangle that is symmetric to the black triangle with respect to the line. Choose your answer from "red", "blue", "green", "yellow", "purple". For example, if black and red triangle are symmetric with respect to the line, your answer should be "red".

<b>Ground Truth</b> Blue	<b>Gemini</b> Green
<b>GPT</b> Purple	<b>LLaVA</b> Blue

Figure 13: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 8/11.

**Question**  
In the image, there are four arrows and three line segments of different colors. The black arrow is the original arrow, and the other arrows are the results of reflections with respect to the line segment that matches the color of the resulting arrow. Among the three colors, one is incorrectly reflected. Which color is not correctly reflected? Choose from blue, gray, and purple.

<b>Ground Truth</b> Gray	<b>Gemini</b> Purple
<b>GPT</b> Blue	<b>LLaVA</b> Purple

**Question**  
In the image, there is a black line and four pairs of curves (each curve has the same color as its partner), each representing a reflection of a figure. Which pair failed to correctly represent a reflection with respect to the black line? Answer with the color of the pair: blue, green, black, or red.

<b>Ground Truth</b> Red	<b>Gemini</b> Red
<b>GPT</b> Blue	<b>LLaVA</b> Blue

**Question**  
In the image, there is a 9 by 9 grid with some blocks colored. Among blue, red, orange, green, and yellow, which color blocks are line-symmetric about the black line?

<b>Ground Truth</b> Blue	<b>Gemini</b> Blue, Red, Orange
<b>GPT</b> Green, Blue, Red	<b>LLaVA</b> Blue and Red

**Question**  
Choose the word in parentheses that correctly describes the image: In the image, line segment (AD/DC/CB) has the same length as line segment AB.

<b>Ground Truth</b> DC	<b>Gemini</b> DC
<b>GPT</b> AD	<b>LLaVA</b> DC

**Question**  
In the image, there are curves labeled a, b, c, x, y, and z. Which curve is the shortest?

<b>Ground Truth</b> y	<b>Gemini</b> y
<b>GPT</b> y	<b>LLaVA</b> c

**Question**  
In the image, there is a black line and 5 different shapes on it. Which shape is the tallest? Choose from: heart, square, triangle, oval, circle.

<b>Ground Truth</b> Circle	<b>Gemini</b> Circle
<b>GPT</b> Heart	<b>LLaVA</b> Circle

**Question**  
Guess which color's line is the only line orthogonal to the bold black line in the image. Choose from: blue, green, red, orange, purple, brown.

<b>Ground Truth</b> Red	<b>Gemini</b> Orange
<b>GPT</b> Green	<b>LLaVA</b> Blue

**Question**  
In the image, there are six points A, B, C, D, E, and O and some line segments. Several pairs of line segments meet at one of the points and form a right angle. Find three points where the right angles are located. List the letters denoting the correct points in the format of 'X, Y, Z', in alphabetical order.

<b>Ground Truth</b> C, D, E	<b>Gemini</b> C, D, E
<b>GPT</b> B,E,O	<b>LLaVA</b> A, B, C, A, E, O, B, D, D

**Question**  
In the image, there are four choices from A to D, each paired with two line segments. Which choice represents a right angle? Answer with a single letter.

<b>Ground Truth</b> A	<b>Gemini</b> B
<b>GPT</b> C	<b>LLaVA</b> D

Figure 14: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 9/11.

**Question**  
In the image, there exists a line segment perpendicular to the line EA. Choose the line segment from the following choices: 1) AB 2) BD 3) BC 4) FB Answer with the number of the choice.

<b>Ground Truth</b> 3	<b>Gemini</b> 2
<b>GPT</b> 2	<b>LLaVA</b> 2

**Question**  
In the image, there are lines with different colors. Which color line is NOT orthogonal to the black line? Choose from: red, blue, orange, green, purple, pink, brown, yellow.

<b>Ground Truth</b> Orange	<b>Gemini</b> Red
<b>GPT</b> Red	<b>LLaVA</b> Purple

**Question**  
In the image, which green shape is least likely to be a rotation of X about the center of the circle? Choose from: a, b, c, d, or e.

<b>Ground Truth</b> c	<b>Gemini</b> d
<b>GPT</b> d	<b>LLaVA</b> b

**Question**  
Which of the shapes labeled 1, 2, 3, or 4 cannot be made by rotating shape A? Provide the corresponding label.

<b>Ground Truth</b> 3	<b>Gemini</b> 4
<b>GPT</b> 2	<b>LLaVA</b> 1

**Question**  
If you rotate the black shape around the center of the circle in the image, which color's shape will fit perfectly with it, as shown in the sample above?

<b>Ground Truth</b> Orange	<b>Gemini</b> Orange
<b>GPT</b> Red	<b>LLaVA</b> Purple

**Question**  
In the image, there is a yellow shape and five other shapes. Choose the shape that is symmetric with the yellow shape with respect to point G. Choose your answer from A, B, C, D, or E.

<b>Ground Truth</b> B	<b>Gemini</b> E
<b>GPT</b> C	<b>LLaVA</b> C

**Question**  
Among shapes A, B, C, D, E, and F, which one does NOT have any 3-fold rotational symmetry?

<b>Ground Truth</b> A	<b>Gemini</b> F
<b>GPT</b> C	<b>LLaVA</b> D

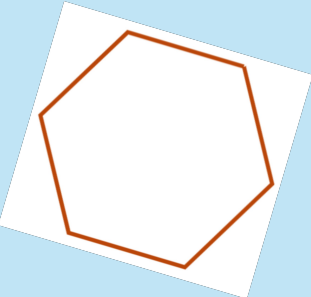
**Question**  
Choose all the options that correctly describe the given purple shape. (1) It is a quadrilateral. (2) It is a square. (3) It is a rectangle. (4) It is a trapezoid. (5) It is a rhombus. (6) It is a parallelogram.

<b>Ground Truth</b> 1	<b>Gemini</b> 1
<b>GPT</b> 1	<b>LLaVA</b> 6

**Question**  
Choose all the shapes you can see in this picture and answer with their labels: (1) triangle (2) square (3) pentagon (4) hexagon (5) arrow (6) star (7) heart (8) circle.

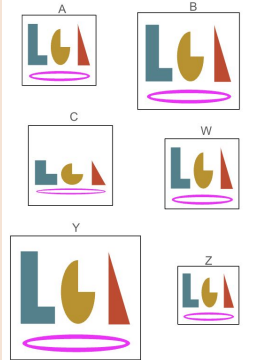
<b>Ground Truth</b> 1, 2, 3, 4, 5, 7	<b>Gemini</b> 1, 4, 5, 7
<b>GPT</b> 1, 2, 3, 4, 5, 7	<b>LLaVA</b> 1, 2, 3, 4, 5, 6, 7, 8

Figure 15: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 10/11.



