MV-MATH: Evaluating Multimodal Math Reasoning in Multi-Visual Contexts

Supplementary Material

Table of Contents in Appendix

A. More Details about MV-MATH

In this chapter, we will introduce MV-MATH in more detail.

A.1. Question Distribution

All questions in MV-MATH are presented in English. As shown in Table 1, the longest question in MV-MATH spans 383 words, while the shortest contains 14 words, with an average length of 80.17 words. Figure 7 further illustrates the distribution of text lengths, highlighting the diversity of MV-MATH.

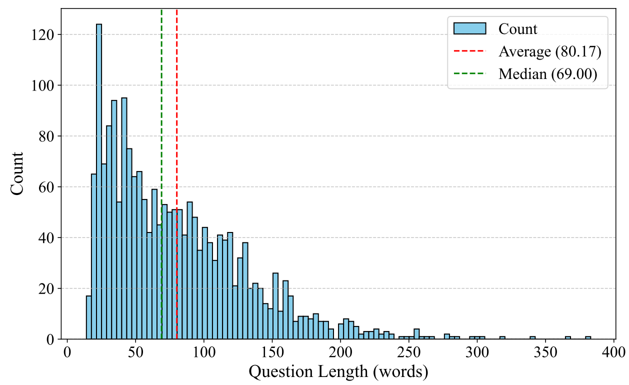


Figure 7. The distribution of the number of words per question in MV-MATH.

A.2. Image Distribution

Figure 8 illustrates the distribution of the number of images associated with each question in the MV-MATH dataset, highlighting its multimodal diversity. The dataset features an average of 3.02 images per question, with a median of 3, demonstrating a balanced and realistic allocation of visual resources. Most questions include 2,3 or 4 images, reflecting the dataset's emphasis on providing sufficient visual context for reasoning tasks. The presence of questions with 5 or more images showcases the dataset's capability to handle complex, multi-visual scenarios. This diversity ensures MV-MATH is suitable for evaluating the ability of models to integrate and reason across multiple interconnected visual elements, further enhancing its value as a benchmark for multimodal mathematical reasoning.

A.3. Division of Difficulty Levels

Since MV-MATH is a meticulously annotated dataset containing both answers and an analysis field, it offers a unique advantage for difficulty classification based on the lengths of the question and analysis. The difficulty classification is conducted in two steps.

Step 1: A weighted average is calculated using the lengths of the *question* and *analysis* fields, assigning a weight of 0.4 to the *question* length and 0.6 to the *analysis*

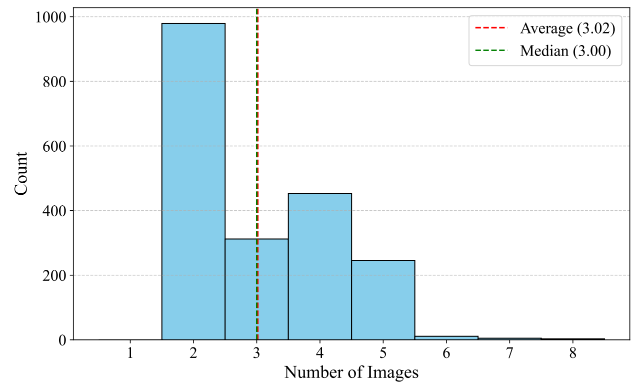


Figure 8. The distribution of the number of images per question in MV-MATH.

length. This weighting reflects the assumption that the solution process (captured in the *analysis* field) is more indicative of a question's difficulty. The weighted length distribution is shown in Figure 8. Using the K-means clustering algorithm, we cluster the weighted lengths into three difficulty levels: easy (0–150), medium (150–500), and hard (> 500).

Step 2: Manual verification is performed on the clustering results from Step 1 by two graduate students. This process adjusts the difficulty classification to account for cases where the weighted length alone may not fully reflect complexity. For instance, a question with a short weighted length but involving intricate formula derivation is reclassified as harder, while a question with a long weighted length but straightforward reasoning is reclassified as easier.

The detailed distribution of weighted lengths can be seen in Figure 9. After the two-step process described above, we classified the questions into 542 as *easy*, 964 as *medium*, and 503 as *hard*.

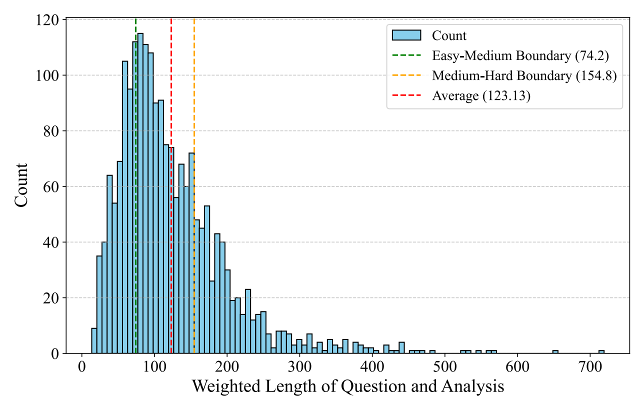


Figure 9. Weighted Length Distribution of Questions and Analysis for Difficulty Classification in MV-MATH

A.4. Image by Subjects

This section presents examples of images from various categories in the MV-MATH dataset. Figures 10 to 20 correspond to images under the categories of *Algebra*, *Arithmetic*, *Combinatorial Geometry*, *Analytic Geometry*, *Combinatorics*, *Descriptive Geometry*, *Logic*, *Metric Geometry*, *Solid Geometry*, *Statistics*, and *Transformation Geometry*, respectively.

The diversity of images across categories reflects the wide range of mathematical concepts covered in MV-MATH. For instance, images in the *Statistics* category primarily include charts, graphs, and tables, providing rich visual representations of statistical data. In contrast, images under *Arithmetic*, often targeted at younger students, feature elements with real-world objects or playful designs to aid comprehension. Similarly, the *Combinatorics* and *Combinatorial Geometry* categories include diagrams with intricate arrangements, requiring detailed reasoning. On the other hand, images in *Analytic Geometry* and *Metric Geometry* are more abstract and geometric, often involving coordinate systems, vectors, or precise measurements.

This variety of visual styles within and across categories highlights the versatility of MV-MATH in evaluating models' ability to interpret and reason over diverse visual contexts, making it a comprehensive benchmark for multi-visual mathematical reasoning tasks.

A.5. Introduction of Subjects

Analytic Geometry. Analytic Geometry integrates algebraic techniques with geometry through the use of a coordinate system. It provides a systematic method to describe geometric shapes using equations and to interpret these equations visually. This branch of mathematics enables a detailed study of geometric properties, such as distance, angles, and tangents.

Algebra. Algebra is a fundamental area of mathematics that explores the use of symbols to represent numbers and quantities in formulas and equations. It encompasses a wide range of topics, from solving linear equations to studying abstract structures such as groups, rings, and fields, playing a crucial role in generalizing mathematical principles and solving problems where specific values are unknown.

Metric Geometry. Metric Geometry focuses on the study of geometric figures based on distances and angles. It examines properties that remain invariant under transformations such as rotations, translations, and reflections. Metric Geometry serves as a foundation for applications in computer graphics, engineering, and physics.

Combinatorics. Combinatorics is the branch of mathematics concerned with counting, arrangement, and combination of objects. It investigates problems related to discrete structures, such as permutations, combinations, and graph theory. Combinatorics is widely used in areas like cryptography, algorithm design, and network theory.

Transformation Geometry. Transformation Geometry studies geometric transformations, such as translations, rotations, reflections, and dilations, to understand how shapes change while preserving certain properties. It provides insights into symmetry, congruence and has significant applications in computer vision.

Logic. Logic in mathematics involves the study of principles of reasoning, including the formulation and analysis of valid arguments. It lays the groundwork for proof techniques and the development of formal systems. Mathematical logic is essential for understanding the structure of mathematical theories.

Solid Geometry. Solid Geometry is the study of three-dimensional figures such as spheres, cubes, cones, and cylinders. It explores their properties, measurements, and spatial relationships, often involving volume, surface area, and intersections.

Arithmetic. Arithmetic is the oldest and most fundamental branch of mathematics, dealing with numbers and basic operations such as addition, subtraction, multiplication, and division. It serves as the foundation for other mathematical disciplines and is essential for everyday problem-solving and quantitative reasoning.

Combinatorial Geometry. Combinatorial Geometry combines concepts from combinatorics and geometry to study the arrangement and interaction of geometric objects. It addresses problems involving configurations of points, lines, and planes, often focusing on optimization and enumeration. This field has applications in graph theory, computational geometry, and optimization.

