

ReasonBench: Benchmarking and Improving Visual Language Models for Complex Graphic Reasoning

Jianyi Zhang^{1*}, Xu Ji¹, Ziyin Zhou¹, Yuchen Zhou¹, Shubo Shi¹, Haoyu Wu¹, Zhen Li¹, Shizhao Liu¹

¹Beijing Electronic Science & Technology Institute

ABSTRACT

Evaluating the performance of visual language models (VLMs) in graphic reasoning tasks has become an important research topic. However, VLMs still show obvious deficiencies in simulating human-level graphic reasoning capabilities, especially in complex graphic reasoning and abstract problem solving, which are less studied and existing studies only focus on simple graphics. To evaluate the performance of VLM on complex graphical reasoning, we propose the first high-granularity evaluation benchmark **ReasonBench** focusing on structured graphical reasoning tasks, which contains 1,613 questions from real-world intelligence tests. ReasonBench covers reasoning dimensions related to location, attribute, quantity, and multi-element tasks, providing a comprehensive evaluation of the performance of VLMs in spatial, relational, and abstract reasoning capabilities. We benchmark 11 mainstream VLMs (including closed-source and open-source models) and reveal significant limitations of current models. Based on these findings, we propose a dual optimization strategy: **Diagrammatic Reasoning Chain (DiaCoT)** enhances the interpretability of reasoning by decomposing layers, and **ReasonTune** enhances the task adaptability of model reasoning through training, all of which improves VLM performance by 33.5%. All experimental data and code are in the Huggingface datasets: <https://huggingface.co/datasets/cistine/ReasonBench>.

ACM Reference Format:

Jianyi Zhang^{1*}, Xu Ji¹, Ziyin Zhou¹, Yuchen Zhou¹, Shubo Shi¹, Haoyu Wu¹, Zhen Li¹, Shizhao Liu¹. 2025. ReasonBench: Benchmarking and Improving Visual Language Models for Complex Graphic Reasoning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (ACM MM Datasets)*. ACM, New York, NY, USA, 28 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

The rapid development of visual language models (VLMs) is reshaping the field of AI, enabling them to jointly understand and reason across visual and textual modalities for the first time. VLMs have made significant advancements on tasks such as image captioning[2] and open-domain visual question answering[4]. However, current benchmark research remains fragmented. Existing datasets (e.g., Raven[35] and CLEVR[17]) lack diversity and cannot fully evaluate

reasoning capabilities. Recent studies have explored dynamic and abstract reasoning, but fail to capture the true complexity of graphical reasoning. Simple graphical reasoning refers to the case where there are only one or a few elements in the graph[10], or the model is used to solve simple geometric angle problems[12]. Complex graphical reasoning refers to the case where there are dimensions such as position, style, attribute, ablation, etc. between graphs, including intersections between these dimensions. And human will be elusive and cramped when dealing with such reasoning. To investigate VLMs' performance in complex graphical reasoning, we perform a comprehensive analysis of existing benchmarks. The results of this study indicate that the ability of VLM to handle complex graph reasoning remains unexplored, as previous studies have mainly focused on simple graph reasoning and the classification and dimensionality of complex graphs are still insufficient. To fill this research gap, we systematically study the complex graphical reasoning capabilities of VLMs in structured settings for the first time, and introduce ReasonBench, a diagnostic benchmark specifically designed to evaluate structured graphical reasoning.

ReasonBench consists of 1,613 standardized test problems and integrates 11 core cognitive reasoning dimensions, including Positional, Stylistic, Attribute, Quantitative, Spatial, and other related dimensions. We systematically evaluate 11 VLMs, which include both closed-source and open-source ones. Additionally, we introduce a triple-controlled evaluation protocol to ensure measurement reliability and cross-model comparability. We also establish a human-level performance baseline, where human participants (mean age = 28.3 ± 6.5 years) achieve an average accuracy of 68.7% on the same test set, offering a crucial reference for future model improvements.

Based on these findings, we identified several key trends that underscore the primary limitations of current VLMs: (1) Even the best-performing models achieve only an average accuracy of 27%. (2) The performance gap between integrated and separated format for VLMs is minimal. Overall, the gap between closed-source and open-source models has not significantly widened.

To enhance VLMs' graphical reasoning capabilities. First, we introduce **Diagrammatic Chain-of-Thought (DiaCoT)**, which leverages visualized, step-by-step reasoning from a layered perspective of the graph to improve interpretability. Second, we propose a fine-tuning strategy **ReasonTune**, which gradually strengthens models' inductive reasoning abilities. Experimental results demonstrate that our optimization framework improves complex graphical reasoning accuracy by 33.5%. Hence, our main contributions include:

- (1) We propose ReasonBench, a more comprehensive benchmark set than previous work, specifically designed to evaluate VLM complex graph reasoning capabilities, covering 11 cognitive dimensions and 29 task types. This dataset addresses the

Correspondence: (zjy@besti.edu.cn).

Permission to make digital or hard copies of all or part of this work for personal or

Unpublished working draft. Not for distribution. Copying, distributing, republishing, or otherwise using this work for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM MM Datasets, 2025, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/XXXXXXX.XXXXXXX>

2025-05-28 14:17. Page 1 of 1–28.

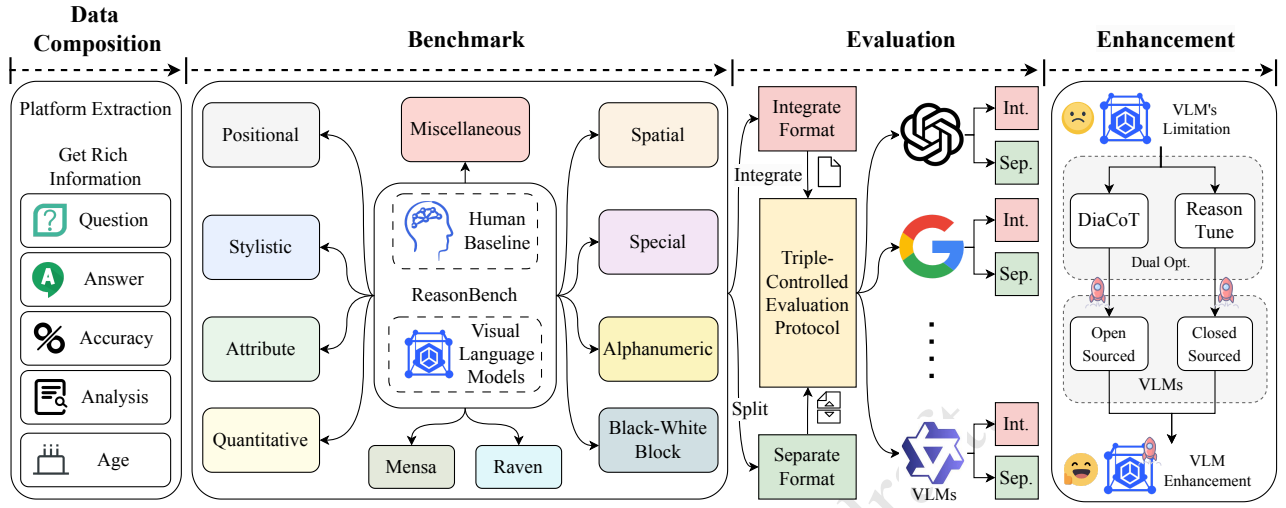


Figure 1: The pipeline of our work. First, in the data combination phase, we obtain rich information. Second, we propose the first benchmark ReasonBench from 11 cognitive dimensions. Third, in the evaluation phase, we evaluate the reasoning ability of VLM by integrating and separating formats. Experiments show that our approach can enhance the ability of VLM.

problem of previous work lacking granularity in practical applications. In addition, we also provide a human performance baseline.

- (2) We systematically test 11 closed-source and open-source VLMs, revealing the main limitations of current VLMs and presenting a series of fascinating findings. We also evaluate VLMs using two input formats, observe and draw a series of conclusions, and we open-source the integrated and separated images and their corresponding URLs.
- (3) We propose a dual optimization framework to enhance VLMs' graphical reasoning abilities, accommodating both open-source and closed-source models. Our DiaCoT and ReasonTune strategies significantly improve performance, achieving a 33.5% accuracy boost in complex graphical reasoning tasks.

2 RELATED WORK

Recent advances in VLMs have enabled significant progress in multimodal tasks such as image captioning[2] and visual question answering[4]. However, evaluating VLMs' capabilities in structured, complex graphical reasoning remains underexplored. Early benchmarks like Raven[35], focused on matrix completion tasks, but lacked diversity in reasoning types. CLEVR[17] emphasized compositional logical reasoning but diverged from real-world testing scenarios. Recent efforts, such as Mementos[29] for sequential dynamic reasoning and PuzzleVQA[10] is used to solve single or double elements reasoning, expanded the scope but still fell short of capturing the complexity of graphical reasoning problems.

Human intelligence tests, such as Mensa and RPM exams, have long served as gold standards for evaluating abstract reasoning. Prior work adapted RPM-style tasks to assess machine intelligence[36], but these efforts often lacked real-world applicability and fine-grained cognitive diagnostics. The Raven dataset[35] primarily focuses on RPM-style matrix completion but lacks diversity in problem types,

thus constrain its ability to evaluate broader reasoning capabilities. The CLEVR dataset[17] emphasizes compositional logical reasoning but deviates significantly from real-world testing scenarios. Existing evaluations[24] prioritized accuracy over limiting insights into models' failure modes. Recent studies also identified weaknesses in VLMs' spatial and relational reasoning, particularly in tasks included occlusion, composite transformations, and 3D geometric operations[18]. Although the current graphical reasoning benchmarks are very extensive, as shown in Table 1, their fine-grained evaluation scope is very limited except for MM-IQ. Mensa and RAVEN are mainly widely used in IQ tests, and do not touch on aspects such as quantitative operations, spatial geometry or attribute dimensions, which is different from this study in terms of evaluation dimensions. G-set[21] still tests RAVEN questions and does not have other dimensions. Although VAP[14] has graphical content, it focuses more on the analogy level. Bongard-LOGO[22] focuses on the realization of images by human action strokes. SVRT[13] is a test established from the perspective of lines and positioning. DOPT[30] is used to evaluate the gradient change of the model's reasoning ability, mainly by scaling up or translating the data. The ARC[11] benchmark focuses on object continuity and basic geometry, which is similar to RAVEN. SPACE[23] focuses on the benchmark of spatial cognition, mainly examining the model's ability to understand the surrounding environment. MARVEL[16] and MM-IQ[7] both focus more on complex graphical reasoning capabilities, but the dimensions they focus on are still insufficient compared to this study. At the same time, these two articles send graphical questions and answers to the model at the same time, ignoring the effect of sending questions and answers to the model separately. Although existing benchmarks and methods have advanced the field, they exhibit fragmented task coverage, limited real-world relevance, and insufficient diagnostic granularity. Similarly, HellaSwag[34], though primarily text-based, inspired multimodal extensions but lacked alignment with complex