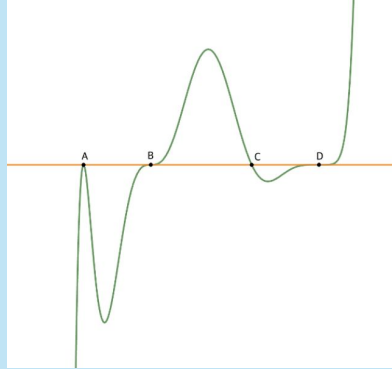
**Question**  
In the given picture, there are line segments which become the sides of some shape. What is the type of the shape in the picture? Choose one the following and answer in a single word: triangle, rectangle, parallelogram, pentagon, hexagon, circle, star, or heart.

<b>Ground Truth</b> Hexagon	<b>Gemini</b> Hexagon
<b>GPT</b> Pentagon	<b>LLaVA</b> Parallelogram



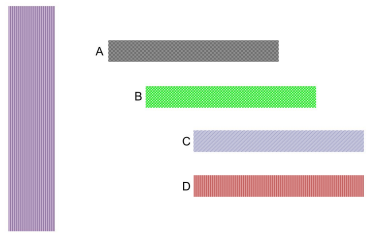
**Question**  
In the image, there are figures of different sizes labeled A, B, C, X, Y, and Z. Which figure is the only one with a different aspect ratio?

<b>Ground Truth</b> C	<b>Gemini</b> Y
<b>GPT</b> W	<b>LLaVA</b> Y



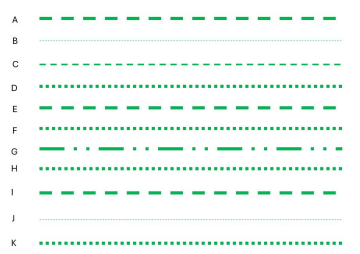
**Question**  
Among the points A, B, C, and D in the given picture, identify the only point where the green curve and the orange line meet but are not tangent.

<b>Ground Truth</b> C	<b>Gemini</b> D
<b>GPT</b> B	<b>LLaVA</b> C



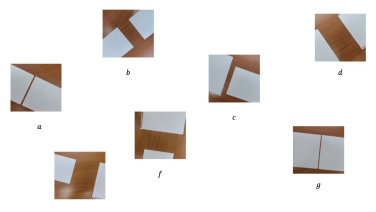
**Question**  
Among A, B, C, and D in the given picture, which rectangle has the same texture as the purple rectangle? The answer should be a single letter.

<b>Ground Truth</b> D	<b>Gemini</b> D
<b>GPT</b> D	<b>LLaVA</b> A




**Question**  
Lines labeled A, B, C, D, E, F, G, H, I, J, and K are drawn in various styles. List all the letters that represent lines with the same style as line K.

<b>Ground Truth</b> D, F, H	<b>Gemini</b> D, F, H
<b>GPT</b> D, F, H	<b>LLaVA</b> A, B, C, D, E, F, G, H, I, J



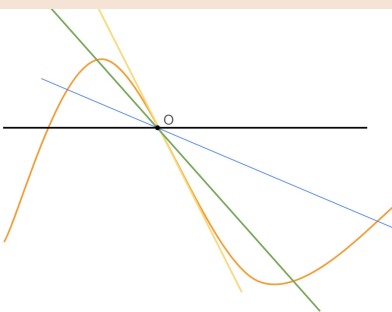
**Question**  
There are two white papers in each option. Among a, b, c, d, e, f, and g, which option has the narrowest gap between the two white papers?

<b>Ground Truth</b> g	<b>Gemini</b> g
<b>GPT</b> g	<b>LLaVA</b> e



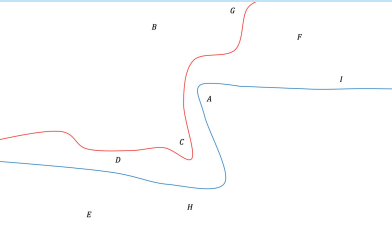
**Question**  
Which color's road is the widest? Choose from: yellow, blue, green, pink, gray.

<b>Ground Truth</b> Yellow	<b>Gemini</b> Gray
<b>GPT</b> Gray	<b>LLaVA</b> Yellow



**Question**  
In the image, which line is tangent to the orange curve at point O among the green, blue, black, and yellow line segments? State the color of the line segment in lowercase.

<b>Ground Truth</b> Yellow	<b>Gemini</b> Black
<b>GPT</b> Black	<b>LLaVA</b> Green



**Question**  
In the image, there are 9 alphabet letters, labeled A to I, along with one red curve and one blue curve. List all the alphabet letters that are located between the red curve and the blue curve.

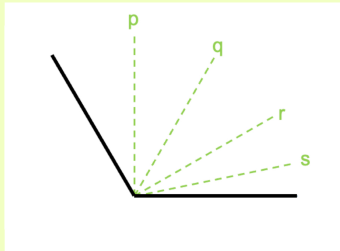
<b>Ground Truth</b> D, F, I	<b>Gemini</b> A, C
<b>GPT</b> A, C, D	<b>LLaVA</b> B, C, D, E, F, G, H

Figure 16: Sample problems from AVSBench and responses from Gemini-1.5-pro, GPT-4o, and LLaVA-Next-13B, 11/11.

## D Descriptions of the atomic visual skills

In this section, we provide detailed definitions of the 36 atomic visual skills, together with a corresponding problem sample from AVSBench.

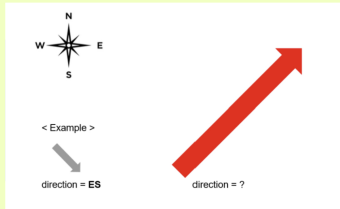
1. **Angle** is a skill to understand how an angle is visually represented. Angle is the primary factor in how a polygon looks, how two or more objects are related, and many other situations.



Q) In the given image, there is an angle represented by two black lines. Also, there are four green dotted lines that divide the angle. Which line divides the angle equally?

A) q

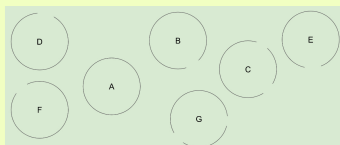
2. **Direction** is an ability to recognize linear direction in an image. It is a fundamental skill in human vision, supporting representation of linearity and multi-dimensional relations.



Q) In the picture, directions are described by north (N), south (S), east (E), west (W), or combinations of these. For example, the gray arrow in the picture indicates the ES (east-south) direction. Now, among N, E, S, W, NE, ES, SW, and SN, which best describes the direction of the red arrow in the picture?