Descriptive Geometry. Descriptive Geometry is a method for representing three-dimensional objects in two dimensions using projections. It involves techniques to visualize and solve spatial problems through precise drawings. This field is widely used in engineering and architecture for designing and visualizing complex structures.

Statistics. Statistics is the study of data collection, analysis, interpretation, and presentation. It involves mathematical techniques to summarize and infer conclusions from data, often focusing on patterns, variability, and uncertainty. Statistics is indispensable in research, economics, and decision-making processes across various domains.

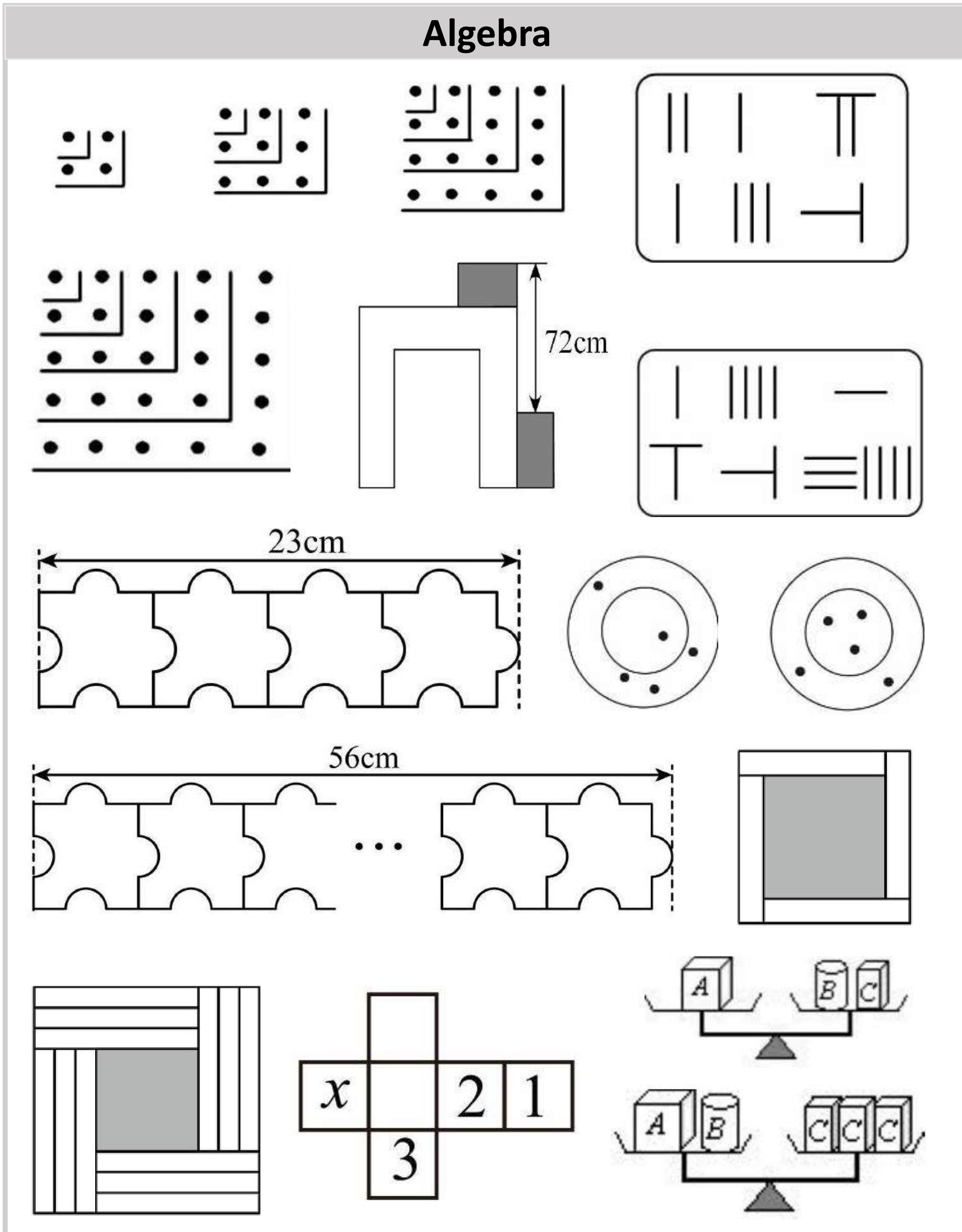


Figure 10. Some images from Analytic Algebra.

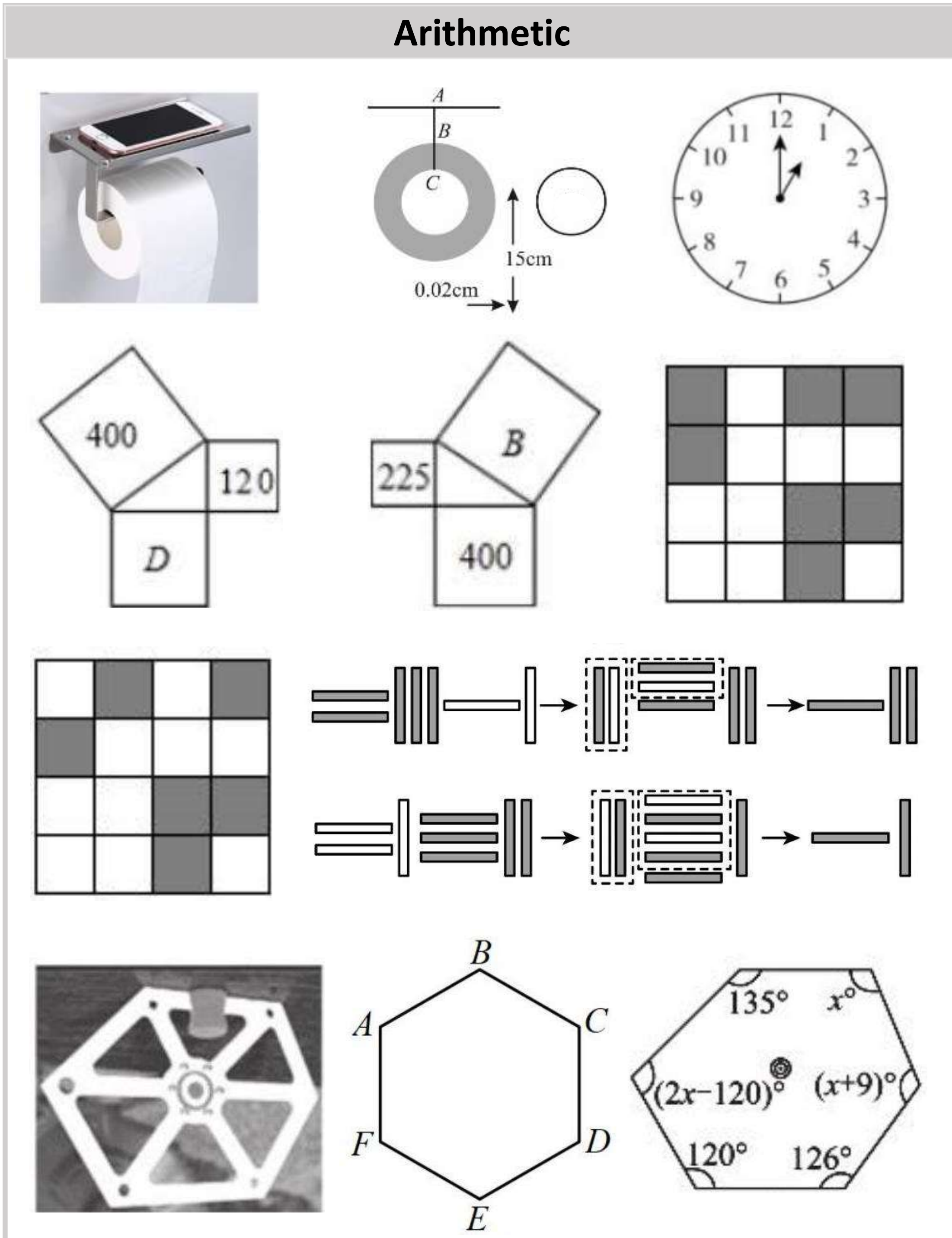


Figure 11. Some images from Arithmetic.