Dimensions	Task	Mensa	RAVEN	G-set[21]	VAP[14]	Bongard-LOGO[22]	SVRT[13]	DOPT[30]	ARC[11]	SPACE[23]	MARVEL[16]	MM-IQ[7]	ReasonBench
Positional	Trans.	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
	Rot.	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓
	Comb.				✓	✓	✓				✓	✓	✓
Stylistic	Trav.												✓
	Add.& Sub.	✓	✓	✓	✓							✓	✓
	B&W											✓	✓
Attribute	Sym.	✓	✓	✓					✓		✓	✓	✓
	O&C											✓	✓
	Comb.											✓	✓
Quantitative	Line											✓	✓
	Surf.										✓	✓	✓
	Point											✓	✓
	Elem.	✓	✓	✓	✓		✓		✓			✓	✓
	Comb.											✓	✓
Spatial	Cube									✓	✓	✓	✓
	3D-Mos.									✓	✓	✓	✓
	Poly.										✓	✓	✓
	3-view									✓	✓	✓	✓
	Sec. View										✓	✓	✓
Special	Q&T										✓	✓	✓
	2D Comp.					✓	✓				✓	✓	✓
Alphanumeric	InterFig.					✓	✓		✓			✓	✓
AlphaNumeric	AlphaNum.	✓	✓	✓								✓	✓
Black&White Block	B&W	✓										✓	✓
Miscellaneous	Misc.											✓	✓
Mensa	Mensa-v1	✓										✓	✓
	Mensa-v2	✓										✓	✓
Raven	Raven-v1		✓									✓	✓
	Raven-v2		✓									✓	✓

Table 1: Comparison table of related work.

graphical reasoning problem-solving. Therefore, our work bridges these gaps by introducing ReasonBench, the first benchmark derived from standardized human intelligence tests.

To address reasoning limitations, methods like Chain-of-Thought (CoT) prompting[31] and neuro-symbolic approaches[20] have been proposed. CoT improved the interpretability of the model in text-based reasoning, but struggled with visual complexity. Neuro-symbolic frameworks integrated symbolic rules with neural networks, but required extensive domain-specific engineering. Fine-tuning strategies, such as progressive training on synthetic data[17], enhanced specific capabilities but failed to generalize across diverse reasoning dimensions. Therefore, we propose DiaCoT and ReasonTune, a dual optimization framework that combines visualized reasoning chains with structured fine-tuning to address the cognitive bottlenecks of VLMs. while addressing their limitations in multimodal, complex graphical reasoning reasoning scenarios.

3 REASONBENCH

3.1 Dataset Composition

ReasonBench is a meticulously curated dataset consisting of 1,613 real-world graphical reasoning problems, systematically gathered from three primary sources: Chinese Civil Service Aptitude Tests[19], Mensa Intelligence Tests¹, and Raven's Progressive Matrices². The dataset includes 11 cognitive dimensions, which are further categorized into 29 distinct task types. This multi-source hierarchical architecture enables comprehensive evaluation of VLMs analytical capabilities across diverse reasoning paradigms. Table 5 shows the details.

¹<https://www.mensa.org/>

²<https://psycho-tests.com/test/raven-matrixes-test>

3.2 Human Performance

To further explore the comparison between VLMs and human reasoning abilities, we extracted data from existing resources. We collect accuracy results for 1439 questions from the FenBi³ platform, with participants in the age group 21 to 35, including the questions themselves, correct answers, human accuracy, and detailed analysis. Additionally, for the Mensa and Raven tests, we invited three participants aged from 22 to 25, who voluntarily completed the tests with the consent of their respective mentors. Our analysis revealed an average baseline human score of 69.76%.

3.3 Evaluation Procedure and Metrics

Our experimental framework systematically evaluates the reasoning capabilities of VLMs on complex graphical reasoning tasks. As illustrated in Figure 1 Evaluation phase. We tested 11 VLMs in batches on ReasonBench. Meanwhile, we investigated the different effects of integrated and segmented visual presentation formats on VLM. Due to space limitations, the specific format is shown in Appendix B.2).

- **Integrated format:** The problem and answer choices are presented in a single graphic.
- **Separated format:** The problem and answer choices are split into multiple graphics and sequentially fed into the model.

We apply the Triple-controlled Evaluation Protocol (see Section 3.4) to both formats to address the lack of real-world applicability of previous benchmarks. Finally, we use automatic keyword extraction to systematically analyze model responses. This evaluation is used to measure the reasoning ability of different VLMs on graphics using the accuracy (ACC) metric. For examples of each task, see Appendix B.3.

³<https://fenbi.com/>

Task	Closed-source VLMs						Open-source VLMs					Human Eval.	Int. Avg	Sep. Avg
	GPT 4o	Gemini 2.0	Gemini 1.5	Claude 3.5	GLM	Yi	Qwen	Intern VL2	PixTral	QVQ	Deep Seek			
Trans.	27.66	22.34	27.66	24.47	23.40	25.53	32.98	14.89	28.72	26.60	30.85	73.28	25.92	26.73
Rot.	<u>28.57</u>	23.21	30.36	21.43	23.21	23.21	30.36	30.36	19.64	25.00	25.00	77.30	25.49	28.79
Comb.	30.00	36.67	23.33	16.67	23.33	20.00	43.33	40.00	23.33	40.00	30.00	75.53	29.70	25.83
Trav.	24.07	31.48	22.22	27.78	31.48	24.07	22.22	24.07	22.22	29.63	22.22	69.07	25.59	25.00
Add.&Sub.	32.84	26.87	23.88	20.90	29.85	28.36	29.85	23.88	32.84	16.42	31.34	73.19	27.00	28.17
B&W	23.81	34.92	26.98	33.33	14.29	25.40	23.81	36.51	20.63	12.70	17.46	70.67	24.53	26.59
Sym.	35.78	<u>28.44</u>	19.27	16.51	34.86	21.10	22.94	22.02	<u>27.52</u>	22.94	20.18	73.61	24.69	25.11
O&C	21.05	26.32	31.58	15.79	36.84	36.84	31.58	47.37	36.84	5.26	31.58	73.89	29.19	30.26
Comb.	16.67	16.67	16.67	16.67	16.67	16.67	16.67	33.33	50.00	33.33	16.67	72.50	22.73	14.58
Line	20.81	31.21	<u>25.43</u>	23.12	<u>25.43</u>	28.32	21.39	21.97	18.50	20.81	<u>24.86</u>	64.71	23.80	24.06
Surf.	23.36	<u>22.63</u>	<u>27.01</u>	28.47	<u>26.28</u>	18.25	<u>27.01</u>	21.17	21.90	<u>27.01</u>	<u>26.28</u>	71.50	24.49	26.55
Point	28.79	15.15	16.67	21.21	28.79	16.67	22.73	30.30	24.24	21.21	19.70	65.24	22.31	25.00
Elem.	21.28	27.66	23.40	24.47	25.53	24.47	30.85	18.09	21.28	17.02	29.79	63.97	23.98	25.80
Comb.	20.00	22.00	16.00	30.00	34.00	38.00	20.00	<u>26.00</u>	26.00	24.00	26.00	64.30	25.64	24.25
Cube	<u>27.52</u>	<u>25.69</u>	23.85	28.44	<u>27.52</u>	28.44	21.10	<u>27.52</u>	30.28	21.10	20.18	64.31	25.60	25.11
3D Mos.	17.39	36.96	<u>30.43</u>	15.22	32.61	23.91	26.09	19.57	23.91	19.57	17.39	54.46	23.91	25.54
Poly.	23.53	17.65	29.41	35.29	17.65	23.53	35.29	35.29	29.41	17.65	29.41	60.18	26.74	22.06
3-View	32.50	25.00	30.00	25.00	25.00	25.00	30.00	25.00	25.00	25.00	17.50	68.50	25.91	23.75
Sec.View	25.71	17.14	22.86	22.86	42.86	28.57	31.43	31.43	34.29	14.29	<u>28.57</u>	55.69	27.27	28.93
Q&T	40.00	30.00	20.00	20.00	30.00	20.00	40.00	10.00	30.00	20.00	10.00	68.80	24.55	32.50
2D Comp.	<u>29.03</u>	25.81	19.35	32.26	19.35	19.35	19.35	38.71	22.58	<u>32.26</u>	22.58	70.58	25.51	25.00
InterFig.	40.00	30.00	27.50	22.50	30.00	32.50	27.50	22.50	32.50	17.50	22.50	68.18	27.73	20.94
AlphaNum.	29.63	<u>33.33</u>	<u>33.33</u>	40.74	25.93	33.33	37.04	22.22	29.63	40.74	18.52	54.96	31.31	32.41
B&W	28.13	21.88	31.25	21.88	21.88	21.88	25.00	18.75	31.25	12.50	25.00	68.47	23.58	23.83
Misc.	<u>35.29</u>	<u>38.24</u>	47.06	<u>44.12</u>	38.24	14.71	23.53	20.59	<u>41.18</u>	32.35	<u>41.18</u>	60.44	34.22	30.88
Mensa-v1	17.14	34.29	20.00	31.43	11.43	5.71	17.14	8.57	11.43	31.43	28.57	80.00	19.74	16.79
Mensa-v2	20.51	23.08	12.82	15.38	20.51	15.38	20.51	10.26	15.38	20.51	25.64	84.62	18.18	14.10
Raven-v1	25.00	<u>30.00</u>	37.50	37.50	25.00	17.50	22.50	25.00	25.00	30.00	10.00	80.00	25.91	16.88
Raven-v2	25.00	31.67	21.67	43.33	35.00	23.33	21.67	16.67	28.33	23.33	30.00	95.00	27.27	19.38
All Avg.	26.22	27.22	25.05	25.85	24.49	23.93	25.73	23.62	25.17	22.88	24.55	69.76	24.97	25.23

The **bolded values** represent the highest accuracy results and the underlined values represent results above the average accuracy for each task.

Table 2: The accuracy of different VLMs on ReasonBench.

3.4 Triple-controlled Evaluation Protocol

In the task design phase of our evaluation, we systematically explored various response formats. Through an in-depth analysis of the characteristics of visual reasoning tasks, we identify significant modality differences between graphic-presented questions and options and the text-based responses generated by the model. To mitigate the impact of non-cognitive factors on the validity of the evaluation, we established a standardized evaluation protocol with triple control, as outlined below:

1. We use multiple choice question (MCQ) as the evaluation question type. Except for the Mensa IQ test and Raven’s reasoning matrix, which retain 6-8 options, all other questions are designed with four options. At the same time, we ensure that the frequency fluctuation of each correct option (A/B/C/D) in all questions is controlled within $25\% \pm 0.7\%$ (see Appendix A.1).