A) NE

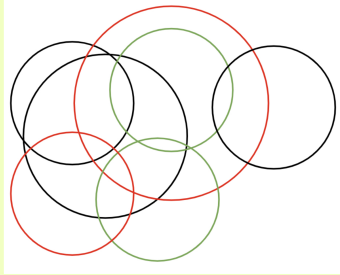
3. **Boundary** is a skill to understand the ends of objects or areas, and to detect visual representation of edges. The skill is used in distinguishing between distinct objects, or detecting boundaries between spaces.



Q) Let's say there are farms called A, B, C, D, E, F, and G, as shown in the image. Black curves represent their fences. From which farm are the sheep unable to escape?

A) A

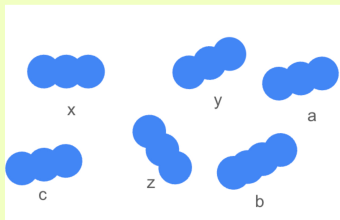
4. Cardinal is a field about counting distinct objects or specified concepts. Mastery of cardinals implies measuring quantities or dealing with multiple objects. Especially, it should take into account everything that satisfies given conditions, giving a difference from the skill of understanding *Ordinals*.



Q) How many circles are in the image?

A) 7

5. Congruence is a skill of detecting objects with the exact same scale and shape, and understanding their correspondence. Congruence is a primary component of visualizing various symmetries including translation, rotation or flipping. Congruence is distinguished from other equivalence because it requires the objects to be equal at all levels of measurement.



Q) In the image, there are different shapes labeled as a, b, c, x, y, and z. Guess which shape is different from the others.

A) b

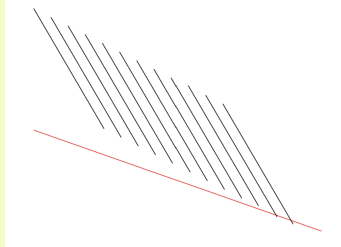
6. Convexity is a skill of understanding convexity of given shapes. The skill is also closely related to detecting bumps or indentations and understanding convex and concave functions.



Q) In the image, there are shapes of different colors. Identify the color of the only convex shape. Choose from: red, purple, orange, green, or blue.

A) Orange

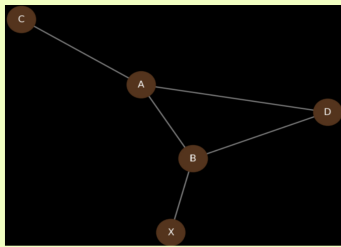
7. Intersection is a mastery of detecting intersections of lines and curves. The skill is necessary for interpreting relationships among 1-dimensional objects, and also among higher dimensional objects from 1-dimensional representations of their boundaries.



Q) How many times does the red line segment intersect with the black line segments in the image?

A) 2

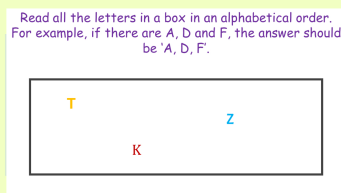
8. Line is a skill to detect line segments and understand their roles in the image. This skill is a fundamental unit in understanding various objects as polygons, graphs and diagrams.



Q) There are five vertices A, B, C, D, and X connected with some edges in the image. Among the following choices, which edge does not exist in the image?

- 1) edge(A,D)
- 2) edge(B,X)
- 3) edge(C,D)
- 4) edge(A,B)

9. OCR is a skill to detect and read characters from visual inputs.

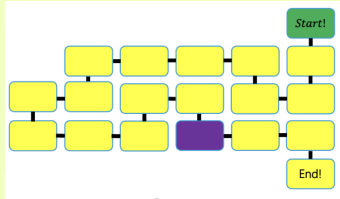


Q) Follow the instructions in the image.

A) K, T, Z



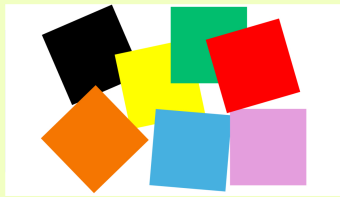
10. Ordinal is a skill to count certain objects or concepts in a given order. Mastery of this skill requires not just counting but also focusing on specific portions and order of targets, giving a difference from Cardinal Understanding.



Q) The given diagram shows the path from the "Start!" box to the "End!" box. Consider the green "Start!" box as the first box. What is the position of the purple box in terms of its sequence number?

**A) 16(th)**

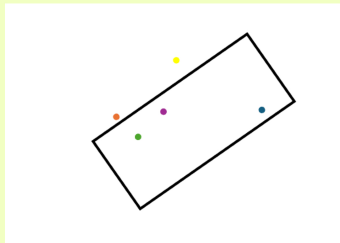
11. Overlap skill is about correctly recognizing two or more objects sharing a common area. The skill is crucial in understanding overlapping shapes or complex shapes such as diagrams.



Q) There are several squares of identical size colored red, orange, yellow, pink, green, blue, and black. Which squares are overlapping with the blue square? Answer in a set of colors.

**A) Yellow, pink**

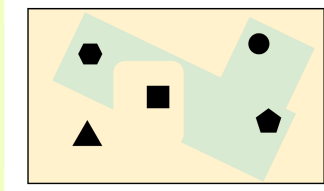
12. Interior is a skill of distinguishing between interior and exterior of the target area. This skill is essential in perceiving different areas.



Q) In the image, there is a black box and five points with different colors (orange, green, yellow, purple, and blue). Choose best option that includes all the colors of the points inside the box.

1. orange, green, purple
- 2. purple, blue, green**
3. blue, purple, yellow

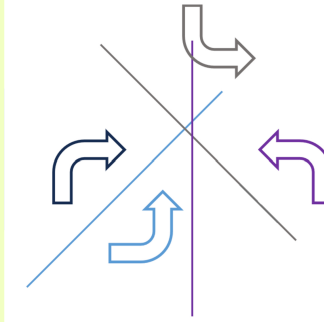
13. **Relative Position** is an ability to identify positional relationships between objects that cannot be simply described such as inside, outside, or moved in a certain direction. This skill requires comprehension of complex relationships such as “positioned in between,” or “at the same side of.”



Q) In the image, there is a box with two disjoint regions: a yellow region and a green region. In that box, there are also 5 shapes: triangle, square, pentagon, hexagon, and circle. Among those shapes, which shape is in the same region as the triangle?

A) Square

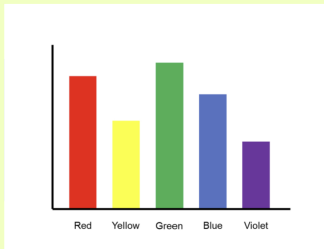
14. **Reflection** is a field of recognizing linear symmetries. It requires detecting the axis of reflection and induced correspondence of objects.