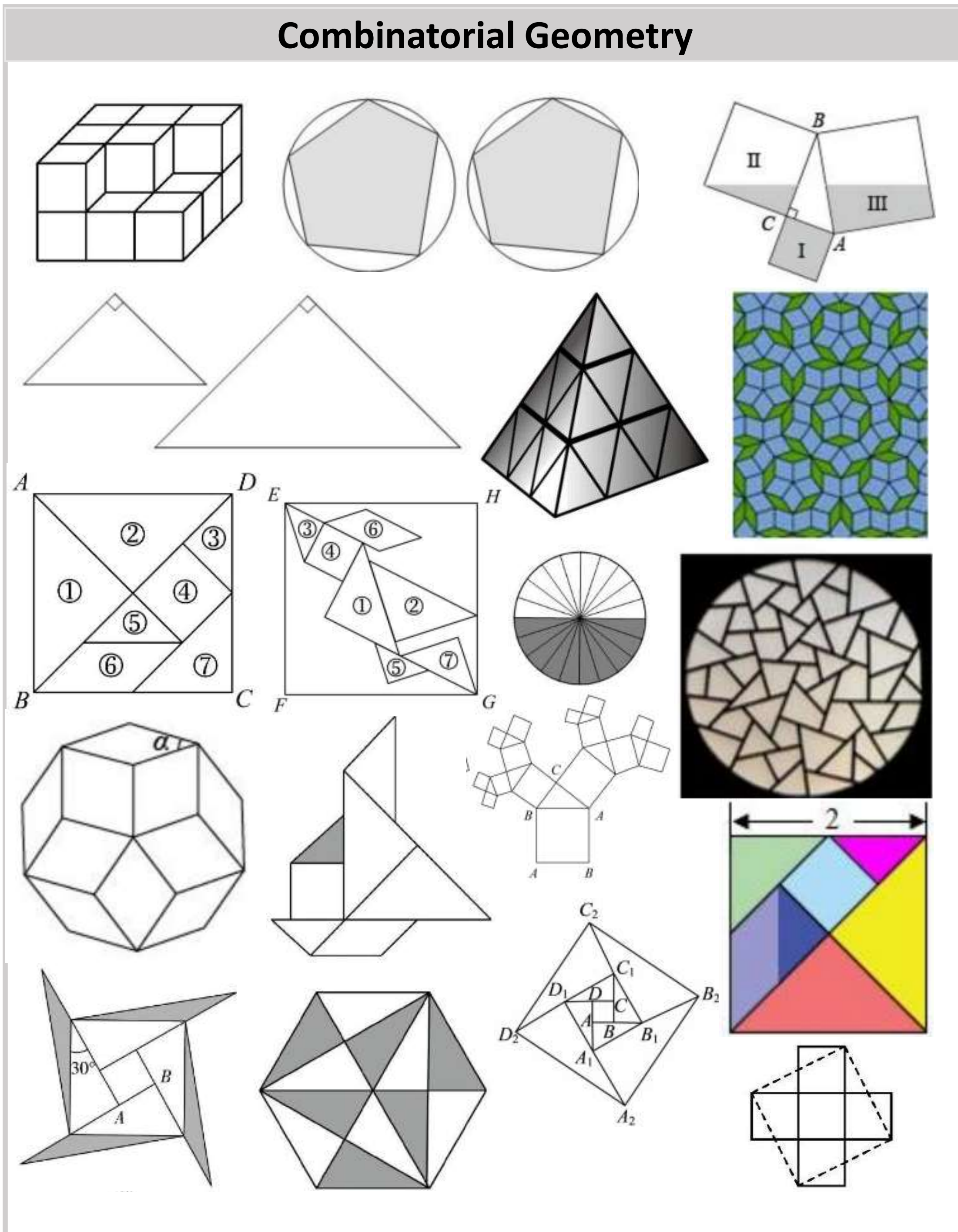


Figure 12. Some images from Combinatorial Geometry.

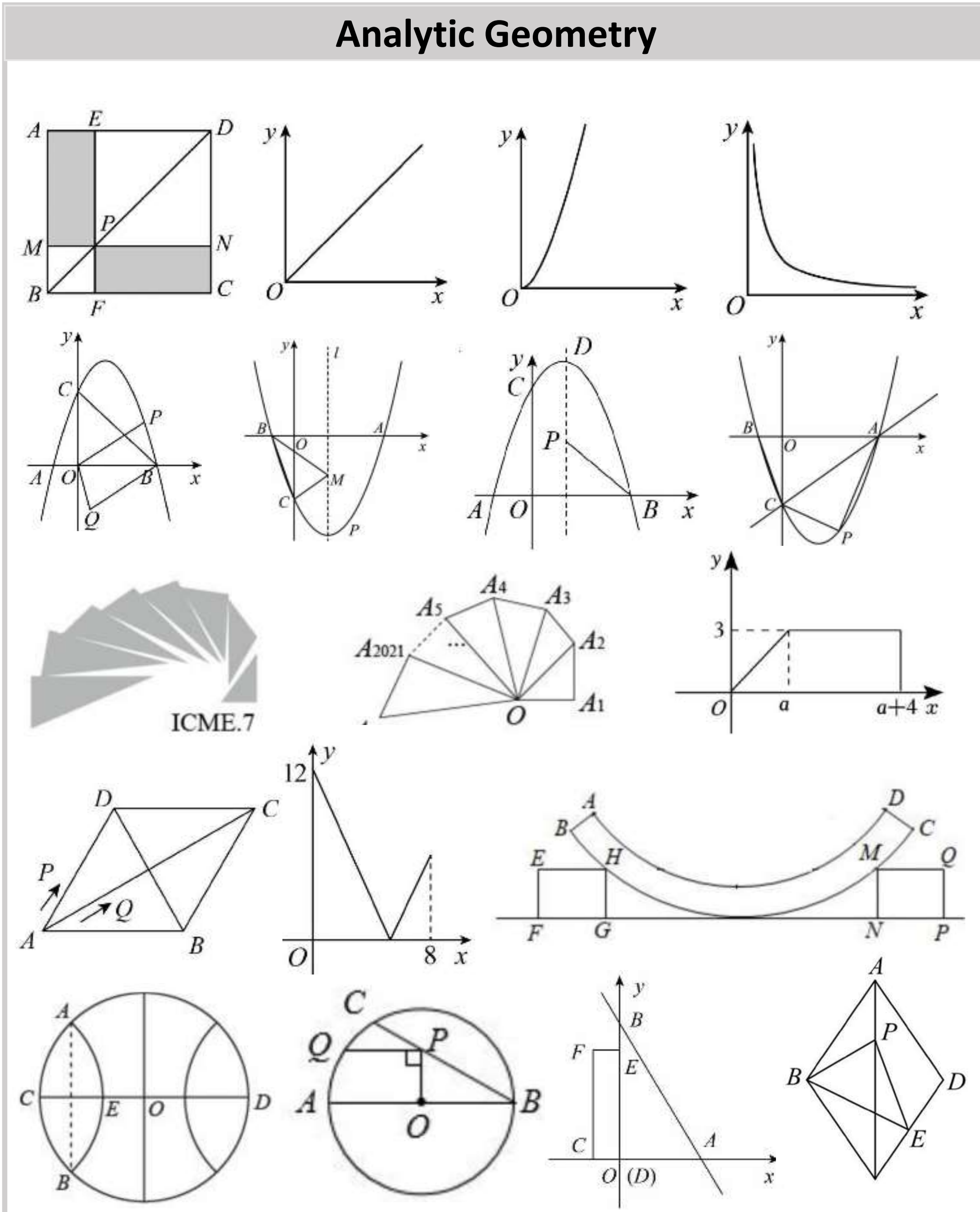


Figure 13. Some images from Analytic Geometry.