2. Both integrated and separated formats use distinct, fixed templates. A fixed prompt template is used to ensure consistent input formatting for each question, while a structured response format (`<ans>OptionX</ans>`) is applied to minimize the impact of non-cognitive factors on model outputs (see Appendix B.1). Regardless

of the format type, whether integrated or separated, graphics are converted into a universally accessible API URL format to ensure compatibility across all models, with all data being made publicly available.

3. A standardized evaluation protocol based on Pass@1 single-attempt scoring [8] is applied to ensure the objectivity and consistency of the evaluation results (see Appendix A.2).

4 EXPERIMENTS

In our experimental analysis, we explore two key Research Questions:

RQ1: How VLMs perform in complex graphical reasoning?

We evaluate closed-source and open-source VLMs on complex graphic-based problems, assessing their reasoning capabilities across multiple cognitive dimensions.

RQ2: What is the impact of input strategies? Using a separated format, we sequentially present independent graphical dimensions to analyze input format effects on model performance.

4.1 Models

This study is based on Chatbot Arena Leaderboard⁴, an authoritative evaluation ranking published on HuggingFace. A total of 11 representative VLMs were selected to construct the benchmarking framework. The selected models are categorized into closed-source and open-source groups, strictly adopting the best-performing versions officially validated within each model series as of December 25, 2024. The closed-source group includes six commercial systems: GPT-4o[15], Gemini-2.0[25], Gemini-1.5[26], Claude-3.5[3], GLM-4V[28], and Yi-Vision[33]. The open-source group consists of five reproducible implementations: Qwen-VL-72B[6], InternVL2-26B[9], Pixtral[1], QVQ-72B-Preview[27], and DeepSeek-VL2[32]. The architectural specifications, training protocols, and implementation details of all models are fully presented in Appendix A.3.

4.2 Benchmark evaluation

We evaluated all baseline VLMs on ReasonBench and summarized the results in Table 2. Human Baseline indicates the accuracy of human completing the ReasonBench questions. Int. Avg refers to inputting the integrated graphic into VLMs and Sep. Avg refers to separated graphical input them into VLMs sequentially. All Avg. represents the average accuracy for each column. Additionally, we provide the results of the separated format in the Table 6 in Appendix. Our findings are summarized as follows:

There are significant differences in the performance of different VLMs in various graphics reasoning tasks. Overall, open-source models outperformed closed-source models in this category. See Table 3 for details.

The overall answer accuracy of VLMs remains consistently around 25%, with even the best-performing model, Gemini-2.0, achieving only 27%. This result is surprising, as even a model that randomly selects answers would achieve a 25% accuracy rate among four options. In terms of overall performance, there is a significant disparity between the capabilities of VLMs and humans. Specifically, even the best-performing model falls far behind the lowest human baseline, with a gap of up to 13%, while the difference with the highest-performing human baseline exceeds 50%. These findings suggest that current VLMs are not yet capable of independently reasoning and accurately selecting answers from complex graphics. Therefore, the ReasonBench benchmark is crucial in highlighting the limitations of VLMs in complex graphical reasoning tasks. In other words, at least in the context of real-world complex graphical reasoning tasks, the performance of VLMs has not yet reached a level of reliability.

Gemini-2.0 and Qwen-72B-VL represent the best-performing closed-source and open-source models, respectively. From the data in the table, it is evident that Qwen-72B performs slightly worse than Gemini-2.0 in the Mensa and Raven tests, with Qwen-72B lagging behind in each task. Although models such as Yi-vision, QVQ-72B, and Deepseek-VL are specifically optimized for graphic analysis tasks, they do not show a more significant advantage over Qwen in overall performance. Notably, while QVQ outperforms Qwen-72B in various metrics in official showcases, it ranks lower in our benchmark tests. In the overall assessment, open-source models

Dimensions	Results
Positional	For the position-pattern task, closed-source models such as Claude-3.5, GLM-4V, and Yi-Vision all performed below average, while the open-source model Qwen-72B achieved the best result in this task. This suggests that closed-source models have certain limitations when handling position-related reasoning tasks.
Stylistic	For the style-pattern task, closed-source models generally outperformed open-source models. The number of closed-source models performing above average was notably higher, while open-source models tended to perform below average in both tasks.
Attribute	Regarding attribute-pattern tasks, Claude-3.5 performed below average in all related tasks, whereas PixTral achieved excellent results in all attribute-based tasks. However, it is worth noting that QVQ performed particularly poorly on the open-closed tasks, achieving only a 5.26% accuracy rate. This may be due to QVQ's insufficient focus on the open-closed reasoning problem during training.
Quantitative	In the quantity-pattern tasks, although GLM-4V did not achieve the best score in any single task, its performance was stable, remaining above average across all tasks. DeepSeek-VL2 also performed well, surpassing the average level in four out of five tasks, whereas GPT-4o and QVQ performed relatively weak, achieving strong results in only one task.
Spatial	Spatial-pattern tasks are the most challenging category for human baselines. The results show that Qwen-72B and GPT-4o performing the best in spatial-pattern tasks, achieving the optimal scores in two tasks and surpassing the average in most tasks. In contrast, QVQ performs below average in all tasks of this category.
Special	In the special-pattern tasks, GPT-4o and Gemini-2.0 shows strong performance overall, exceeding the average and achieving the best results in certain tasks, while Gemini-1.5, Qwen-72B, and DeepSeek-VL2 performs relatively poorly.
Alphanumeric	In word and letter-pattern tasks, Claude-3.5 and Qwen-72B stood out, achieving the best results among all models
Black&White Block	In the black-and-white operation tasks, Gemini-1.5 and PixTral performs the best.
Miscellaneous	In Miscellaneous tasks, closed-source models like GPT-4o and Gemini-2.0 outperformed open-source models, with almost all closed-source models achieving results above average.
Mensa	We explored the performance of VLMs in Mensa and Raven's tests. The results show that Gemini-2.0 performed exceptionally well in these types of tasks, exceeding the average in all four tasks and achieving the best scores in certain tasks. Meanwhile, Claude-3.5, QVQ, and DeepSeek-VL2 each excelled in different aspects of the Mensa and Raven's tests, while Yi-Vision and InternVL2 performed relatively poorly, showing weaker reasoning abilities.
Raven	

Table 3: An overview of the VLMs' evaluation results under different Dimensions. Dimensions correspond to those in Table 5 and Result reflect the experiment details in Table 2.

generally outperform closed-source models across most tasks. However, it is noteworthy that Qwen and pixtral have surpassed GLM and Gemini-1.5, achieving leading positions in multiple tasks. Despite this, when comparing across different task categories, closed-source models still lead in a greater number of tasks, including cases of

⁴<https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard>

shared first place. The recently popular Deepseek, we also tested its VL2 version, which only achieved a leading position in one task.

Overall, the difference in performance between the integrated and separated approaches for VLMs is minimal. In general, the integrated approach achieves an accuracy of 25.26%, while the separated approach reaches an accuracy of 25.23%. According to the comparison data in Table 1 and Appendix A.4 Table 6, in tasks involving positional dimensions and style dimensions, the separated questioning method outperforms the integrated method for models like Gemini-1.5, Claude-3.5, and Yi-vision. However, Qwen shows superior performance in positional dimension tasks when using the integrated method, significantly outperforming the separated method. For attribute dimension tasks, Gemini-2.0 and Claude-3.5 perform better with the separated method than with the integrated method. In numerical dimension tasks, Gemini-1.5 and InternVL perform better with the separated method. For spatial dimension tasks, Gemini-1.5 and Qwen show better results with the integrated method compared to the separated method. In black-and-white arithmetic and comprehensive dimension tasks, the integrated method outperforms the separated method for Gemini-2.0, Gemini-1.5, and Claude-3.5. Among the 232 sets of data compared, 38 sets were completely consistent.

5 IMPROVEMENT METHODS

We propose a dual optimization solution, which is suitable for closed-source and open-source models, Diagrammatic Chain-of-Thought (DiaCoT) and fine-tuning strategy (ReasonTune).

5.1 DiaCoT

To enhance the generalization of the method across both open-source and closed-source models, we first consider prompt engineering and explore effective problem-solving strategies based on Chain-of-Thought[31]. While strategies such as visual perception, inductive reasoning, and deductive reasoning are highly effective in addressing simple single-element and two-element problems[10], their performance is limited when dealing with more complex multi-element problems. Therefore, we propose a novel reasoning method—**DiaCoT**.

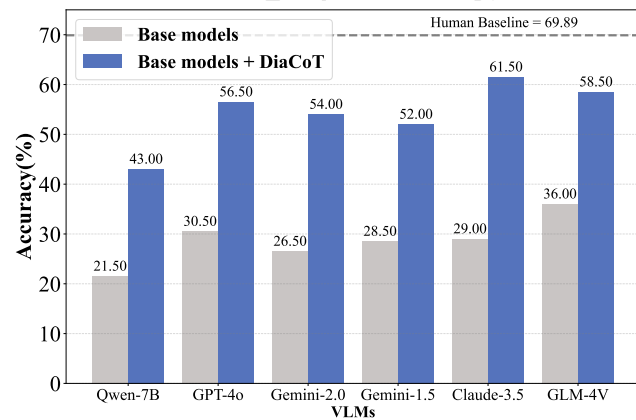


Figure 2: Comparison between Base models and DiaCoT method.

Specifically, in designing prompts, we guide the VLMs to explain the content of the question and options individually and decompose multi-element diagrams layer by layer from a graphical perspective. In addition, DiaCoT integrates the contextual information of the problem environment, providing the model with the necessary context, defining the problem scope, and enhancing the model's abstract reasoning capabilities through the prompts. Detailed implementations and examples can be found in Appendix B.1.

In experimental testing, we compare DiaCoT with the baseline methods. As shown in Figure 2. All models show an apparent improvement in accuracy with an average increase of 25.58% after applying DiaCoT and narrowing the gap with human-level performance. Although the accuracy has not yet achieved human-level precision, it has successfully exceeded the limitation where models typically perform below 25% accuracy or rely on random guessing.

5.2 ReasonTune

We selected Qwen-7B[5] as the open source model for ReasonTune optimization (due to its poor performance), and fine-tuned and optimized it using datasets other than the validation set. As shown in Figure 3, the accuracy of Qwen-7B is only 21.5%. Through ablation experiments, we verified the synergy of dual optimization: (1) ReasonTune on the base model improved the accuracy by 6.5%. (2) The introduction of the DiaCoT method improved the performance by 21.5%. (3) The proposed dual optimization framework achieved an overall improvement of 33.5%. This gradual performance improvement not only verifies the effectiveness of our optimization, but also provides a new optimization paradigm for models that handle complex graph reasoning.