Q) In the image, there are four arrows and three line segments of different colors. The black arrow is the original arrow, and the other arrows are the results of reflections with respect to the line segment that matches the color of the resulting arrow. Among the three colors, one is incorrectly reflected. Which color is not correctly reflected? Choose from blue, gray, and purple.

A) Gray

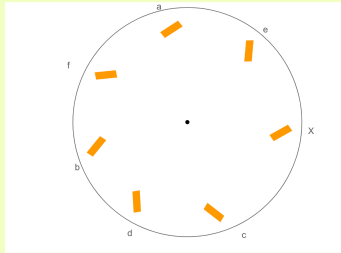
15. **Length** is a skill to handle lengths of different objects. It involves comparing different lengths and measuring distances of objects.



Q) Choose the word in parentheses that correctly describes the image: In the given bar graph, the length of the red bar is (longer/shorter) than the length of the blue bar.

A) longer

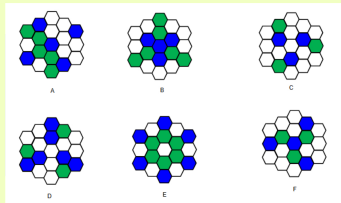
16. Rotation is an ability to identify changes in positions and angles induced by rotation, and detecting the axis of rotation.



Q) In the image, which orange shape is least likely to be a rotation of X about the center of the circle? Choose from: a, b, c, d, e, f.

A) a

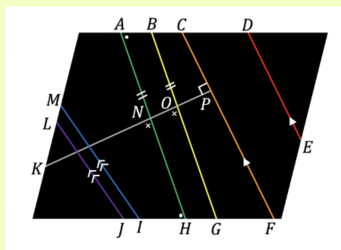
17. Rotational Symmetry is a field of symmetric representations with respect to rotations. The skill involves understanding invariant geometric features under specified rotations.



Q) Among shapes A, B, C, D, E, and F, which one does NOT have any 3-fold rotational symmetry?

A) A

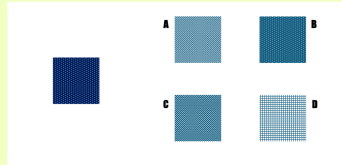
18. Symbol is a skill to detect symbols, understand their roles in the image, and combine them with other visual information to attain the correct interpretation of the image.



Q) Choose all the correct meanings of the symbols in the given picture.

- (1) The angle JHA has the same size as the angle DAN.
- (2) The angle KNH has the same size as the angle KOB.
- (3) Two lines MI and CF are parallel.
- (4) Two lines NP and CF are perpendicular.

19. Texture is a skill to understand textures of objects in the image. The skill is essential as texture is another main component of visual representation of objects, and is used to distinguish different objects with same shapes, such as line styles.



Q) Among A, B, C, and D in the given picture, which square has a same texture with the leftmost square?

A) B

20. Width is a skill to understand thickness and width of objects or areas. The skill is essential in measuring area or proportion of images together with length understanding.



Q) Choose the widest ladder in the given picture. Choose one from the orange ladder, yellow ladder, green ladder, purple ladder, or the blue ladder.

A) Purple ladder


21. Adjacency is a skill to recognize when two or more objects are next to each other. The skill is crucial in understanding features induced by close positions such as forming clusters.



Q) In the image, there are black, red, orange, green, blue, and purple regions, with some regions sharing sides. Choose all the regions that share at least one side with the red region, and list their colors.

A) Green

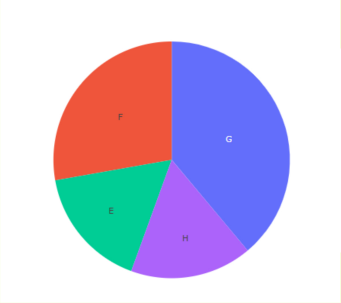
22. Absolute Position is a skill to correctly understand where the objects are represented as a part of the visual input, independently of other objects. This involves recognizing objects posited at corners of an image, or comparing heights of objects represented in the image.



Q) In the image, there are several different shapes. Which shape is in the upper right corner?

- (1) star
- (2) triangle
- (3) pentagon
- (4) rectangle**
- (5) circle

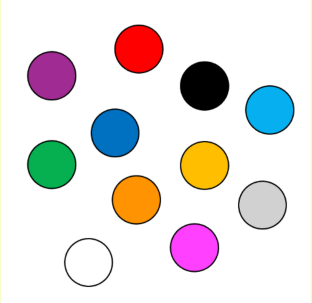
23. Area is a skill to handle 2-dimensional volumes, including comparing areas.



Q) In the image, there is a pie graph with four categories. Which category has the largest ratio? Choose from E, F, G, and H.

**A) G**

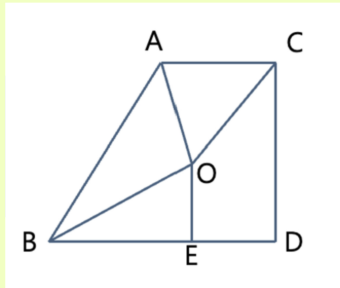
24. Cardinal Direction is a skill to understand primary directions including up, down, left, right, or diagonals. This involves recognizing North, South, West, and East directions.



Q) In the image, there are many circles with different colors (red, blue, green, purple, black, yellow, orange, white, gray, pink, and skyblue). Choose the color of the circle that is to the left above the blue circle.

**A) Purple**

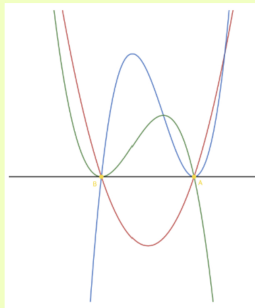
25. Orthogonality is a skill to identify orthogonal relations of objects in the image, including a right angle formed by two lines. Understanding orthogonality is fundamental in geometry, design, and engineering.



Q) In the image, there are six points A, B, C, D, E, and O and some line segments. Several pairs of line segments meet at one of the points and form a right angle. Find three points where the right angles are located.

A) C, D, E

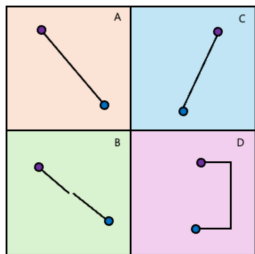
26. Tangency is a skill to detect tangent objects. This skill focuses on geometric representation of tangent curves or boundaries, and is different from understanding adjacency that rather focuses on positional information.



Q) In the given picture, there are three curves: red, blue, and green. Which curve is tangent to the black line at point B? State the color of that curve.