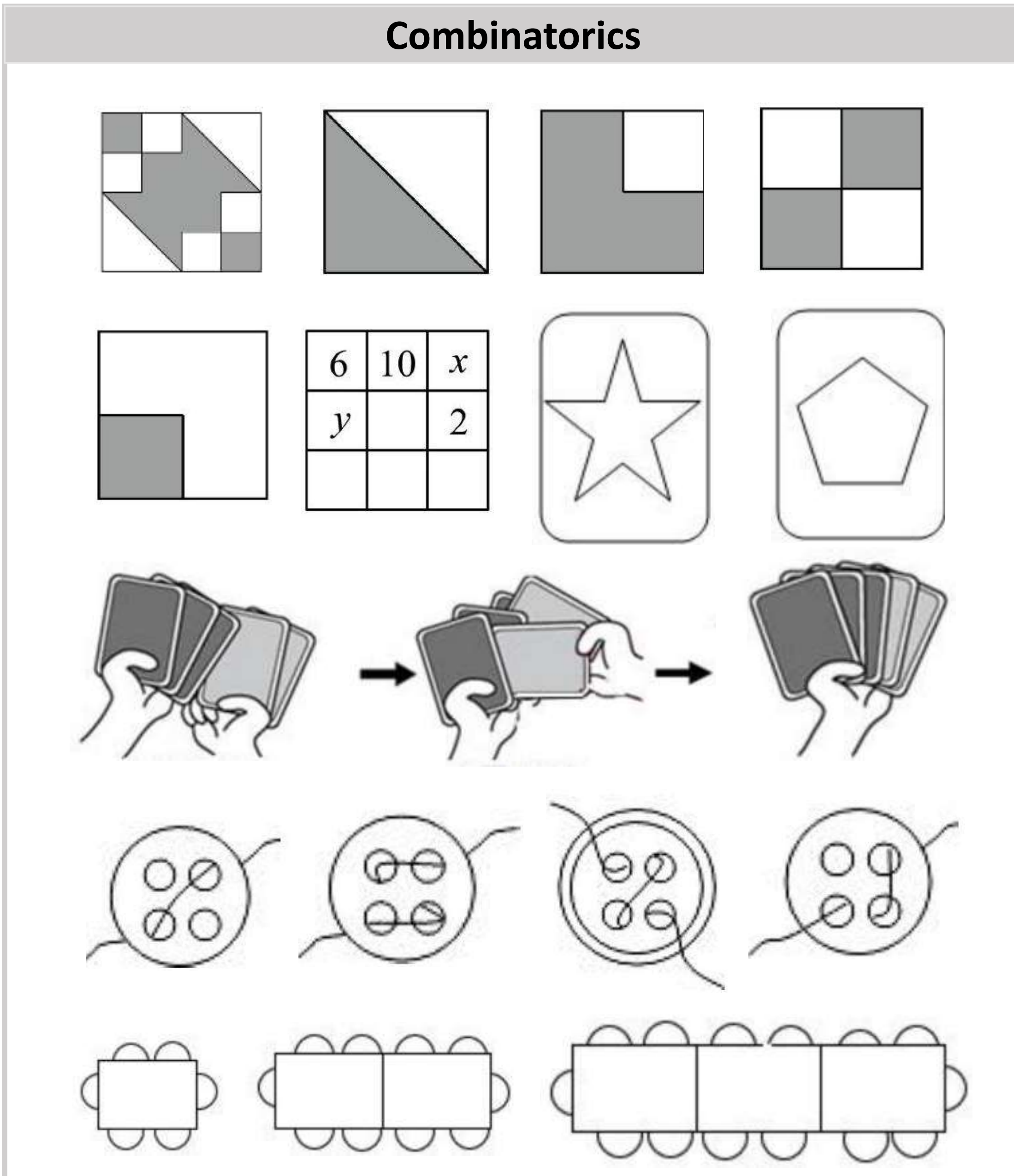


Figure 14. Some images from Combinatorics.

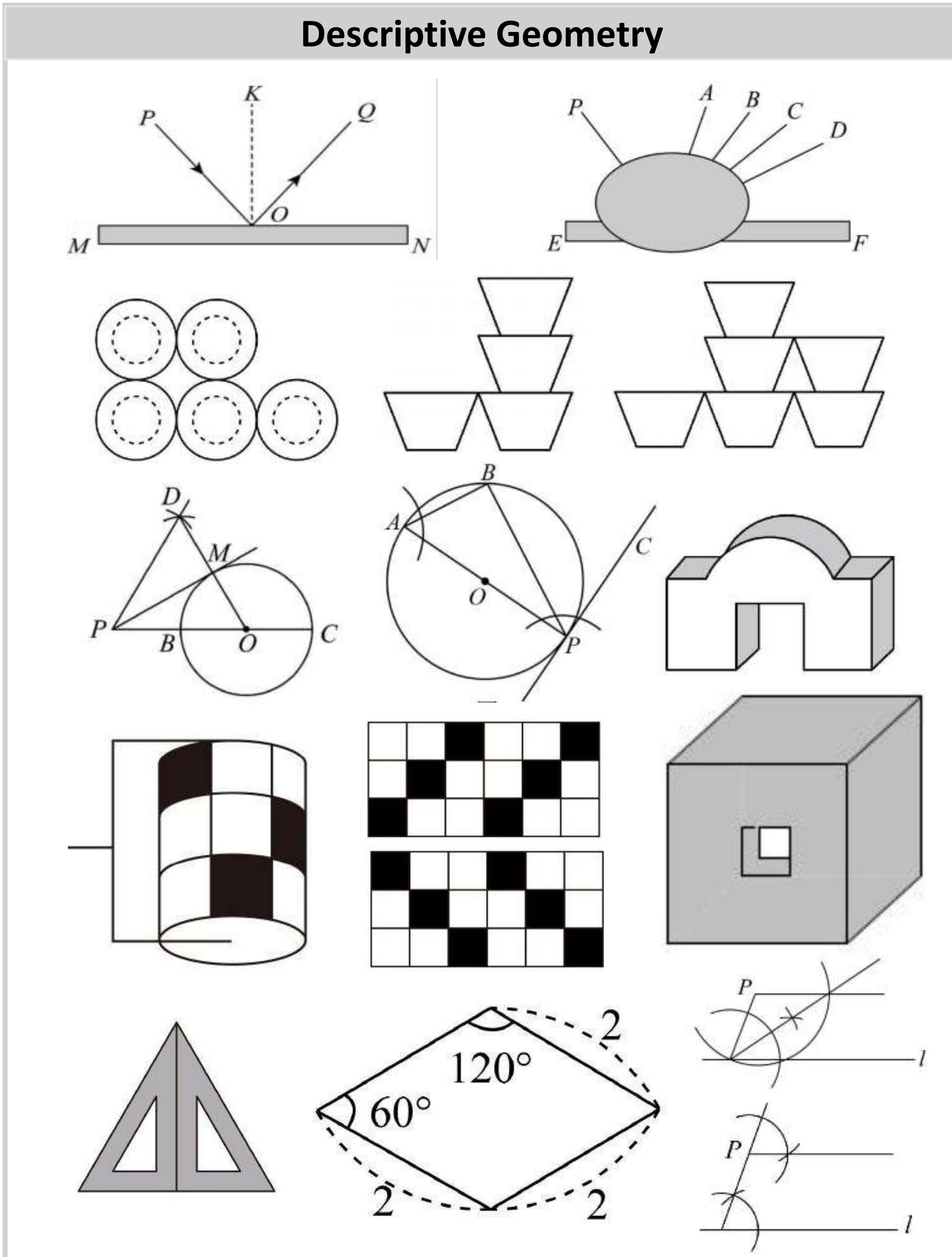


Figure 15. Some images from Descriptive Geometry.