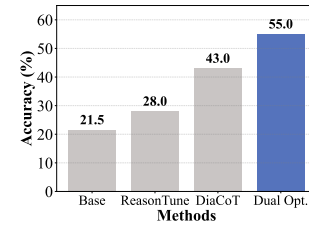


Figure 3: Ablation results of Dual Optimization.

6 CONCLUSION

In order to fairly and effectively evaluate the reasoning ability of VLMs on complex graph reasoning task, we develop a new benchmark named ReasonBench, which contains 1,613 questions across 11 cognitive dimensions and 29 task types. We select 11 mainstream VLMs and evaluate their performances on ReasonBench. The results reveal that the existing models have limitations: even the best-performing model can only achieve 27% accuracy, far below the human baseline of 68.7%. This gap highlights that the ability of VLMs to reason about complex graphs needs to be improved. To address these issues, we propose a dual optimization framework. The DiaCoT method enhances the interpretability of reasoning by decomposing layers, while ReasonTune enhances reasoning task adaptability through training. Experimental results show that the overall reasoning accuracy is improved by 33.5%, verifying the effectiveness of our framework. Our study lays a foundation for developing VLMs with complex graph reasoning.

REFERENCES

- [1] Praveesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12B. *arXiv preprint arXiv:2410.07073* (2024).
- [2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3674–3683.
- [3] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966* (2023).
- [7] Huanqia Cai, Yijun Yang, and Winston Hu. 2025. MM-IQ: Benchmarking Human-Like Abstraction and Reasoning in Multimodal Models. *arXiv preprint arXiv:2502.00698* (2025).
- [8] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.
- [10] Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. 2024. PuzzleVQA: Diagnosing Multimodal Reasoning Challenges of Language Models with Abstract Visual Patterns. *arXiv preprint arXiv:2403.13315* (2024).
- [11] François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547* (2019).
- [12] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. 2024. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885* (2024).
- [13] François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. 2011. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences* 108, 43 (2011), 17621–17625.
- [14] Felix Hill, Adam Santoro, David GT Barrett, Ari S Morcos, and Timothy Lillicrap. 2019. Learning to make analogies by contrasting abstract relational structure. *arXiv preprint arXiv:1902.00120* (2019).
- [15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [16] Yifan Jiang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, Jay Pujara, et al. 2024. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems* 37 (2024), 46567–46592.
- [17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.
- [18] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.
- [19] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmalu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212* (2023).
- [20] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* 1, 8 (2024), A10a2400196.
- [21] Jacek Mańdziuk and Adam Żychowski. 2019. DeepIQ: A human-inspired AI system for solving IQ test problems. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
- [22] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. 2020. Bongard-logo: A new benchmark for human-level concept learning and reasoning. *Advances in Neural Information Processing Systems* 33 (2020), 16468–16480.
- [23] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. 2024. Does Spatial Cognition Emerge in Frontier Models? *arXiv preprint arXiv:2410.06468* (2024).
- [24] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).
- [25] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
- [26] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
- [27] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* (2024).
- [28] Weihai Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079* (2023).
- [29] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529* (2024).
- [30] Taylor Webb, Zachary Dulberg, Steven Frankland, Alexander Petrov, Randall O'Reilly, and Jonathan Cohen. 2020. Learning representations that support extrapolation. In *International conference on machine learning*. PMLR, 10136–10146.
- [31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [32] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *arXiv:2412.10302 [cs.CV]* <https://arxiv.org/abs/2412.10302>
- [33] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652* (2024).
- [34] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830* (2019).
- [35] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5317–5327.
- [36] Chi Zhang, Baoxiong Jia, Feng Gao, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. 2019. Learning perceptual inference by contrasting. *Advances in neural information processing systems* 32 (2019).

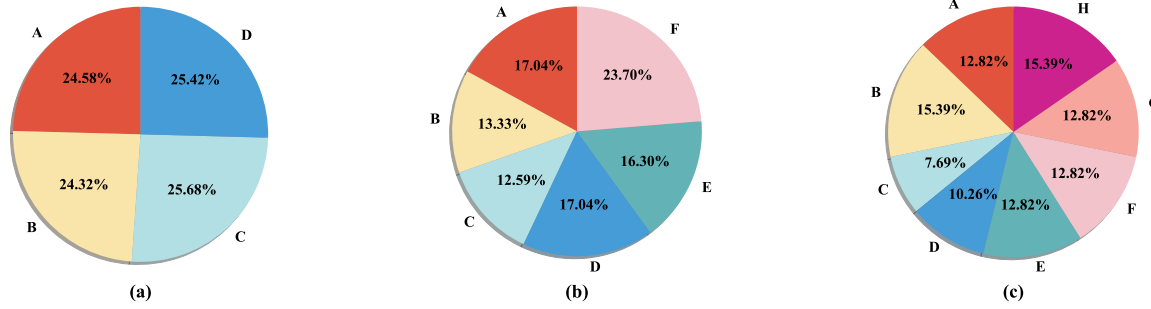


Figure 4: Pie charts of the MCQ distribution. These three charts represent the probability distribution of selecting each choice when the number of choices is 4, 6, and 8, respectively. We label the probability of selecting different choice within the sectors and use different colors for distinction.

Model Type	Model Name	Vendor	Remarks
Closed-sourced	GPT-4o	OpenAI	Multimodal extension of GPT-4 series with cross-modal understanding
	Gemini-2.0-flash-exp1219	Google	Flagship multimodal model released December 2024
	Gemini-1.5-pro-002		Early version supporting long-context (up to 1M tokens) and multimodal I/O
	Claude-3.5-sonnet	Anthropic	State-of-the-art model competitive with GPT-4o on multimodal benchmarks
	GLM-4v-plus	Zhipu AI	Commercial multimodal model with enhanced visual reasoning capabilities
	Yi-vision	01.AI	Proprietary 179B parameter model with hybrid open/closed strategy
Open-sourced	Qwen2-VL-72B	Alibaba	Open-source multimodal model from Qwen series with 72B parameters
	QVQ-72B-Preview		Experimental preview model, community-developed variant with 72B parameters
	InternVL2-26B	SenseTime	Vision-language model with 26B parameters
	Mistral-8x22B	Mistral AI	Sparse mixture-of-experts model with 176B total parameters
	DeepSeek-VL2	DeepSeek	High-performance multimodal model for AI R&D applications

Table 4: Model Information Comparison. We classify the models that used in our paper based on whether they are open-source and label the vendor of each one along with a brief remark.

A APPENDIX

A.1 Multiple Choice Question (MCQ) Distribution

The distribution of MCQ is shown in Figure 4. Here, Figure 4(a) represents the probability distribution of selecting each option when there are four choices, Figure 4(b) represents the probability distribution when there are six choices, and Figure 4(c) shows the probability distribution when there are eight choices. It can be observed that when the number of choices is four or eight, the probabilities of selecting each choice are very close. When there are six, the probability of selecting choice F is slightly higher. Overall, as the number of choice increases, the probabilities of selecting each one remain fairly balanced. And there is no significant preference for any specific choice or a few ones.

A.2 Pass@k

To evaluate the graphical reasoning capabilities of VLMs, we introduce Pass@k metric from the code generation domain. We choose **Pass@1** because it directly reflects the model's deterministic understanding of the problem. A high Pass@1 rate indicates that the model consistently generates the correct solution, rather than relying on "luck", while a high Pass@100 may rely on diversity in generation to mask the model's deficiencies.

A.3 VLMs

Model version&company information, see Table 4.

A.4 Separation format evaluation results

Due to the fact that some VLM API services were interrupted due to force majeure and GLM does not support uploading 7 graphics at the same time, we conducted the test on 8 VLMs. See Table 6 for the evaluation results of the separation format.

A.5 Effect of Few-Shot Demonstrations

In the evaluation, we use zero-shot demonstrations to evaluate how VLM handles complex graph reasoning tasks. In addition, we explore the performance of the model when using few-shot demonstrations. To ensure the diversity of demonstrations, we randomly select questions of different dimensions from the database for each few-shot demonstration.

The results show that the performance improves with the number of demonstrations. However, we find that for GPT-4o, the performance does not reach a peak even with the largest number of demonstrations. This suggests that the relationship between reasoning ability and contextual learning may be proportional but nonlinear when dealing with similar complex graph reasoning tasks in the real

world. Nevertheless, we still prove that few-shot demonstrations are a reliable method, based on the performance of Qwen-7B, which highlights the effectiveness of this method in complex graph reasoning tasks.

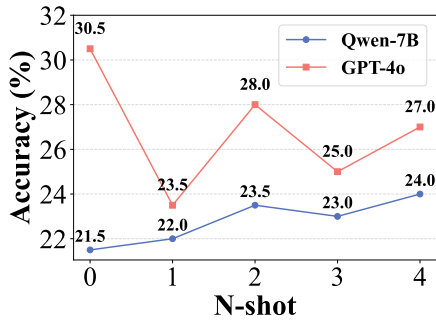


Figure 5: Analysis of the impact of few-shot demonstrations on model performance in complex graph reasoning.

B EXAMPLE

B.1 Format and DiaCoT Example

Integrated and separated format prompts, look at Figures 6-7. See Figure 8 for an example of DiaCoT.

B.2 Int&Sep Example

See Figure 9 for the specific forms of integration and separation formats.

B.3 Task Example

For specific examples of our benchmark, please see Figure 10-38.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please make sure your answer is accurate and meets the requirements of graphic reasoning.

Requirements:

1. Observe the picture carefully and find the pattern of the pattern.
2. Determine the correct answer based on the four options ABCD, and make sure to fill in the question mark.
3. The answer must be accurate and meet the requirements of graphic reasoning.
4. Clear, you must give the answer options, for example, if A is selected for a question, your answer should be:

<begin>

Write the reasoning process here

<ans>optionA</ans>

<end>

5. Let's describe this image first and think step by step.

Figure 6: The prompt words of the integrated format. The blue part represents the system prompt words; the black part represents the requirements; the green part is the fixed model answer format; the red part uses the COT method.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please make sure your answer is accurate and meets the requirements of graphic reasoning.

In the label:

1. question: This is a graphic reasoning question.
2. optionA: This is the picture of option A.
3. optionB: This is the picture of option B.
4. optionC: This is the picture of option C.
5. optionD: This is the picture of option D.

Requirements:

1. Observe the picture carefully and find the pattern of the pattern.
2. Determine the correct answer based on the four options ABCD, and make sure to fill in the question mark.
3. The answer must be accurate and meet the requirements of graphic reasoning.
4. Clear, you must give the answer options, for example, if A is selected for a question, your answer should be:

<begin>

Write the reasoning process here

<ans>optionA</ans>

<end>

5. Let's describe this image first and think step by step.

Figure 7: Prompt words in separated format. Blue indicates system prompt; black parts with red frames indicate graphic labels after cutting, and others indicate requirements; green parts have fixed model answer formats; red parts use COT method.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please make sure your answer is accurate and meets the requirements of graphic reasoning.