A) Green

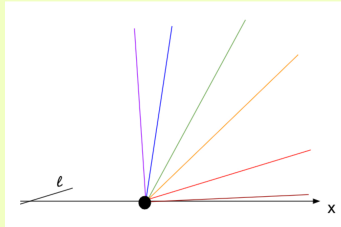
27. Connectedness is a skill to identify connected components and detect links between objects. This is crucial in understanding interactions and distinguishing distinct components.



Q) Among option A, B, C, and D in the picture, choose one option where purple and blue points are not connected by a black line.

A) B

28. Parallel is a skill to recognize parallel lines or curves. This is essential in identifying fundamental objects like squares.



Q) Which color's line is parallel to line  $l$ ?  
Choose from: blue, orange, green, purple, red, brown.

A) Red

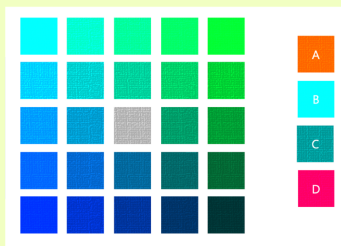
29. Similarity is a skill to understand equivalence of geometric representations independent of scale. It also involves understanding of rescaling or comparing aspect ratios.



Q) There are five stars in the image.  
Choose the shape that is geometrically similar to the black star (rightmost star).  
What is the number indicating that shape? Your answer should be "1", "2", "3", or "4".

A) 3

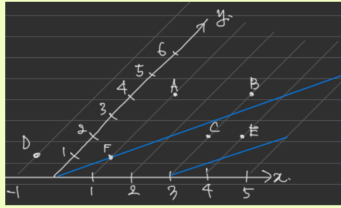
30. Color is an ability to perceive, distinguish different colors, and understand the change in saturation and brightness.



Q) Choose the most appropriate color to fill in the gray box in the image. The answer is one of 'A', 'B', 'C', or 'D'.

A) C

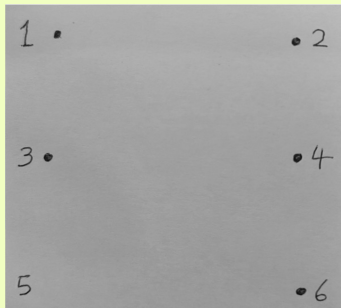
31. Coordinate is a skill to recognize and acquire correct information upon coordinate systems. We provide and acquire information about different systems such as polar coordinates.



Q) The picture describes six dots labeled as A, B, C, D, E and F on the x-y coordinate. Which dot has the same x-y coordinate as dot A?

- <1> dot B
- <2> dot F
- <3> dot E
- <4> none of the above

32. Point is a fundamental capability to detect points and understand their roles in the image. It also involves understanding nodes in different graphs.



Q) In the given picture, there are five points and six numbers, meaning that one of the six numbers does not have a corresponding point. Identify the number that does not have a corresponding point.

A) 5

33. Shape is a skill to understand details of shapes and compare different shapes independently of positions or tilts. It also involves identifying popular shapes such as triangles, rectangles, circles, and stars.

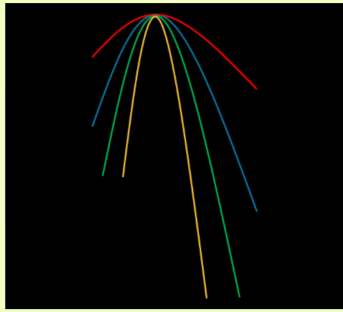


Q) In the image, there are diverse polygons with different colors and textures. Which of the sentences describes the image worst? Choose only one sentence.

- 1) There is a triangle with a hole inside.
- 2) There is a square with a hole inside.
- 3) There is a pentagon with a hole inside.
- 4) There is a hexagon with a hole inside.



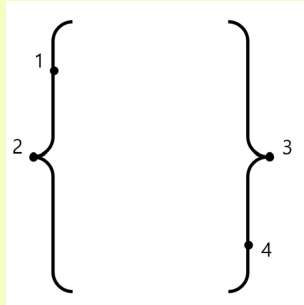
34. Curvature is an ability to measure and compare curvatures of different curves. This involves distinguishing between straight lines and wavy curves, and detecting bends in a shape.



Q) In the image, there are four curves that all meet at a single point. Which curve is the most sharply bent at that point? Choose from red, orange, and blue. Answer with the color of the curve.

A) Orange

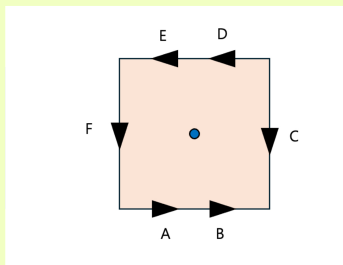
35. Sharpness is a skill to detect pointy parts of a shape. This is essential in understanding the representations of non-smooth objects such as points of a function that are not differentiable.



Q) There is a shape in the image. Among the shape's four parts labeled from 1 to 4, choose all the parts that are pointy.

A) 2, 3

36. Orientation is a skill to correctly distinguish clockwise and counterclockwise tendencies induced by not only rotations but also other movements that result in clockwise and counterclockwise directional change. The name originated from the mathematical definition of orientation in differential geometry.



Q) In the image, there are six crosses with arrows. Among them, one arrow is rotating in the opposite direction compared to the other five. Find the one that rotates in a different orientation from the others. Answer with the letter that denotes it.

A) E

## E Model versions

We evaluated closed-source models ChatGPT [40], Gemini [49] and open-weight models LLaVA-NeXT [29], LLaVA-OneVision [26], Math-LLaVA [47], Table-LLaVA [60], Phi-3.5-Vision [1], InternVL2 [10], DeepSeek-VL [31]. Tables 1 and 2 describe further details about the model sizes and versions. For closed-source models, we used the commercial APIs. All models’ temperatures were set to 0.

Table 1: Versions of closed-source models

Model Name	Version
ChatGPT	gpt-4o-2024-05-13
	gpt-4o-mini-2024-07-18 (For scoring)
Gemini	gemini-1.5-pro-001

Table 2: Versions and model sizes of open-weight models

Version	Model Size(s)
LLaVA-NeXT	7B, 13B
LLaVA-OneVision	7B
Math-LLaVA	13B
Table-LLaVA	7B
Phi-3.5-Vision-Instruct	4B
InternVL2	8B
DeepSeek-VL	7B

## F Further details on evaluation process

We used GPT-4o mini to extract answers from model responses and to judge correctness. Few-shot in-context learning prompts we provided to GPT-4o mini as described in Tables 3 and 4. To verify the reliability of this pipeline, we randomly selected 128 problems from our dataset and compared the scores from GPT-4o mini with human annotations. Reassuringly, GPT-4o mini and the human annotators agreed on the scoring of the 128 problems. We attribute this high level of reliability, in part, to the straightforward and clear design of our questions and answers.