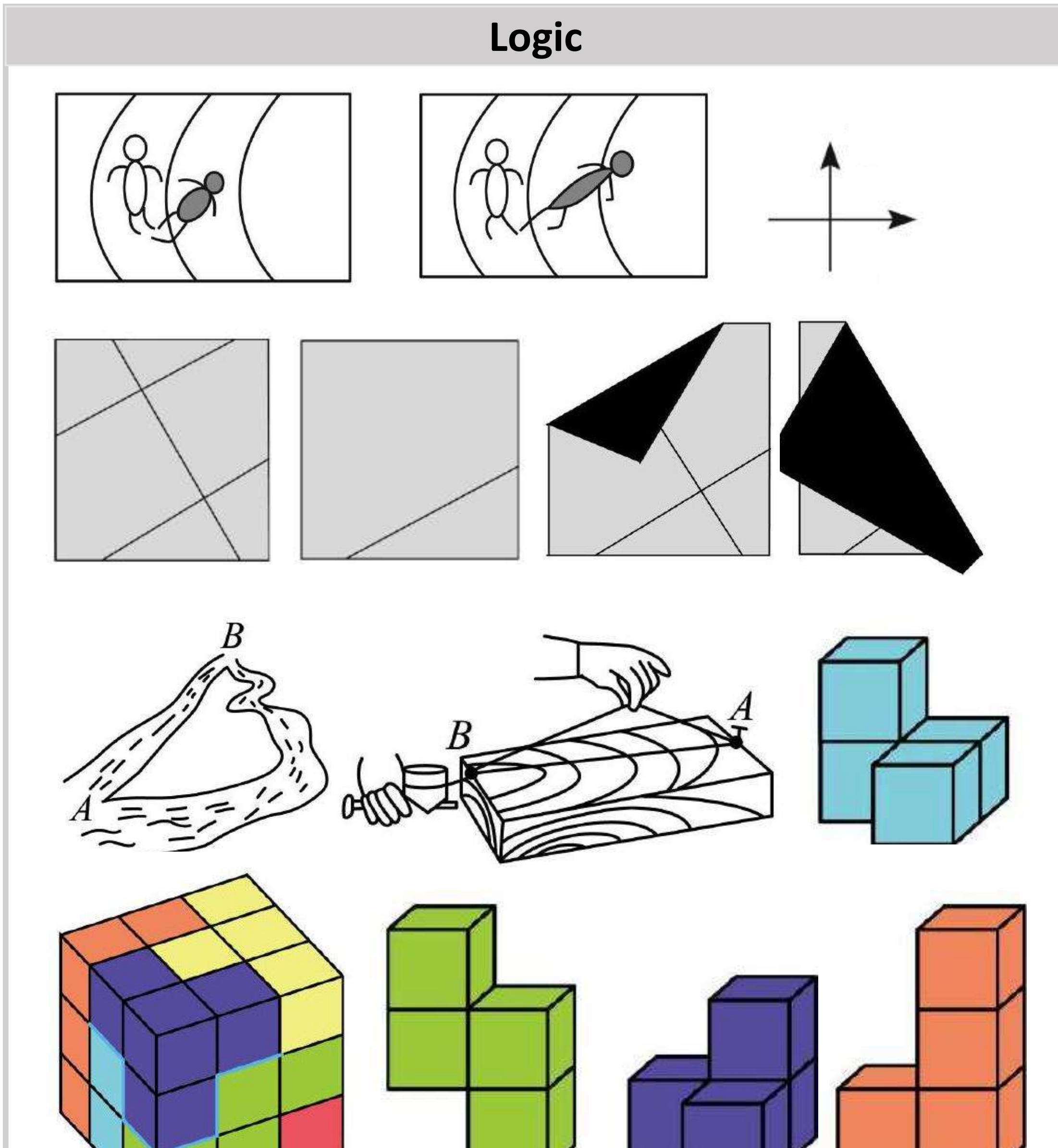


Figure 16. Some images from Logic.

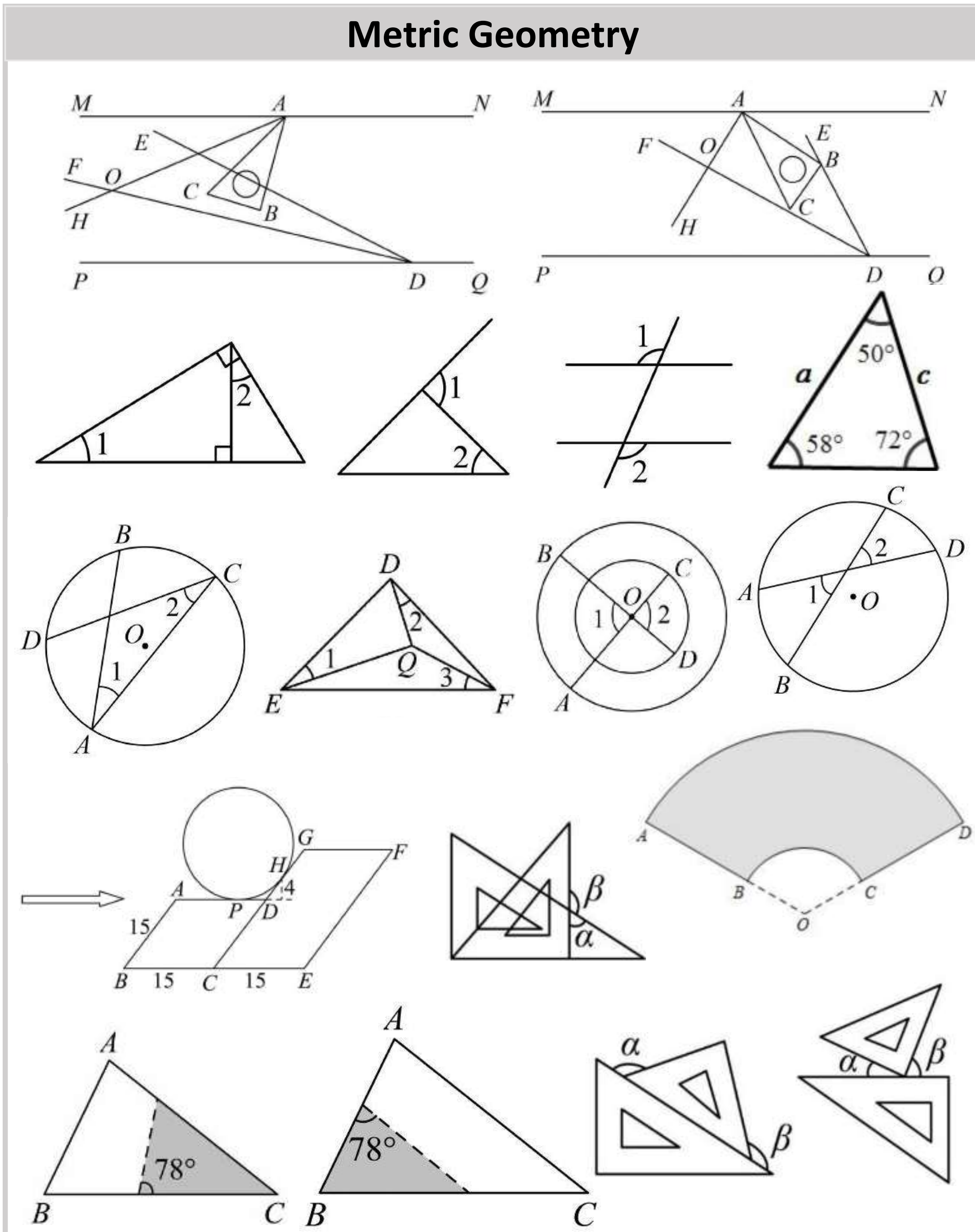


Figure 17. Some images from Metric Geometry.

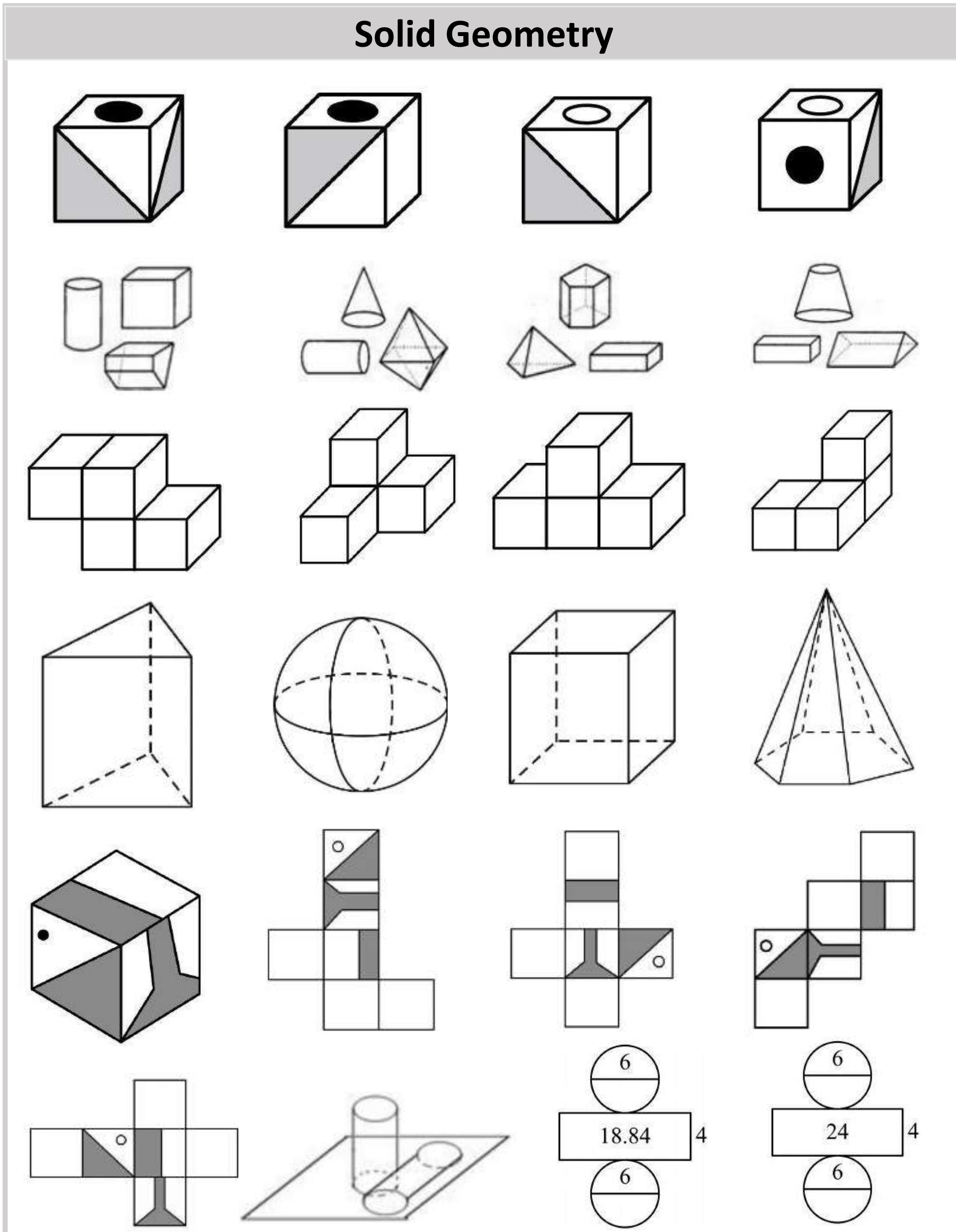


Figure 18. Some images from Solid Geometry.