DiaCoT: First, observe whether the element composition is the same, and determine whether there is any obvious positional pattern, considering adjacent comparison; secondly, decompose the figures and analyze from the black block level. In the first group of figures, Figure 1 and Figure 2 have a black block in the same position, and Figure 2 and Figure 3 also have a black block in the same position; again, verify and apply this rule to the second group of figures to determine whether Figure 1 and Figure 2 have black blocks in the same position; finally, please determine what pattern should be at the "?" and give an answer.

Requirements:

1. Observe the picture carefully and find the pattern of the pattern.
2. Determine the correct answer based on the four options ABCD, and make sure to fill in the question mark.
3. The answer must be accurate and meet the requirements of graphic reasoning.
4. Clear, you must give the answer options, for example, if A is selected for a question, your answer should be:

<begin>

Write the reasoning process here

<ans>optionA</ans>

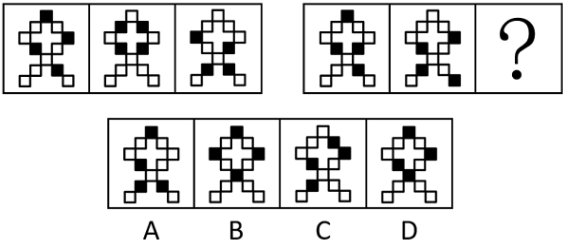
<end>

5. Let's describe this image first and think step by step.

Figure 8: Taking the graph in Figure 10 as an example, the specific application of the DiaCoT method is shown.

Prompt:

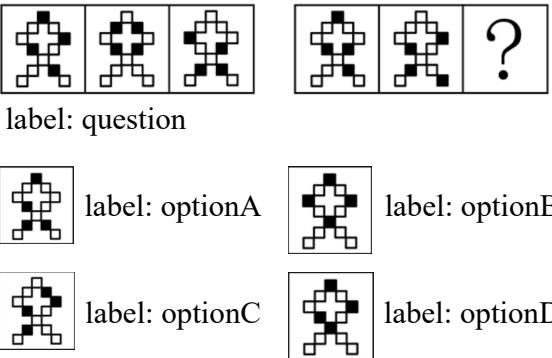
You are a person with high IQ and graphic reasoning ability....



Integrated format.

Prompt:

You are a person with high IQ and graphic reasoning ability....



Separated format.

Figure 9: Examples of integrated and separated formats

Dimensions	Task	Num	Introduction	Purpose
Positional	Translation	94	Translation refers to the movement of a figure along a fixed direction and distance	Evaluate whether VLMs could accurately recognize movement patterns and infer spatial transformations.
	Rotation	56	Rotation involves figures rotating around a fixed point or axis, requiring the model to understand angular changes and symmetry.	
	Combination	30	Combination tasks incorporate both translation and rotation, increasing complexity by requiring models to jointly infer multiple transformation types.	
Stylistic	Traversal	54	Traversal involves elements appearing sequentially across images, requiring the model to extract common patterns and predict missing components.	Evaluate VLMs’ sensitivity to graphical transformations and their ability to process visual modifications across different dimensions.
	Additive & Subtractive	67	Additive & subtractive changes refer to figures incrementally merging or eroding, necessitating inference of transformation rules.	
	Black&White Operations	63	Black&white operations test the model’s ability to interpret changes in color, shading, and contrast.	
Attribute	Symmetry	109	Symmetry tasks require models to recognize axial or central symmetry in figures.	Evaluate whether VLMs can accurately identify and reason about geometric attributes.
	Open & Closed States	19	Open&closed states involve distinguishing whether a shape is enclosed or has open boundaries.	
	Combination	6	Combination tasks integrate both properties, increasing complexity.	
Quantitative	Lines	173	The number of points, lines, planes, or elements may follow an increasing or decreasing pattern, requiring the model to detect these numerical changes.	Evaluate VLMs' ability to recognize numerical relationships and infer implicit rules.
	Surface	137		
	Points	66		
	Elements	94		
	Combination	50	Combination tasks incorporate multiple dimensions of quantity variation, demanding higher-level abstraction and reasoning.	
Spatial	Cube	109	Cube matching requires models to identify which option corresponds to a given unfolded cube.	Evaluate VLMs’ spatial modeling and 3D reasoning capabilities.
	3D-Assembly	46	3D assembly involves selecting the correct solid formed by assembling given geometric components.	
	Polyhedral	17	Polyhedral folding requires models to determine which unfolded net corresponds to a given polyhedral shape.	
	Threeview	40	Three-view Drawing test the model’s ability to match a 3D shape with its correct top, front, or side view.	
	Sectional View	35	Sectional view require models to infer possible cross-sections of a given 3D object.	
	Spatial Quantities& Transformations	10	Spatial quantities and transformations assess how many cubic units compose a shape or how it transforms in space.	
Special	2D Composition	31	2D composition tasks require models to predict the final shape resulting from multiple plane figures merging according to a given rule.	Evaluate VLMs’ understanding of figure relationships and ability to perform logical reasoning across multiple elements.
	Inter-figure Relations	40	Inter-figure relations assess the model’s ability to analyze relationships such as intersections, unions, and differences between figures, requiring logical set operations.	
Alphanumeric	Alpha-Numeric	27	Involves recognizing letter and number sequences that follow specific patterns.	Evaluate VLMs' ability to infer the next element based on explicit rules.
Black&White Block	Black&White Block	32	Evaluates the arrangement, exchange, and transformation of black and white blocks.	Evaluate VLMs' ability to infer logical transformations within binary visual structures.
Miscellaneous	Miscellaneous	34	Combines multiple reasoning rules, requiring the model to apply diverse inference strategies in a single problem.	Evaluate VLMs' ability to handle complex, multi-rule reasoning scenarios.
Mensa	Mensa v1	35	This subset provides benchmark comparison with human intelligence assessment metrics.	Evaluate VLMs' reasoning and problem-solving abilities with increasing difficulty and include logical reasoning, mathematical abilities, and spatial perception.
	Mensa v2	39		
Raven	Raven v1	40	These tasks preserve design for evaluating abstract reasoning capabilities.	
	Raven v2	60		

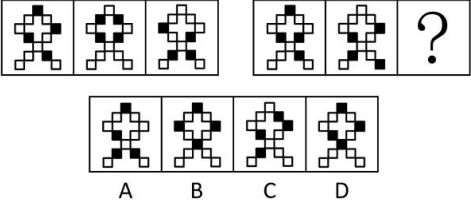
Table 5: The overview of ReasonBench. Different colors represent different Dimensions. Task indicates one or more tasks under each dimension. Num refers to the number of test questions for each task. Introduction provides a brief overview of each task. Purpose describes the utility and impact of the different dimensions we designed.

Task	GPT 4o	Gemini 2.0	Gemini 1.5	Claude 3.5	GLM 4V	Intern VL2	Qwen 7B	Yi Vision
Trans.	29.79	27.66	27.66	25.53	28.72	27.66	21.28	25.53
Rot.	30.36	23.21	33.93	33.93	23.21	23.21	26.79	35.71
Comb.	23.33	30.00	26.67	26.67	20.00	33.33	20.00	26.67
Trav.	27.78	16.67	25.93	35.19	22.22	14.81	31.48	25.93
Add.&Sub.	28.36	37.31	25.37	28.36	22.39	20.90	31.34	31.34
B&W	26.98	28.57	31.75	19.05	31.75	25.40	30.16	19.05
Sym.	27.52	33.03	23.85	31.19	14.68	23.85	24.77	22.02
O&C	36.84	31.58	31.58	42.11	31.58	15.79	26.32	26.32
Comb.	33.33	16.67	0.00	16.67	33.33	0.00	0.00	16.67
Line	24.28	24.86	23.12	19.65	25.43	23.12	26.01	26.01
Surf.	22.63	26.28	30.66	28.47	26.28	24.82	27.01	26.28
Point	27.27	21.21	24.24	24.24	22.73	30.30	28.79	21.21
Elem.	23.40	27.66	30.85	22.34	28.72	25.53	22.34	25.53
Comb.	28.00	22.00	30.00	18.00	26.00	32.00	18.00	20.00
Cube	26.61	19.27	24.77	22.02	31.19	27.52	28.44	21.10
3D Mos.	23.91	28.26	26.09	21.74	28.26	26.09	26.09	23.91
Poly.	23.53	52.94	17.65	5.88	17.65	17.65	17.65	23.53
3-View	15.00	20.00	20.00	22.50	25.00	27.50	20.00	40.00
Sec.View	22.86	37.14	20.00	31.43	17.14	28.57	25.71	48.57
Q&T	30.00	60.00	20.00	30.00	30.00	10.00	30.00	50.00
2D Comp.	29.03	25.81	19.35	35.48	22.58	22.58	16.13	29.03
InterFig.	20.00	20.00	22.50	17.50	7.50	30.00	27.50	22.50
AlphaNum.	29.63	40.74	33.33	22.22	29.63	33.33	29.63	40.74
B&W	28.13	18.75	28.13	18.75	18.75	31.25	25.00	21.88
Misc.	32.35	29.41	26.47	38.24	38.24	29.41	23.53	29.41
Mensa-v1	20.00	25.71	17.14	17.14	/	20.00	20.00	14.29
Mensa-v2	17.95	20.51	25.64	10.26	/	10.26	12.82	15.38
Raven-v1	30.00	32.50	22.50	12.50	/	10.00	22.50	5.00
Raven-v2	38.33	30.00	30.00	16.67	/	13.33	13.33	13.33
All Avg.	26.29	26.91	26.16	24.12	24.88	24.05	24.55	24.86

Table 6: The evaluation results in separated format. We omit the percentage sign(%). Each Task corresponds to the one in Table 2

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please



GPT-4o

To identify the pattern, let's analyze the sequence step by step:

1. In each frame, the blackened squares move one step clockwise.
2. Notice that each shape consists of small square sections, and the black pattern shifts consistently in each step.

Given this pattern, the sequence for the black squares is: Top, Right, Bottom.

Next should be the left side of the shape. **Therefore, the shape that fits this pattern from the options is:**

<ans>optionA</ans>

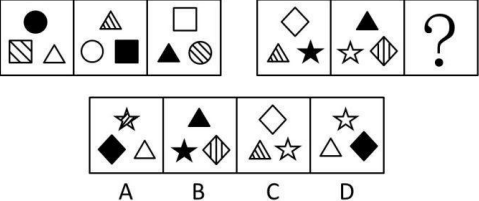
Human annotation

The elements are composed identically, but there is no obvious positional pattern, so adjacent comparisons are considered. Observing the first group of figures, it is found that Figure 1 and Figure 2 share one black block in the same position, and Figure 2 and Figure 3 also share one black block in the same position. Applying this rule to the second group of figures, Figure 1 and Figure 2 share one black block in the same position, so Figure 2 and the "?" should also share one black block in the same position. Only option B fits this pattern. **Therefore, the correct answer is B.**

Figure 10: Example of a Translation task in the Positional dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please



GPT-4o

1. Each box in the sequence follows a pattern of progression based on shape and shading:
 - The sequence involves a specific transformation in shape types and/or shading styles across the rows.
 - The logic in the transitions between the elements of the rows indicates that the missing element needs to conform to the sequence pattern.
2. Comparing the options (A, B, C, and D) against the observed logic, **<ans>optionC</ans>** fits as the correct shape and shading pattern that completes the sequence.