Element	Prompt
System prompt	Imagine you are an intelligent teacher. Thoroughly read the provided instruction to ensure a solid understanding of the information provided.
Task description	Please read the following example. Then extract the answer from the model response and type it at the end of the prompt. If the question requires a full sentence with a correct word filled in, please provide the word only. <i>{examples}</i> Question: <i>{question}</i> Model response: <i>{model response}</i> Extracted Answer:
Examples	<p><b>Question:</b> There is a single rectangle with multiple color layers in the image. What is the color of the boundary of the rectangle? The answer should be one of 'red', 'yellow', 'green', or 'blue'.  <b>Model response:</b> The color of the boundary of the circle is red.  <b>Extracted answer:</b> red</p> <p><b>Question:</b> How many line segments are in the image? Answer should be a number.  <b>Model response:</b> There are 4 dashed line segments in the image.  <b>Extracted answer:</b> 4</p> <p><b>Question:</b> Choose the word in parentheses that correctly describes the image. Rewrite the sentence with the chosen word.            In the image, shape (A/B) has sides curved inward. (Unit: \$)  <b>Model response:</b> In the image, shape B has sides curved inward.  <b>Extracted answer:</b> B</p> <p><b>Question:</b> Choose the phrase in parentheses that correctly describes the image. Rewrite the sentence with the chosen phrase.            In the given image, the green arrow (is longer than/has the same length as/is shorter than) the black arrow.  <b>Model response:</b> In the given image, the green arrow is longer than the black arrow.  <b>Extracted answer:</b> is longer than</p> <p><b>Question:</b> In this image, choose the path which is a single line segment between points A and B from the following options. Provide your answer as a single uppercase letter: (A) the purple path (B) the blue path (C) the green path (D) the red path  <b>Model response:</b> B  <b>Extracted answer:</b> B</p> <p><b>Question:</b> Choose the most appropriate color to fill in the box marked with '?' in the image. The answer is one of 'a', 'b', 'c', or 'd'.  <b>Model response:</b> The correct color to fill in the box marked with '?' is (a) blue. The colors are following a gradient pattern from red, to a more purple hue, and finally to blue. The logical next color in the sequence would be blue, as it extends the progression seen in the previous squares.  <b>Extracted answer:</b> a</p> <p><b>Question:</b> There is a book in the image. What is the color of the book in the image? Choose answer from the number of the option and give your answer in "1", "2", "3", or "4". (1) red (2) yellow (3) blue (4) green  <b>Model response:</b> The color of the guitar in the image is (2) yellow.  <b>Extracted answer:</b> 2</p>

Table 3: System prompt, task description, and examples used to prompt GPT-4o mini for answer extraction.

<b>Element</b>	<b>Prompt</b>
System prompt	Imagine you are an intelligent teacher. Thoroughly read the provided instruction to ensure a solid understanding of the information provided.
Task description	<p>The [Standard Answer] is the correct answer to the question, and the [Model Answer] is the answer generated by a model for that question. Thoroughly read both the [Standard Answer] and the [Model Answer]. Assess the consistency of the information provided in these two responses.</p> <p>Although you do not know the specific question, you can still assess the consistency between the two responses by checking for logical conflicts if both responses are assumed to be correct.</p> <p>If the [Model Answer] is consistent with the [Standard Answer], please answer '1'. Otherwise, answer '0'.</p> <p>When the [Standard Answer] is provided as a list, answer '1' if the [Model Answer] is consistent with at least one item on the list. Otherwise, answer '0'.</p> <p>Below are the examples of the correct consistency judgment.</p> <p><i>{examples}</i></p> <p>Now, below are two answers to a question. What is your judgment?</p> <p>[Standard Answer] <i>{standard answer}</i></p> <p>[Model Answer] <i>{extracted answer}</i></p> <p>Judgment:</p>
Examples	<p>[Standard Answer] a [Model Answer] a <b>Judgment: 1</b></p> <p>[Standard Answer] 1 [Model Answer] 4 <b>Judgment: 0</b></p> <p>[Standard Answer] circle [Model Answer] the circle <b>Judgment: 1</b></p> <p>[Standard Answer] 4 [Model Answer] shape 4 <b>Judgment: 1</b></p> <p>[Standard Answer] line segment B and C [Model Answer] B, C <b>Judgment: 1</b></p> <p>[Standard Answer] ac [Model Answer] ca <b>Judgment: 0</b></p> <p>[Standard Answer] 2 [Model Answer] two <b>Judgment: 1</b></p> <p>[Standard Answer] three [Model Answer] 3 <b>Judgment: 1</b></p> <p>[Standard Answer] ['ac', 'ca'] [Model Answer] ca <b>Judgment: 1</b></p>

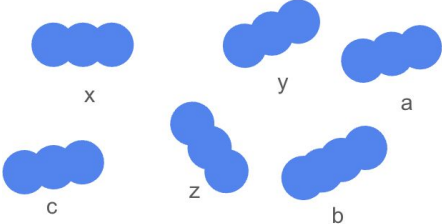
Table 4: System prompt, task description, and examples used to prompt GPT-4o mini for judgment.

## G Results on 36 skills and Further Analysis

In this section, we provide full details of Section 3 and further analysis of Findings 3.

Table 5 presents our full evaluation results on AVSBench. “Random Chance” represents the expected accuracy of randomly choosing multiple-choice questions and scoring 0 from all non-multiple-choice questions. The subcolumn named “overall” indicates the accuracy across all problems of its corresponding skill. The column named “TOTAL” describes the overall accuracy across all problems in AVSBench, involving all skills.

As mentioned in Findings 3, CoT did not provide meaningful performance gains when used with GPT-4o and Gemini-1.5-pro. Gemini-1.5-pro had only a 0.1% gain from applying CoT, and GPT-4o had a 1% gain. However, CoT worsened performances of skills including OCR, Length, and Symbol. As in the case of Figure 17, by inspecting the responses of GPT-4o with and without CoT prompting, we observe that the additional reasoning steps afforded by CoT are not helpful in comprehending visual inputs.