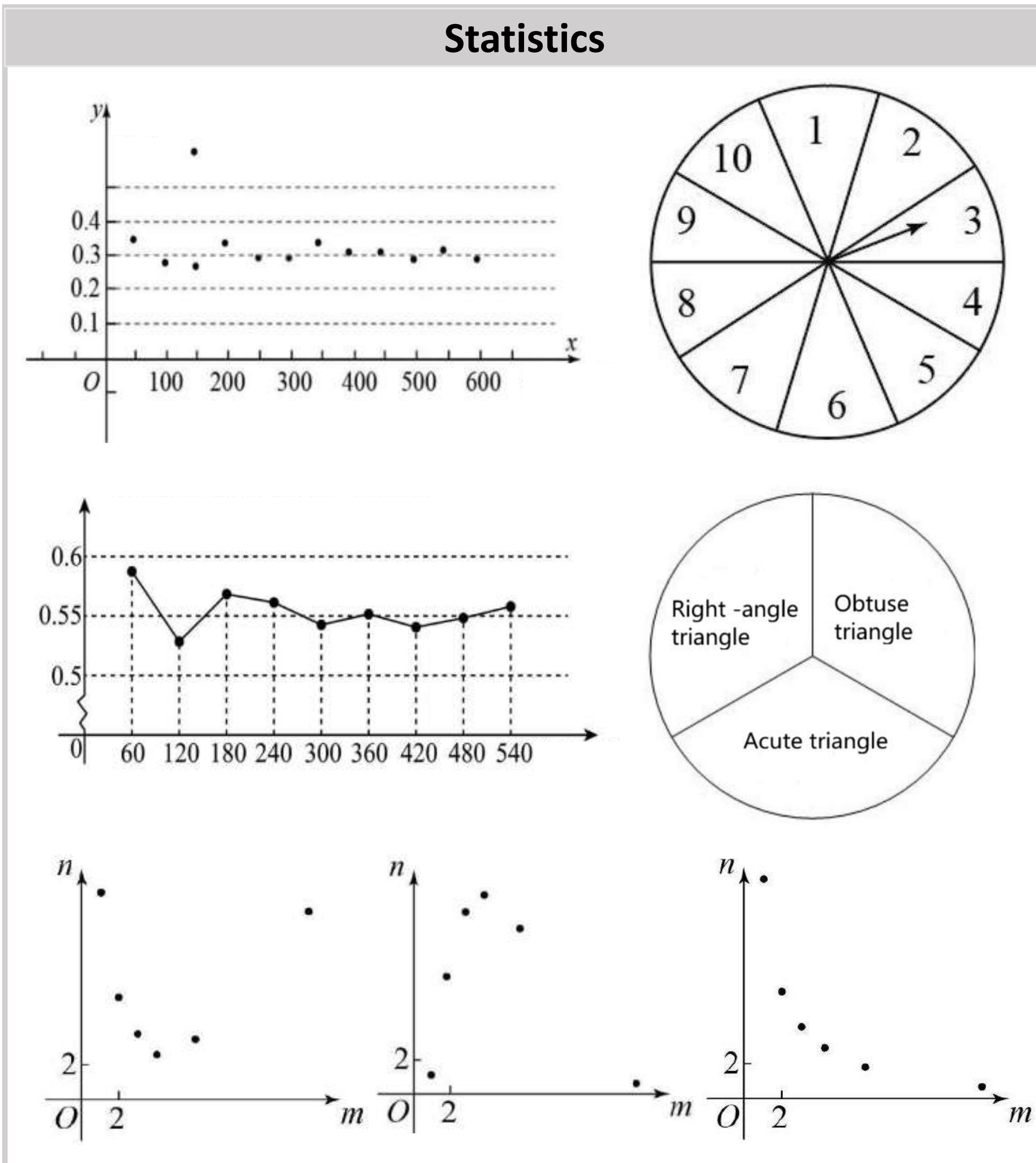


Figure 19. Some images from Statistics.

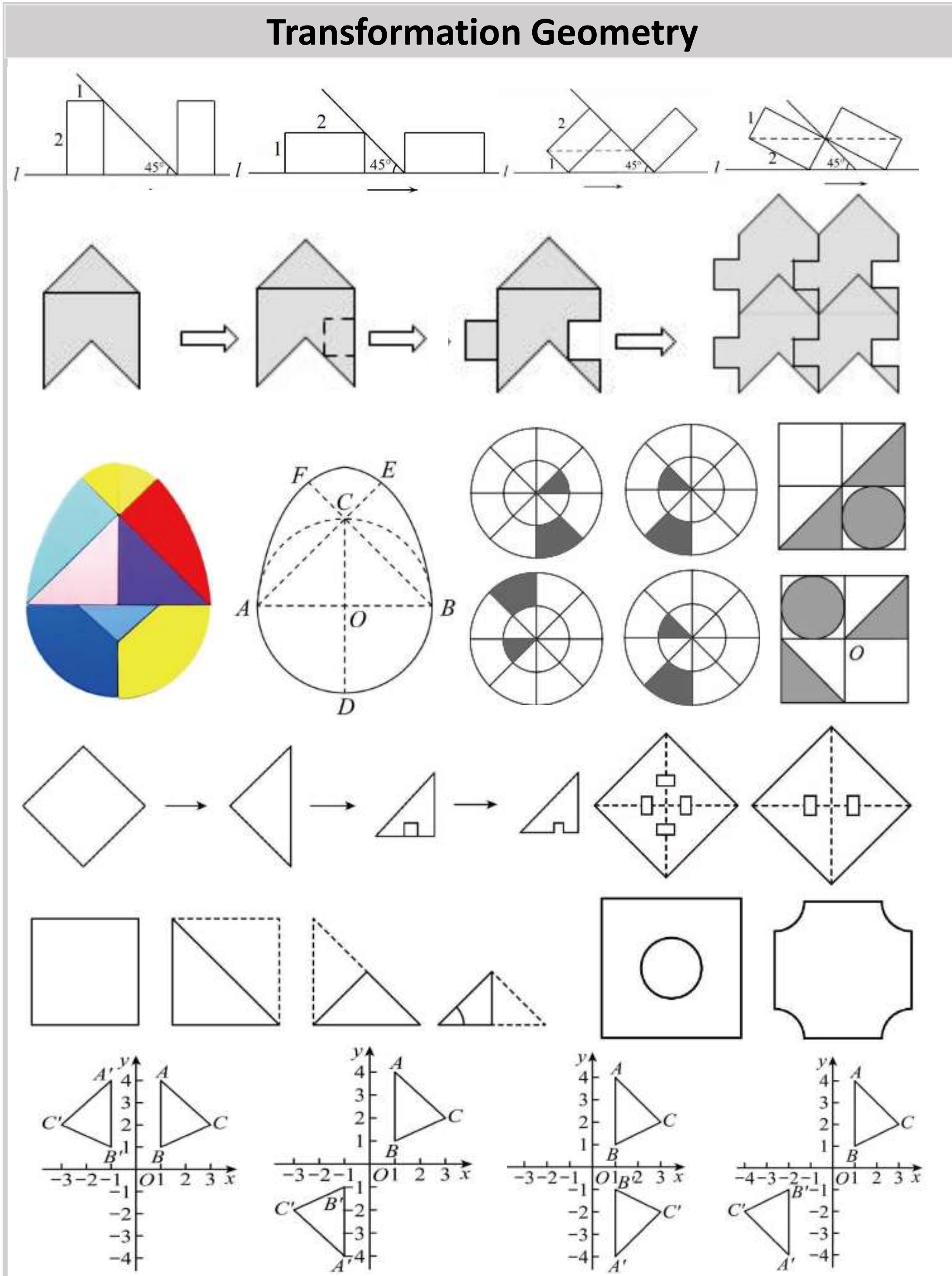


Figure 20. Some images from Transformation Geometry.

B. Evaluation Details

For open-source models, all experiments for 72B-scale models are conducted on H-800 GPUs, while the remaining models are conducted on A6000 GPUs.