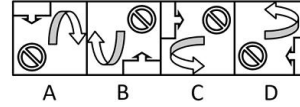
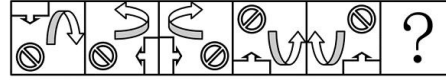
Human annotation

The elements are the same, but the positions of the elements are rotated. In each group of shapes, the position of the shape itself is rotated, while the shadows of different colors also rotate in the opposite direction. In the second group, the shape rotates clockwise, while the shadow rotates counterclockwise. **Therefore, the correct answer is A.**

Figure 11: Example of a Rotation task in the Positional dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

The correct answer is `<ans>optionB</ans>`. Here's the reasoning:

1. **Pattern Analysis**:

- Each panel alternates between incorporating specific symbols (e.g., arrows, "no entry" signs, and lines).
- The direction of arrows and their interaction with other shapes progress systematically through each step.

2. **Matching the Sequence**:

- Option B aligns with the continuation of the alternating pattern in the given series.

Human annotation

The elements are the same, but the position pattern should be prioritized. Upon observation, it is found that:

- Figure 1 becomes Figure 2 after a horizontal flip and clockwise rotation.
- Figure 2 becomes Figure 3 after a horizontal flip.
- Figure 3 becomes Figure 4 after a horizontal flip and clockwise rotation.
- Figure 4 becomes Figure 5 after a horizontal flip.

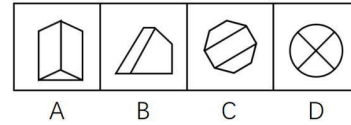
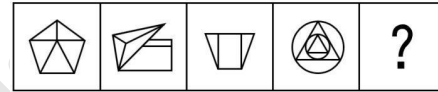
Therefore, the figure at the ? position should be the result of Figure 5 after a horizontal flip and clockwise rotation, and only option C follows this pattern.

Thus, the correct answer is C.

Figure 12: Example of a Combination task in the Positional dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

The correct answer is `<ans>optionD</ans>`.

Explanation: The progression in the given sequence follows a transition in geometric structures and their relationships. The first three figures show transformations of shapes with an increasing focus on symmetry and the integration of simpler geometric elements. The fourth figure highlights the combination of a triangle and circle, leading to a fully symmetrical circular design with intersecting lines, which corresponds to option D.

Human annotation

The elements are different, and the four figures in the question consist of multiple faces. The pattern of the number of faces should be prioritized. Upon observation, the number of faces is 5, 5, 3, and 10, which doesn't show a clear pattern.

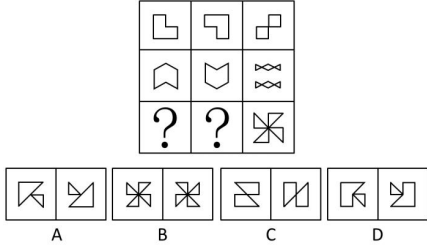
However, further observation reveals that each figure contains triangles, and among the options, only option A contains a triangle.

Thus, the correct answer is A.

Figure 13: Example of a Traversal task in the Stylistic dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please



GPT-4o

The correct answer is <ans>optionB</ans>.

Explanation:

- In each column of the grid, the shapes progressively transform by following a visual pattern.
- In the third column, for example, the top shape becomes simplified in the second row and further evolves in the third row.
- Similarly, the empty space and rotation follow consistent logic in the grid. Among the given options, **B** is the one that best fits the pattern in the context of the grid.

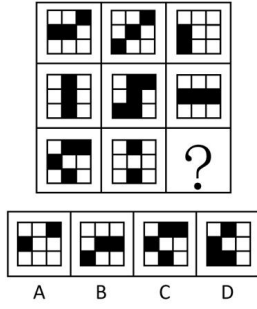
Human annotation

The elements of the figures are similar, so the pattern of the style should be prioritized. Upon observation, the first two figures in each row of the grid follow the "keep the same, discard the different" pattern to form the third figure. The figures in the third row should also follow this rule. Among the options, only option C forms a windmill pattern by keeping the same and discarding the different. Thus, the correct answer is C.

Figure 14: Examples of Additive & Subtractive tasks in the Stylistic dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please



GPT-4o

The pattern follows a rotation pattern over the rows. Observing the image:

1. Row 1 shows the pattern rotating 90 degrees clockwise each step.
2. Row 2 follows the same rotation logic.
3. Therefore, in Row 3, the first shape needs to rotate 90 degrees clockwise to lead to the second, and the second shape to the third.

Following this logic, the second shape in Row 3 rotated 90 degrees clockwise will match Option B.

So, the answer is <ans>optionB</ans>.

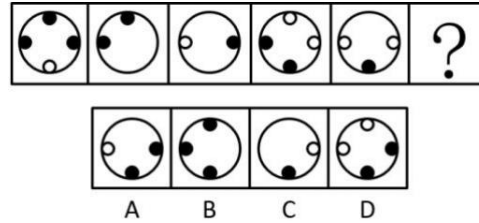
Human annotation

The elements are similar, so the pattern of the style should be prioritized. The shapes in the question have the same outline and divided areas, but the number of black blocks differs, so the black-and-white operation should be considered. When looking horizontally at the grid, no clear pattern emerges, so we examine it vertically. In the first column, the operation rule is: black + black = white, black + white = black, white + black = black, white + white = white. This rule is verified in the second column. Applying this rule to the third column, only option B fits. Thus, the correct answer is B.

Figure 15: Example of Black-White Operations tasks in the Stylistic dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

To determine the correct pattern for the sequence of shapes:

1. The circles follow a pattern regarding the arrangement of the shaded (dark) and unshaded (white) segments.
 2. Observing the sequence, the shaded and unshaded segments rotate clockwise in each new frame.
 3. This rotation alternates between moving the shaded spots one position clockwise each time.
- Following this pattern, the next circle in the sequence (where the question mark is) should have the configuration of shaded and unshaded segments as shown in **<ans>optionC</ans>**.

Human annotation

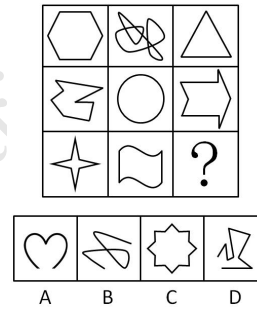
The elements are different, so the pattern of attributes should be prioritized. Upon observation, all the figures in the question are symmetrical and axis-symmetric, with the axis of symmetry rotating counterclockwise by 45° sequentially. Based on this pattern, only option D fits.

Thus, the correct answer is D.

Figure 16: Example of Symmetry tasks in the Attribute dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

To determine the pattern in the grid, we can observe the shapes in each row. Here's the breakdown:

- First row: The shapes transition from a hexagon to a scribble to a triangle. The sequence involves a more structured polygon to less structured or open forms.
- Second row: The sequence goes from an angular to a rounded shape, and then an angular shape.
- Third row: It starts with a star-like shape and a wavy form.

To follow this emerging pattern logic in the third row, the sequence seems to be alternating between angular to rounded/wavy forms.

The best match from the options provided is **<ans>optionC</ans>** (the star-like shape with more segments), as it fits the need for more angular features after the wavy form.

Human annotation

Given that all the figures are closed shapes, upon analysis, only the shape in option C is a closed figure.

Thus, the correct answer is C.

Figure 17: Example of Open&Closed States tasks in the Attribute dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

GPT-4o

To determine the correct answer, let's analyze the pattern:

- Shapes**: Each box has a main shape: circle, square, heart, hexagon, crescent.
- Arrows**: Each shape contains an arrow. The direction of these arrows changes in a clockwise or counter-clockwise manner.
 - The arrow in each shape moves from pointing horizontally, to diagonally, to vertically, etc.

Observing the arrows: - Circle: Horizontal arrow- Square: Diagonal arrow- Heart: Vertical arrow- Hexagon: Diagonal arrow (pointing opposite direction to the square)- Crescent: Horizontal arrow (pointing opposite direction to the circle). The movement follows a sequence of rotational direction changes.

Next in the sequence: - The pattern alternates directions for each type, continuing with a diagonal arrow (in a new shape). The option that fits: **<ans>optionC</ans>** shows a diagonal arrow inside a triangle, continuing the sequence pattern.

Human annotation

Given that all the figures are closed shapes, upon analysis, only the shape in option C is a closed figure. Thus, the correct answer is C.

Figure 18: Example of Combination tasks in the Attribute dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

GPT-4o

The correct answer is ****D****.

Reasoning:

- Shape Progression**: The series adds a side to the polygon in each step: starting with a triangle (3 sides), then a square (4 sides), a pentagon (5 sides), and finally a hexagon (6 sides).
- Symbols Around Shapes**: Each shape is progressively surrounded by symbols. The triangle has three "D"s, the square has four triangles, the pentagon has five squares, and the hexagon has six circles.
- Pattern Completion**: The missing shape should follow the sequence logically:
 - A circle (2D shape, filled) should be surrounded by squares, as squares follow the same progression.

Thus, **<ans>optionD</ans>** fits the given logic.

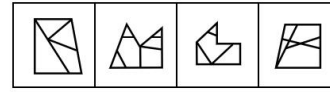
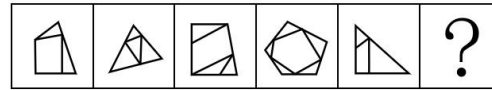
Human annotation

The shaded shapes in the question are a triangle, quadrilateral, pentagon, and hexagon, with the number of sides increasing by one. Therefore, the shape at the ? position should have two sides, which eliminates options C and D. Upon comparing options A and B, we observe that the small shapes in the top-left and bottom-left corners of the shaded area have different numbers of sides. From the question, it is clear that the number of sides of the small shapes increases sequentially: 2, 3, 4, and 5. Therefore, the small shape in the ? position should have one side, which is satisfied by option A. Thus, the correct answer is A.

Figure 19: Example of the Lines task in the Quantitative dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

To solve this pattern problem, let's analyze the sequence of shapes:

1. The shapes progress from having more sides and showing more complexity.
2. Each shape is a polygon with internal lines dividing it into triangles or other smaller polygons.

Let's match the level of complexity and style:

The correct option to fit this sequence based on the progression is <ans>optionA</ans>.