<p style="text-align: center;"><b>Question</b></p> <p>In the image, there are different shapes labeled as a, b, c, x, y, and z. Guess which shape is different from the others.</p>	
<p style="text-align: center;"><b>Ground Truth</b></p> <p style="text-align: center;">b</p>	
<p><b>GPT-4o</b></p> <p>The shape labeled "z" is different from the others. While all the other shapes (a, b, c, x, y) are arranged horizontally, the shape labeled "z" is arranged vertically.</p>	
<p><b>GPT-4o + CoT</b></p> <p>Let's analyze the shapes step by step:\n\n1. **Shape x**:\nConsists of three circles in a horizontal line.\n2. **Shape y**:\nConsists of three circles in a horizontal line.\n3. **Shape a**:\nConsists of three circles in a horizontal line.\n4. **Shape c**:\nConsists of three circles in a horizontal line.\n5. **Shape b**:\nConsists of three circles in a horizontal line.\n6. **Shape z**:\nConsists of three circles in a vertical line.\n\nAll shapes except for shape z consist of three circles in a horizontal line. Shape z is the only one with circles arranged vertically.\n\nTherefore, the shape that is different from the others is <b>shape z</b>.</p>	

Figure 17: Responses of GPT-4o on a problem from AVSBench with and without chain-of-thought (CoT) prompt.