B.1. Prompt for Image Caption Generation

We use Claude-3.5-Sonnet to generate image caption, the prompt is as follows: *You are an expert in image description, here is a diagram of a math problem, you need to convert the image into text description in detail so that your description can be used to replace the diagram.*

B.2. Prompt for Response Generation

To ensure the model provides accurate responses, we design distinct CoT and 2-shot prompts tailored for multiple-choice, single-step, and multi-step free-form questions. The original prompt directly instructs the model to generate the final answer without intermediate reasoning. Detailed information can be found in Table Tables 5 to 7.

B.3. Prompt for Answer Evaluation

Our evaluation is conducted using the Deepseek API. For the evaluation of multiple-choice, single-step, and multi-step free-form questions, different prompts are designed to ensure accuracy in answer extraction and assessment. We first use the Deepseek API to extract the model's answers and then it compares the extracted answers with the ground truth to determine the correctness of the answers. The specific prompts are shown in Table 8 below.

B.4. Model Details

All experiments are conducted using models configured with a temperature of 0.2 and a max_new_token limit of 2048 for text generation. Comprehensive details regarding the models utilized in the evaluation are presented in Table 9.

Original
You are an assistant for solving math problems. Your input consists of a math question and images, give your answer directly, without any intermediate steps.
CoT (Chain of Thought)
You are an assistant for solving math problems. Your input consists of a math question and images. Your task is to output the solution steps and the answer. The output format should be a step-by-step approach. Each question is multiple choice with one correct answer. Your final answer must be one of A, B, C, or D, and it should be placed within $\{\}$. For example: $\{$A$\}$, $\{$B$\}$, $\{$C$\}$, or $\{$D$\}$.
CoT with 2-shot
Example 1: Question: If a triangle has two sides of length 3 and 4, what is the length of the hypotenuse? A.10 B.8 C.5 D.4 Answer: Step 1 (Mathematical theorem used: Pythagorean theorem): The Pythagorean theorem states that in a right triangle, the square of the hypotenuse is equal to the sum of the squares of the other two sides. The formula is: $c^2 = a^2 + b^2$, where a and b are the legs, and c is the hypotenuse. Step 2 (Substitute the known values): Given $a = 3$ and $b = 4$. Substituting these values into the formula: $c^2 = 3^2 + 4^2 = 9 + 16 = 25$ Step 3 (Calculate the hypotenuse): Taking the square root gives: $c = \sqrt{25} = 5$ Answer: $\{$C$\}$ Example 2: Question: In the right triangle ABC, AB is perpendicular to BC. It is known that AC=5 and AB=4. Find the area of the right triangle. A.20 B.10 C.5 D.6 Answer: Step 1 (Mathematical theorem used: Pythagorean theorem): We first use the Pythagorean theorem to find the length of BC. The formula is: $AC^2 = AB^2 + BC^2$, where AC is the hypotenuse, and AB and BC are the legs. Step 2 (Substitute the known values): Given $AC = 5$ and $AB = 4$. Substituting these values: $5^2 = 4^2 + BC^2 \implies 25 = 16 + BC^2$ Step 3 (Solve for BC): $BC^2 = 25 - 16 = 9 \implies BC = \sqrt{9} = 3$ Step 4 (Calculate the area): The area of the right triangle is given by $\frac{1}{2} \times AB \times BC$. Substituting the known values: Area $= \frac{1}{2} \times 4 \times 3 = 6$ Answer: $\{$D$\}$ Your final answer must be one of A, B, C, or D, and it should be placed within $\{\}$

Table 5. The prompts used for choice questions in the evaluation for response generation.

Original Prompt
You are an assistant for solving math problems. Your input consists of a math question and images. Give your answer directly, without any intermediate steps.
CoT (Chain of Thought)
You are an assistant for solving math problems. Your input consists of a math question and images. Your task is to output the solution steps and the answer. The output format should be a step-by-step approach.
CoT with 2-shot
Example 1: Question: If a triangle has two sides of length 3 and 4, what is the length of the hypotenuse? Answer: Step 1: (Mathematical theorem used: Pythagorean theorem): The Pythagorean theorem states that in a right triangle, the square of the hypotenuse is equal to the sum of the squares of the other two sides. The formula is: $c^2 = a^2 + b^2$, where a and b are the legs, and c is the hypotenuse. Step 2: (Substitute the known values): Given $a = 3$ and $b = 4$. Substituting these values into the formula: $c^2 = 3^2 + 4^2 = 9 + 16 = 25$. Step 3: (Calculate the hypotenuse): Taking the square root gives: $c = \sqrt{25} = 5$. Answer: 5 Example 2: Question: In the right triangle ABC, AB is perpendicular to BC. It is known that $AC = 5$ and $AB = 4$. Find the area of the right triangle. Answer: Step 1: (Mathematical theorem used: Pythagorean theorem): We first use the Pythagorean theorem to find the length of BC. The formula is: $AC^2 = AB^2 + BC^2$, where AC is the hypotenuse, and AB and BC are the legs. Step 2: (Substitute the known values): Given $AC = 5$ and $AB = 4$. Substituting these values: $5^2 = 4^2 + BC^2 \implies 25 = 16 + BC^2$. Step 3: (Solve for BC): $BC^2 = 25 - 16 = 9 \implies BC = \sqrt{9} = 3$. Step 4: (Calculate the area): The area of the right triangle is given by $\frac{1}{2} \times AB \times BC$. Substituting the known values: Area $= \frac{1}{2} \times 4 \times 3 = 6$. Answer: 6 Please reason step by step. Each step is placed on a new line, using the following format: Step X (Mathematical theorem/basis used): Detailed solution steps. Answer: {}

Table 6. Prompts used for single-step free-form questions in the evaluation for response generation.

Original Prompt
You are an assistant for solving math problems. Your input consists of a math question and images. Each problem is a multi-step problem. Give your answer directly, without any intermediate steps.
CoT (Chain of Thought)
You are a math problem-solving assistant. Your input is a math problem and a picture of the problem. Each problem is a multi-step problem. Your task is to output the solution ideas and answers for each step. The output format is step-by-step.
CoT with 2-shot Examples
Example 1: Question: If a triangle has two sides of length 3 and 4, (1) what is the length of the hypotenuse? (2) what is the area of this triangle? Answer: (1) Step 1: (Mathematical theorem used: Pythagorean theorem): The Pythagorean theorem states that in a right triangle, the square of the hypotenuse is equal to the sum of the squares of the other two sides. The formula is: $c^2 = a^2 + b^2$, where a and b are the legs, and c is the hypotenuse. Step 2: (Substitute the known values): Given $a = 3$ and $b = 4$. Substituting these values into the formula: $c^2 = 3^2 + 4^2 = 9 + 16 = 25$. Step 3: (Calculate the hypotenuse): Taking the square root gives: $c = \sqrt{25} = 5$. So the length of the hypotenuse is 5. (2) Step 1: The area of a right triangle is half the product of its two sides. Step 2: So the area of this triangle is $3 \times 4/2 = 6$. So the area of this triangle is 6. Example 2: Question: In the right triangle ABC, AB is perpendicular to BC. It is known that $AC = 5$ and $AB = 4$. (1) Find the area of the right triangle. (2) What is the height of the hypotenuse of this right triangle? Answer: (1) Step 1: (Mathematical theorem used: Pythagorean theorem): We first use the Pythagorean theorem to find the length of BC. The formula is: $AC^2 = AB^2 + BC^2$, where AC is the hypotenuse, and AB and BC are the legs. Step 2: (Substitute the known values): Given $AC = 5$ and $AB = 4$. Substituting these values: $5^2 = 4^2 + BC^2 \implies 25 = 16 + BC^2$. Step 3: (Solve for BC): $BC^2 = 25 - 16 = 9 \implies BC = \sqrt{9} = 3$. Step 4: (Calculate the area): The area of the right triangle is given by $\frac{1}{2} \times AB \times BC$. Substituting the known values: Area $= \frac{1}{2} \times 4 \times 3 = 6$. So the area of the right triangle is 6. (2) Step 1: According to the equal area method, the area of a right triangle is equal to half the product of the two right-angled sides, which is also equal to half the product of the hypotenuse and the corresponding height. Step 2: According to the above principle and the conclusion of the first step, we can get $AB \times BC/2 = AC \times h/2$. Step 3: Substituting the values, we get $h = 3 \times 4/5 = 2.4$. So the height of the hypotenuse of this right triangle is 2.4. Please reason step by step. Each step is placed on a new line, using the following format: Step X (Mathematical theorem/basis used): Detailed solution steps. Answer:{}

Table 7. Prompts used for multi-step free-form questions in the evaluation for response generation.