Human annotation

The elements are different, and there is no obvious pattern in attributes, so the number pattern should be considered. The shapes in the question are segmented, and the enclosed areas are clearly visible, so the number of faces should be prioritized. The total number of faces is 3, 4, 4, 6, and 3, which shows no clear pattern.

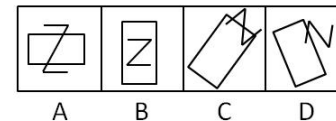
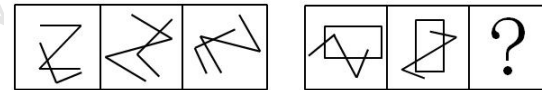
Upon further analysis, it is observed that each shape in the question has a prominent largest face, and the number of edges of the largest face matches the number of edges of the outer frame. Only option C fits this condition.

Thus, the correct answer is C.

Figure 20: Example of the Surface task in the Quantitative dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

The correct answer is <ans>optionB</ans>.

Reasoning:

1. In the top sequence, each figure alternates between a zigzag pattern and the addition of a geometric rectangle or boundary around it.
2. The first figure in the second sequence has a zigzag line with a rectangle added to it.
3. The next figure continues the pattern by alternating back to a plain zigzag line.
4. Therefore, the final figure (question mark) should return to the pattern of a zigzag line with a rectangle added around it, matching <ans>optionB</ans>.

Human annotation

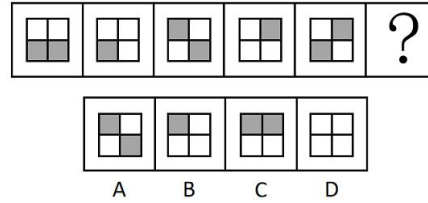
In the first group of shapes, each element intersects, and the number of intersection points is the same, which is 2. The second group of shapes should follow the same pattern, meaning each shape's elements should have the same number of intersection points. Observing the first two shapes in the second group, the elements intersect, and the number of intersection points is 3. Therefore, the third shape should also have 3 intersection points, and only option C satisfies this condition.

Thus, the correct answer is C.

Figure 21: Example of the Points task in the Quantitative dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

To solve this problem, let's observe the given pattern:

1. The sequence alternates between two positions of the shaded square:

- First square: bottom left
- Second square: bottom right
- Third square: top left
- Fourth square: bottom left

Based on this alternating sequence, the fifth position (where the question mark is) should follow this pattern.

Hence, the correct answer is `<ans>optionB</ans>`, where the shaded square is in the top right position.

Human annotation

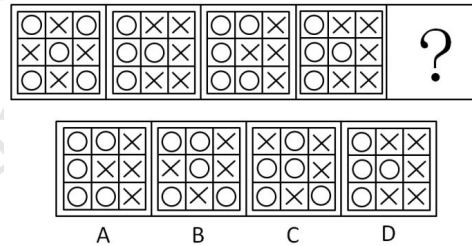
The shapes in the question show no obvious pattern in position or attributes, so the number pattern should be considered. Upon observation, the shapes are made up of gray and white small squares. The number of gray squares follows the pattern: 2, 1, 2, 1, 2. Therefore, the shape at the ? position should have 1 gray square, which is satisfied by option B.

Thus, the correct answer is B.

Figure 22: Example of the Elements task in the Quantitative dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

The image presents a logical sequence puzzle based on visual patterns. After carefully analyzing the grid patterns for changes, the correct answer is `<ans>optionC</ans>`.

Explanation:

The sequence shows a consistent progression of swapping or replacing elements (X or O) in specific rows or columns across the grids. Option C aligns perfectly with the continuation of this pattern.

Human annotation

The elements are the same, but their quantities differ, so the number pattern should be considered. Upon observation, the number of circles in the shapes are 5, 4, 5, and 4. Therefore, the fifth shape should contain 5 circles, which eliminates option D.

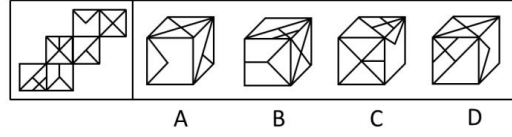
Options A, B, and C differ only in the position of the circles, but there is no movement pattern. From the overall observation of the question, all the shapes are symmetrical. Thus, we should choose a symmetrical shape, and only option A satisfies this condition.

Thus, the correct answer is A.

Figure 23: Example of the Combination task in the Quantitative dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

The sequence of cubes appears to follow a pattern where each successive cube is a rotation or transformation of the previous one. The lines and shapes inside the cube share a consistent arrangement that slightly alters between each step.

Looking at the options:

The transformation seems to be a rotation or rearrangement of the internal lines, consistent with the sequence provided. The correct answer should maintain this pattern of alteration.

<ans>optionC</ans> fits this transformation pattern, continuing the sequence accurately.

Human annotation

By labeling each face as shown in the diagram:

****Option A**:** The common points of faces d, e, and f emit 2 lines in the unfolded diagram, but only 1 line is emitted in the option, so it does not match the unfolded diagram. ****Eliminate****.

****Option B**:** In face b, the Y-shaped opening faces the unique edge, and when drawn clockwise, edge 2 is next to face a in the unfolded diagram, but in the option, edge 4 is next to face a. This does not match the unfolded diagram. ****Eliminate****.

****Option C**:** The common points of faces a, c, and e emit 2 lines in the unfolded diagram, but 3 lines are emitted in the option, so it does not match the unfolded diagram. ****Eliminate****.

****Option D**:** The common points of faces a, e, and f match the unfolded diagram, so it is correct.

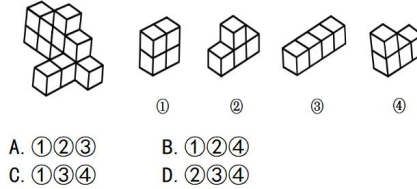
Thus, the correct answer is D.

Figure 24: Example of the Cube task in the Spatial dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

The 3D shape shown in the left image can be composed of the following three shapes: ()

**GPT-4o**

The correct answer is ****A (①②③)****.

Explanation:

1. The goal is to identify which three shapes from the given options (①, ②, ③, ④) can combine to form the target shape on the left.

2. Upon analysis:

- Shape ① contributes the $2 \times 2 \times 1$ cuboid in the lower left.
- Shape ② provides the L-shaped structure located on the upper right.
- Shape ③ completes the structure by forming the $2 \times 1 \times 1$ rectangle on the lower front side.

3. Shape ④ cannot be part of the solution because its linear structure does not fit anywhere in the target figure.

Thus, the three shapes required are ****①, ②, and ③****, making **<ans>optionA</ans> the correct answer.**

Human annotation

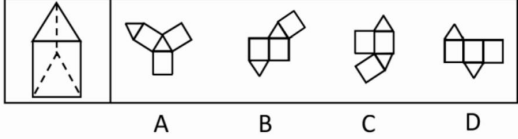
This question tests the concept of solid shape assembly. The 3D shape can be formed by combining shapes ①, ②, and ④. **Thus, the correct answer is B.**

Figure 25: Example of the 3D assembly task in the Spatial dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

Given a paper box, which of the following options cannot be folded into the given paper box?



GPT-4o

The correct answer is ****A****.

Explanation:

The given net of the paper box (in the leftmost part of the image) can fold into the shape of a house with a triangular roof. To determine which of the options (A, B, C, D) cannot form this shape:

- ****Option A****: The net in this option cannot form the given paper box because the triangular sections are arranged incorrectly, making it impossible to fold into the triangular roof of the house.
- ****Options B, C, and D****: These nets can all fold correctly into the given paper box shape.

Thus, the answer is **<ans>optionA</ans>**.

Human annotation

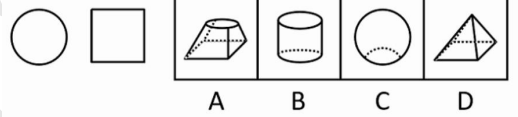
This question tests the concept of paper folding into a box. The rectangle in the lower-left corner of C is incorrectly positioned; it should be placed on the right side. Options A, B, and D can all be folded into the given box. Thus, the correct answer is C.

Figure 26: Example of the Polyhedral task in the Spatial dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

The image shows two different views of an object. What is the most likely shape of the object?



GPT-4o

The given shapes are a circle and a square, which suggests a three-dimensional object that appears as a circle from one view and as a square from another.

The most likely 3D object that fits these descriptions is a cylinder, as a cylinder presents a circular face from a top view and a rectangular shape from a side view.

So, the correct option is **<ans>optionB</ans>**.

Human annotation

This question tests the concept of orthographic projection (three-view drawing). Let's analyze the options one by one:

- ****Option A****: From the bottom view, a square can be seen, but the circle cannot be seen, so it is eliminated.
- ****Option B****: From the top view, a circle is visible; from the front view, a square is visible. This matches the conditions, so it is correct.
- ****Option C****: From the top view, a circle is visible, but the square cannot be seen, so it is eliminated.
- ****Option D****: From the bottom view, a square is visible, but the circle cannot be seen, so it is eliminated.

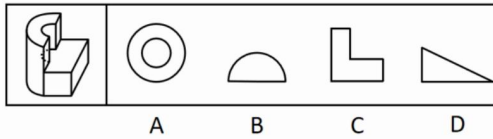
Thus, the correct answer is B.

Figure 27: Example of the Three-view task in the Spatial dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

The left image shows the given 3D shape. Which of the following could be the cross-section obtained by slicing the shape along any of its faces?

**GPT-4o**

The problem requires determining the correct cross-sectional shape of the given 3D object when it is cut from any plane. After careful analysis of the 3D object and the four options (A, B, C, D), the correct answer is:

****A****.

This is because the given 3D object contains a circular cross-section, which matches <ans>optionA</ans>. Other shapes (B, C, and D) do not align with possible cross-sections of this 3D figure.

Human annotation

This question tests the concept of cross-sectional views. Let's analyze the options one by one:

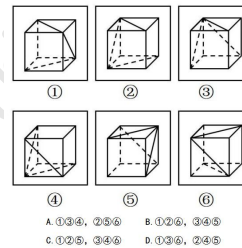
- ****Option A****: It is not possible to obtain a complete ring through the section, so it is eliminated.
- ****Option B****: It is not possible to obtain a semicircle through the section, so it is eliminated.
- ****Option C****: As shown in the diagram, a proper cross-section can be obtained. This matches the conditions, so it is correct.

Thus, the correct answer is C.

Figure 28: Example of the Sectional view task in the Spatial dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

The correct answer is ****A. ①③④****.

Explanation:

The task involves grouping the cube diagrams based on their internal line patterns and geometric arrangements. Observing carefully:

1. In ****Group ①③④****, the internal lines represent a consistent triangular connection or a similar structure formed inside the cube. These cubes share the same visual logic in their internal configuration.
2. In ****Group ②⑤⑥****, the internal lines and connections form a distinctively different pattern. These cubes align with a different visual rule.