	TOTAL				OCR				Absolute position				Adjacency				Angle				Area				Boundary						
	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall			
Random Chance	22.4	19.1	11.7	19.2	4.1	6.9	0.8	4.6	20.8	19.0	12.5	18.2	4.7	10.6	2.6	7.0	25.1	20.2	16.8	22.8	28.7	29.8	13.9	26.1	22.6	17.5	15.4	19.1			
Closed Source Models																															
GPT-4o [40]	75.4	61.6	32.3	62.5	97.3	96.1	47.1	87.4	100.0	96.3	30.0	81.9	79.6	71.7	29.4	68.9	56.8	37.0	17.6	47.3	90.8	68.1	42.9	71.6	95.0	66.7	41.2	72.3			
GPT-4o (+CoT)	75.5	64.6	30.9	63.5	94.5	96.1	44.1	85.8	100.0	96.3	35.0	83.1	83.7	81.1	29.4	74.8	63.1	24.1	17.6	47.3	87.7	79.7	54.3	77.5	98.3	70.4	44.1	75.7			
Gemini-1.5-pro [49]	71.8	57.4	26.9	58.3	94.2	90.8	32.4	80.9	100.0	96.3	20.0	79.5	71.4	79.2	23.5	68.1	60.4	18.5	23.5	44.5	93.8	58.0	22.9	64.5	91.7	81.5	32.4	74.3			
Gemini-1.5-pro (+CoT)	70.8	58.6	27.0	58.4	90.4	92.1	29.4	79.8	100.0	96.3	25.0	77.1	81.6	79.2	11.8	70.6	61.3	27.8	23.5	47.8	92.3	59.4	40.0	68.0	95.0	79.6	26.5	73.6			
Open Source Models																															
LLaVA-NeXT (7B) [29]	36.4	23.8	15.0	27.6	68.5	46.1	8.8	48.1	66.7	59.3	15.0	51.8	16.3	11.3	5.9	12.6	28.8	20.4	17.6	25.3	50.8	33.3	5.7	34.3	31.7	44.4	26.5	35.1			
LLaVA-NeXT (13B)	41.1	28.6	16.4	31.8	79.5	53.9	8.8	55.7	75.0	74.1	10.0	59.0	28.6	32.1	11.8	27.7	22.5	27.8	11.8	23.1	73.8	43.5	11.4	48.5	35.0	40.7	14.7	32.4			
LLaVA-OneVision (7B) [26]	51.0	37.8	18.1	40.0	79.5	69.7	23.5	65.0	94.4	85.2	15.0	72.3	36.7	20.8	5.9	25.2	28.8	22.2	17.6	25.8	58.8	40.6	20.0	50.9	65.0	42.6	8.8	43.9			
Table-LLaVA (7B) [60]	32.5	24.3	13.5	25.9	53.4	26.3	5.9	33.3	47.2	63.0	20.0	45.8	20.4	11.3	5.9	14.3	22.5	25.9	11.8	22.5	41.5	26.1	11.4	29.0	25.0	61.1	26.5	38.5			
Math-LLaVA (13B) [47]	37.7	27.1	15.7	29.6	54.8	38.2	8.8	39.3	50.0	70.4	15.0	48.2	16.3	24.5	17.6	20.2	17.1	18.5	11.8	17.0	67.7	33.3	17.1	43.2	33.3	38.9	11.8	30.4			
Phi-3.5-Vision-Instruct (4B) [11]	49.0	34.8	16.5	37.7	78.1	51.3	8.8	54.1	86.1	70.4	10.0	62.7	38.8	47.2	11.8	38.7	22.5	33.3	23.5	25.8	86.2	39.1	25.7	54.4	51.7	38.9	11.8	37.8			
InternVL2 (8B) [10]	43.0	31.6	13.5	33.3	65.8	52.6	8.8	49.7	30.6	63.0	15.0	37.3	32.7	30.2	17.6	29.4	28.8	18.5	0.0	23.1	60.0	30.4	11.4	37.9	45.0	40.7	0.0	33.1			
DeepSeek-VL (7B) [31]	45.9	30.2	15.1	34.3	69.9	50.0	17.6	51.9	91.7	88.9	25.0	74.7	20.4	22.6	5.9	19.3	32.4	22.2	17.6	28.0	66.2	42.0	14.3	45.6	56.7	42.6	0.0	38.5			
Cardinal				Cardinal Direction				Color				Congruence				Connectedness				Convexity				Coordinate							
easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall				
5.4	3.1	0.0	3.3	27.6	14.9	13.5	20.6	19.9	18.8	13.3	18.6	25.1	25.2	15.5	23.1	8.1	11.5	8.0	9.7	37.7	22.7	17.8	30.6	16.4	17.6	9.3	16.3				
Closed Source Models																															
84.4	66.7	22.7	64.0	86.8	93.8	25.0	80.5	90.4	85.7	46.9	82.7	81.5	26.9	31.8	49.8	71.2	67.1	25.0	60.7	82.3	50.0	69.2	70.6	71.7	52.9	21.4	56.9				
88.3	73.1	31.8	70.1	86.8	93.8	41.7	82.9	91.3	84.5	46.9	82.7	79.0	26.9	20.5	46.3	53.8	48.6	28.6	46.7	85.5	55.9	38.5	70.6	67.9	65.7	21.4	62.0				
84.4	62.4	36.8	65.0	92.1	78.1	41.7	79.3	87.8	79.8	40.6	78.4	67.9	34.6	25.0	45.8	59.6	55.7	21.4	50.7	66.1	41.2	53.8	56.9	45.0	57.1	21.4	61.3				
83.1	59.1	25.0	60.7	92.1	75.0	33.3	76.8	85.2	81.0	56.2	79.7	66.7	34.6	18.2	43.8	69.2	60.0	25.0	56.7	61.3	44.1	53.8	55.0	83.0	61.4	7.1	64.2				
Open Source Models																															
54.5	23.7	6.8	31.3	52.6	28.1	25.0	39.0	58.3	32.1	12.5	42.4	28.4	21.8	18.2	23.6	36.5	5.7	14.3	18.0	48.4	32.4	30.8	41.3	28.3	20.0	14.3	22.6				
66.2	25.8	9.1	36.9	57.9	37.5	50.0	48.8	64.3	41.7	21.9	50.2	29.6	16.7	11.4	20.7	50.0	21.4	21.4	31.3	50.0	35.3	23.1	42.2	22.6	24.3	14.3	22.6				
70.1	39.8	6.8	43.9	60.5	50.0	25.0	51.2	81.7	61.9	18.8	65.8	30.9	25.6	9.1	24.1	50.0	28.6	17.9	34.0	54.8	35.3	38.5	46.8	37.7	24.3	14.3	28.5				
41.6	15.1	6.8	22.9	44.7	25.0	8.3	31.7	49.6	28.6	18.8	37.7	17.3	29.5	11.4	20.7	32.7	11.4	25.0	21.3	41.9	35.3	15.4	36.7	15.1	21.4	0.0	16.8				
54.5	29.0	6.8	33.6	52.6	31.2	25.0	40.2	64.3	27.4	15.6	44.2	25.9	19.2	11.4	20.2	34.6	8.6	17.9	19.3	51.6	23.5	15.4	38.5	22.6	25.7	21.4	24.1				
62.3	34.4	9.1	39.3	73.7	53.1	0.0	54.9	75.7	57.1	15.6	60.6	29.6	24.4	18.2	25.1	25.0	34.3	21.4	28.7	43.5	44.1	46.2	44.0	28.3	34.3	21.4	30.7				
50.6	18.3	6.8	27.6	50.0	40.6	25.0	42.7	65.2	39.3	12.5	48.5	29.6	16.7	27.3	24.1	36.5	22.9	21.4	27.3	54.8	23.5	7.7	39.4	24.5	35.7	14.3	29.2				
66.2	31.2	13.6	40.2	63.2	37.5	33.3	48.8	68.7	38.1	12.5	49.8	30.9	15.4	13.6	21.2	28.8	31.4	25.0	21.2	51.6	32.4	30.8	43.1	18.9	27.1	0.0	21.2				
Curvature				Direction				Interior				Intersection				Length				Line				Overlap							
easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall				
26.2	27.3	16.7	24.2	23.3	17.2	11.8	17.9	23.7	15.0	17.9	19.2	28.8	24.9	8.2	24.6	31.4	21.5	17.1	24.5	20.5	12.9	5.3	16.6	24.4	26.9	2.3	21.6				
Closed Source Models																															
84.6	65.5	67.6	71.8	86.8	51.3	16.7	55.1	93.9	79.7	33.3	80.4	64.3	50.0	25.0	52.4	72.9	69.4	40.0	64.8	76.4	60.5	90.0	72.0	92.9	71.2	29.2	72.7				
89.7	65.5	64.7	72.5	86.8	60.5	12.5	59.4	92.4	84.4	44.4	83.1	61.9	51.4	33.3	53.2	70.8	65.3	36.0	61.5	75.0	79.1	80.0	76.8	89.3	76.3	25.0	72.7				
82.1	41.4	38.2	42.7	81.6	51.3	16.7	53.6	92.4	81.2	33.3	80.4	61.9	50.0	25.0	51.6	66.7	65.3	44.0	61.5	80.6	55.8	70.0	71.2	83.9	71.2	12.5	66.2				
84.6	39.7	38.2	52.7	89.5	50.0	8.3	53.6	95.5	79.7	33.3	81.1	64.3	54.2	25.0	54.8	64.6	57.1	32.0	54.9	79.2	79.1	80.0	79.2	83.9	69.5	25.0	67.6				
Open Source Models																															
12.8	29.3	20.6	22.1	39.5	26.3	12.5	27.5	53.0	35.9	11.1	40.5	38.1	26.4	16.7	29.4	45.8	36.7	24.0	37.7	26.4	18.6	20.0	23.2	10.7	20.3	12.5	15.1				
23.1	29.3	17.6	24.4	39.5	14.5	29.2	23.9	48.5	37.5	22.2	40.5	19.0	19.4	25.0	19.0	45.8	40.8	20.0	38.5	37.5	37.2	10.0	35.2	35.7	33.9	8.3	30.2				
66.7	34.5	17.6	39.7	68.4	31.6	29.2	41.3	71.2	56.2	38.9	60.8	33.3	20.8	0.0	23.0	52.1	49.0	16.0	43.4	61.1	37.2	50.0	52.0	41.1	33.9	12.5	33.1				
17.9	22.4	14.7	19.1	28.9	15.8	12.5	18.8	48.5	32.8	33.3	39.9	26.2	23.6	25.0	24.6	45.8	32.7	12.0	33.6	18.1	11.6	10.0	15.2	12.5	13.6	4.2	11.5				
20.5	25.9	20.6	22.9	36.8	25.0	12.5	26.1	43.9	31.2	16.7	35.1	31.0	20.8	16.7	23.8	45.8	42.9	20.0	39.3	29.2	18.6	10.0	24.0	44.6	23.7	16.7	30.9				
61.5	29.3	17.6	35.9	55.3	32.9	12.5	35.5	59.1	35.9	38.9	46.6	45.2	23.6	16.7	30.2	54.2	55.1	16.0	46.7	54.2	32.6	30.0	44.8	41.1	28.8	8.3	30.2				
43.6	24.1	8.8	26.0	50.0	26.3	8.3	29.7	68.2	43.8	11.1	50.7	33.3	25.0	16.7	27.0	45.8	51.0	8.0	40.2	40.3	30.2	30.0	36.0	46.4	39.0	16.7	38.1				
56.4	36.2	17.6	37.4	60.5	22.4	16.7	31.9	59.1	46.9	16.7	48.6	38.1	31.9	16.7	32.5	39.6	40.8	8.0	33.6	50.0	20.9	20.0	37.6	48.2	16.9	12.5	28.8				
Ordinal				Orientation				Orthogonality				Parallel				Point				Reflection				Relative Position				Rotation			
easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall	easy	medium	hard	overall				
23.3	16.5	2.7	15.4	27.6	33.0	26.0	29.2	24.4	13.5	13.7	18.3	24.3	22.0	18.7	22.4	24.7	20.3	3.9	18.8	26.7	25.2	17.6	24.8	28.3	26.4	18.0	26.3	32.9	27.9	15.5	24.8
Closed Source Models																															
89.1	78.2	30.3	70.1	42.4	42.9	30.0	40.5	60.0	48.4	21.4	48.8	41.1	34.9	16.1	34.0	92.5	87.1	25.0	75.9	53.2	37.1	20.0	41.2	83.3	50.0	40.0	65.5	37.1	21.1	18.0	23.9