Thus, the graphic reasoning rule identifies two separate groups, and the correct option is <ans>optionA</ans>.

Human annotation

This is a grouping and classification problem. The six shapes in the question are all composed of a regular hexahedron and three diagonals within each face. Let's analyze the relationship between the three diagonals on each face.

- ****Analysis****: For the faces in diagrams ①, ②, and ⑥, the faces with the three diagonals represent ****opposite faces**** of the hexahedron. For the faces in diagrams ③, ④, and ⑤, the faces with the three diagonals represent ****adjacent faces**** of the hexahedron, with each pair of faces being adjacent to each other.

Thus, based on this analysis, the correct grouping is ****B****.

Figure 29: Example of the Spatial Quantities&Transformations task in the Spatial dimension.

Among the four figures in the options, only one is made up of the four figures in the question (only by shifting up, down, left or right). Please find it out.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

GPT-4o

The correct answer is **D**.

Reasoning:

- The task is to identify which one of the options (A, B, C, D) can be formed by combining the four shapes given in the top row. The combination can only involve translation, rotation, or flipping of the shapes.
- Careful analysis shows: **Option A**: The outline cannot be formed by the given shapes, as some parts are too distorted compared to the top shapes. **Option B**: Similarly, the shape cannot be formed due to mismatched edges. **Option C**: While closer, it still fails to match perfectly when attempting to combine the shapes through allowable transformations. **Option D**: This shape can be perfectly formed by arranging, rotating, or flipping the four shapes given in the top row. Thus, the answer is **optionD**.

Human annotation

This question tests the plane splicing. For curve-like figures, eliminating the lines with the same curvature and equal length and then splicing them together will give a contour diagram, which is option D. Therefore, the correct answer is **optionD**.

Figure 30: Example of the 2D Composition task in the Special dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

GPT-4o

The correct answer to fill in the question mark is **optionD**.

Reasoning:

- Each graphic in the sequence undergoes a specific transformation step by step. These transformations could include flipping, rotating, or adding/removing elements.
- Observing the sequence in the top row:
 - The changes from one graphic to the next appear systematic, involving geometric alterations and shifts in orientation or design.
 - When applying the observed transformations logically to the options in the bottom row, only **optionD** fits as a continuation of the pattern.

Human annotation

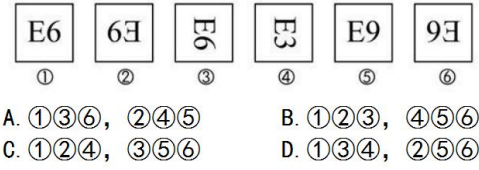
The elements in the figures are disorganized, but they are all composed of two shapes. Let's analyze the relationship between the two shapes.

- Observation**: The pattern follows an odd-even sequence. The odd-numbered figures (1, 3, 5) have separate shapes, while the even-numbered figures (2, 4) have intersecting shapes.
- Conclusion**: Based on the pattern, the figure at the "?" position should have two shapes that intersect, as it follows the pattern of the even-numbered figures. Thus, the correct answer is **optionC**.

Figure 31: Example of the Inter-Figure task in the Special dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

The correct answer is <ans>optionD</ans> ①③④, ②⑤⑥.

Reasoning:

The images are divided into two groups based on the orientation and symmetry of the characters:

- Group ①③④: The characters maintain their correct orientation and are readable (e.g., "E6", "E6", "E3").
- Group ②⑤⑥: The characters are flipped or inverted in some way (e.g., "6E", "E9", "9E").

Human annotation

The figures in the question feature open letters and numbers. The key to solving this is to analyze the direction of the openings in the letters and numbers.

- **Observation**:

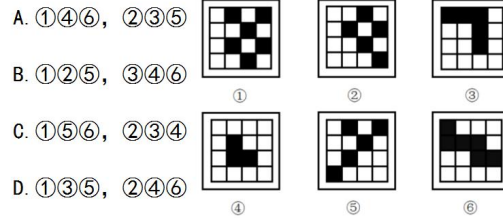
- In figures ①, ③, and ⑥, the direction of the openings in both the letters and numbers is consistent.
- In figures ②, ④, and ⑤, the direction of the openings in the letters and numbers are opposite.

Therefore, the correct answer is **A**.

Figure 32: Example of the Alphanumeric dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please

**GPT-4o**

The correct answer is **D. 135, 246**.

Explanation:

The images follow a pattern where the black cells are rearranged in a specific manner between groups. Analyzing each image group reveals that this sequence matches the arrangement provided by <ans>optionD</ans>.

Human annotation

The figures in the question consist of black-and-white block shapes. There is no obvious pattern when considering the total number of black blocks, so we need to observe the arrangement more closely.

- **Observation**:

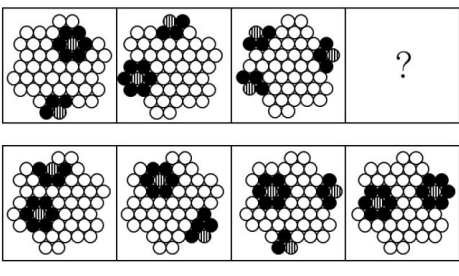
- In figures ①, ②, and ⑤, the black blocks are arranged in a more scattered or irregular manner, where the blocks connect at points.
- In figures ③, ④, and ⑥, the black blocks are arranged in a more neat and compact way, where the blocks connect along their edges.

Thus, the correct answer is **B**.

Figure 33: Example of the Black-White Block dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the four options ABCD, and make sure to fill in the question mark. Please



GPT-4o

The correct answer is ****D****.
 Explanation: The patterns follow a rotational movement of the striped and solid black circles around the central grid. The sequence demonstrates a consistent clockwise shift of the highlighted elements, and **<ans>optionD</ans>** aligns with this rule to fill in the question mark accurately.

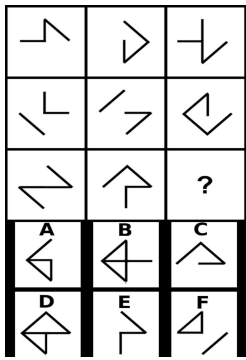
Human annotation

Let's analyze the figures based on the given elements: black circles, bar-shaped circles, and white circles.
 Observation:
 Each bar-shaped circle in the given figures is only adjacent to black circles, not white circles, so we can eliminate A and C.
 Each black circle in the figures is adjacent to a bar-shaped circle, but in B, there is one black circle that is not adjacent to a bar-shaped circle, so we can eliminate B as well.
 This leaves D as the correct option.
 Thus, the correct answer is D.

Figure 34: Example of the Miscellaneous dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the six options ABCDEF, and make sure to fill in the question mark. Please



Human annotation

The non-overlapping lines of the first and second pictures in each row form the third picture. Therefore, the correct answer is A.

GPT-4o

To solve this visual pattern puzzle, let's analyze the sequences in the given grid:

- Row Analysis:** - First Row: Each shape maintains a consistent structure with one additional segment being added from left to right. - Second Row: A similar pattern to the first row and rotated variations of the shapes. - Third Row: Consistency with adding linear segments and slight rotations.
- Column Analysis:** - Each column also follows a pattern of transformation through rotation and segment addition.
- Pattern Deduction:** - The observed pattern includes adding segments and rotating the shapes to match a sequence as seen in both rows and columns. Given these observations, the missing shape should logically follow the rotation and addition patterns seen in its row and column. By inspecting all options (A, B, C, D, E, F), we focus on those that best fit the transformation sequence. Therefore, the correct answer to fill in the question mark is **<ans>optionE</ans>**.

Figure 35: Example of the Mensa v1 task in the Mensa dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the eight options ABCDEFGH, and make sure to fill in the question mark. Please

Human annotation

The small square moves clockwise around the small circle in the middle. So the correct answer is F.

GPT-4o

To solve the puzzle, observe the grid and note the patterns:

1. Each row has a pattern where the position of the black square and the circle follow a consistent movement.
2. Each column seems to progress by shifting or rotating the black squares or circles according to a specific order. Let's analyze: - First row: The black square moves diagonally from top-left to a combination of top-right positions. - Second row: The black square starts moving down vertically, and the circle moves to the right. - Third row: The black square is moving in a diagonal manner, similar to the first row but starting from a different point. Based on these observations, the missing piece in the 3rd row should continue this diagonal pattern with the black square and black circle. Looking at the options, **F** fits this pattern, as it continues the diagonal progression observed in the arrangement. Therefore, the answer is <ans>optionF</ans>.

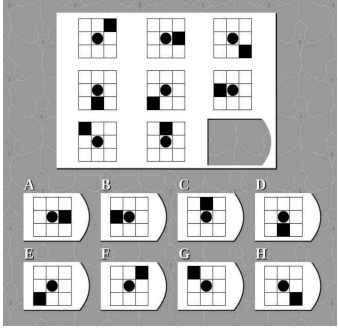


Figure 36: Example of the Mensa v2 task in the Mensa dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the six options ABCDEF, and make sure to fill in the question mark. Please

Human annotation

In the first column, A is the first of 26 letters, C is the third, and F is the sixth. Therefore, the correct answer is F.

GPT-4o

To solve the problem, we need to determine the pattern among the given elements: - Observe the letters: A, C, F. - These letters are the 1st, 3rd, and 6th letters of the English alphabet. Now, observe the numbers associated with each letter: - A corresponds to 1. - C corresponds to 3. Following this pattern, it appears the number correlates to the position of the letter in the alphabet. Thus, F, which is the 6th letter, should correspond to 6. Therefore, the correct answer is <ans>optionF</ans>.




Figure 37: Example of the Raven v1 task in the Raven dimension.

Prompt

You are a person with high IQ and graphic reasoning ability. Please find the pattern of the pattern according to the given picture, give the correct answer directly from the six options ABCDEF, and make sure to fill in the question mark. Please

Human annotation

The appearance is similar to that of the same industry, and each type of figure appears 3 times inside. Therefore, the correct answer is F.

GPT-4o

To solve this, let's identify the pattern in each column: 1. **First Column**: - Top: Circle. - Middle: Triangle shape. - Bottom: Box with a plus inside. 2. **Second Column**: - Top: Triangle shape. - Middle: Plus sign. - Bottom: Box with a circle inside. 3. **Third Column**: - Top: Plus sign. - Middle: Circle. - Bottom: Box with a triangle shape inside. The pattern appears to involve rotating the shapes through each position from top to bottom. Each column shifts the symbol format between the positions. The missing shape must match the pattern from these observations. Following the rotation: - For the unshaded, outlined shape (like the missing one), the contents should rotate from a plus sign in the top to a circle in the middle and a triangle shape in the bottom. - Following this logic in the third row (bottom), the shape should be a box with a triangle shape inside, and the box has a double outline. Based on this, the correct answer that fits the pattern is `<ans>optionF</ans>`.

Figure 38: Example of the Raven v2 task in the Raven dimension.