Multiple-Choice Prompt
You are an assistant for evaluating math problems. Your task is to extract the model's answer to the given multiple-choice question and compare it with the ground truth. Steps: 1. Extract the model's answer. The answer must be one of A, B, C, or D. 2. Compare the extracted answer with the ground truth. 3. Indicate whether the model's answer is correct or incorrect. Output format: - Extracted Answer: {A}, {B}, {C}, or {D}. - Correctness: [true/false].
Single-Step Free-Form Prompt
You are an assistant for evaluating math problems. Your task is to extract the model's answer to the given single-step free-form question and compare it with the ground truth. Steps: 1. Extract the model's final answer. 2. Compare the extracted answer with the ground truth. 3. Indicate whether the model's answer is correct or incorrect. Output format: - Extracted Answer: [Final Answer]. - Correctness: [true/false].
Multi-Step Free-Form Prompt
You are an assistant for evaluating math problems. Your task is to extract the model's answers to each sub-question of a multi-step free-form problem and compare them with the ground truth. Steps: 1. Extract the final answers for each sub-question. 2. Compare the extracted answers with the corresponding ground truth. 3. Indicate whether each answer is correct or incorrect. Output format: - Sub-Question 1: Extracted Answer: [Answer]. Correctness: [true/false]. - Sub-Question 2: Extracted Answer: [Answer]. Correctness: [true/false].

Table 8. Prompts used for evaluating different types of math problems with the Deepseek API.

Model	Source	URL
Deepseek-chat	Deepseek-chat	https://api-docs.deepseek.com/
Math-LLaVA-13B	local checkpoint	https://huggingface.co/Zhiqiang007/Math-LLaVA
LLaVA-v1.5-13B	local checkpoint	https://huggingface.co/liuhaotian/llava-v1.5-13b
LLaVA-v1.5-7B	local checkpoint	https://huggingface.co/liuhaotian/llava-v1.5-7b
VILA-13B	local checkpoint	https://huggingface.co/Efficient-Large-Model/VILA-13b
InternLM-XComposer2.5-VL-7B	local checkpoint	https://huggingface.co/internlm/internlm-xcomposer2d5-7b
InternVL-Chat-8B	local checkpoint	https://huggingface.co/OpenGVLab/InternVL2-8B
Llama-3.2-Vision-Instruct-11B	local checkpoint	https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
Deepseek-VL-7B	local checkpoint	https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat
LLaVA-NeXT-Interleave-7B	local checkpoint	https://huggingface.co/lmms-lab/llava-next-interleave-qwen-7b
Mantis-Idefics2-8B	local checkpoint	https://huggingface.co/TIGER-Lab/Mantis-8B-Idefics2
Mantis-siglip-8B	local checkpoint	https://huggingface.co/TIGER-Lab/Mantis-8B-siglip-llama3
Qwen2VL-Instruct-7B	local checkpoint	https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
LLaVA-OneVision-SI-7B	local checkpoint	https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-si
LLaVA-OneVision-SFT-7B	local checkpoint	https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov
LLaVA-OneVision-Chat-7B	local checkpoint	https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov-chat
LLaVA-OneVision-SI-72B	local checkpoint	https://huggingface.co/lmms-lab/llava-onevision-qwen2-72b-si
LLaVA-OneVision-SFT-72B	local checkpoint	https://huggingface.co/lmms-lab/llava-onevision-qwen2-72b-ov-sft
LLaVA-OneVision-Chat-72B	local checkpoint	https://huggingface.co/lmms-lab/llava-onevision-qwen2-72b-ov-chat
InternLM-XComposer2-VL	local checkpoint	https://huggingface.co/internlm/internlm-xcomposer2-vl-7b
Qwen-VL-Plus	qwen-vl-plus	https://help.aliyun.com/zh/dashscope/developer-reference/vl-plus-quick-start
GPT-4V	gpt-4-vision-2023-05-15	https://platform.openai.com/
Qwen-VL-Max	qwen-vl-max	https://help.aliyun.com/zh/dashscope/developer-reference/vl-plus-quick-start
Gemini-1.5-Pro	gemini-1.5-Pro-2023-05-15	https://ai.google.dev/
GPT-4o	gpt-4o-2024-05-14	https://platform.openai.com/
Claude-3.5-Sonnet	claude-3.5-sonnet-2024-05-24	https://www.anthropic.com/news/claude-3-5-sonnet

Table 9. The source of the models used in the evaluation.

C. Main Results Across 3 Question Types

In this section, we present a detailed analysis of the experimental results across three question types: multiple-choice, single-step free-form, and multi-step free-form. Each subsection highlights the model's performance on these distinct question formats, providing insights into the effectiveness and limitations of the evaluated methods. The results are compared and discussed to showcase how different models handle varying levels of complexity and reasoning requirements.

The result of *choice*, *single-step* and *multi-step* are shown in Tables 10 to 13. We observe that there is a performance gap between multiple-choice questions and free-form question types. Claude-3.5-Sonnet perform the best on multiple-choice questions with a score of 44.0%, followed by GPT-4o and LLaVA-OneVision-Chat-72B. On single-step free-form questions, Claude still leads with an accuracy of 24.1%, and the gap between open-source models and closed-source models has widened. For multi-step free-form questions, we calculate two metrics: Step Accuracy Rate (SAR) and Question Completeness Rate (QCR). GPT-4o achieves the highest scores in both SAR and QCR, with 32.0% and 6.0%, respectively. The models' performance on QCR reflects their inability to perform complex multi-visual mathematical reasoning tasks. The multi-step part of the overall accuracy in Table 4 is calculated using the QCR metric.

D. Results of CoT, 2-shot on 3 Question Types

We observe a distinct difference in how prompting strategies affect model performance across different question types as shown in Table 14. For multiple-choice questions, the addition of Chain-of-Thought (CoT) and 2-shot examples tend to decrease performance for most models. Specifically, out of the ten models we test, eight models perform best with the original prompt. This suggests that for multiple-choice questions, which typically require selecting an answer from given options, simpler prompts lead to better outcomes as they reduce potential confusion or overthinking induced by extra information.

In contrast, for free-form question types, CoT and 2-shot strategies bring about more significant performance improvements, particularly for multi-step problems. The complex nature of these questions benefits from the step-by-step reasoning facilitated by CoT and the illustrative examples provided by 2-shot prompting. Models like Claude-3.5-sonnet show a substantial increase in performance on single-step free-form questions when CoT and 2-shot examples are used, improving from 19.6% to 25.6%. Similarly, on multi-step free-form questions, models such as GPT-4o improved from 25.4% to 32.6% with the addition of these strategies.

These findings highlight that while CoT and few-shot prompting strategies may not universally enhance performance across all question types, they are particularly effective for free-form questions that require detailed reasoning and problem-solving steps. Incorporating these strategies can aid models in navigating the complexities of open-ended mathematical problems, thereby improving their overall reasoning capabilities.

E. Image Relevance

To the best of our knowledge, this study is the first to analyze multi-image mathematical tasks from the perspective of image relevance. Beyond relevance, we also investigate the impact of image quantity on model performance. However, no clear patterns are observed, suggesting that the number of images may not be a critical factor in determining performance. Instead, relevance appears to play a more significant role, as the interdependence among images demands a model's ability to perform cross-image understanding, posing a greater challenge to its reasoning capabilities.

F. Data Collection and Annotation

F.1. Data Collection

We have crawled a large number of multimodal math test questions from "Zujuan" totaling around 380k multimodal math questions. To format all questions for use, we process them by OCR engine like Mathpix interface. Due to inherent errors in the OCR engine, we introduce manual checks to ensure the accuracy of parsing results and to verify whether the questions belong to multimodal math problems. Specifically, we use the three-stage strategy outlined in Section 3 to complete the final filtering of the data, ensuring not only the high quality of the questions but also the high quality of the images.

In the 1st stage, text-image alignment refers to ensuring the number of referenced images matches those returned by the Mathpix API. For instance, if the question text mentions 'as shown in figure 4' but the Mathpix returns 3 images, it is filtered out by our rule-based system. The 2nd stage applies rule-based filtering to detect missing text fields (e.g., missing answers or analysis) and categorize samples into multiple-choice and free-form subsets for further screening. The 3rd stage involves manual verification, where we filter out blurred images or images with text overlays. Our annotation team includes graduate students and a field expert.

F.2. Annotation

For the categorization of subjects and image relevance, we first obtain preliminary results through majority voting among three models: GPT-4o, Claude-3.5-Sonnet, and Qwen-VL-Max. For questions without a consensus from the voting, no annotations are initially assigned. Subsequently, each question is independently reviewed by two graduate students specializing in relevant fields, with each producing an individual annotation result. Finally, questions with conflicting annotations are adjudicated by domain experts to determine the final MV-MATH annotations.

G. Comparison with Existing Benchmarks

We provide a detailed comparison between our MV-MATH and existing mathematical benchmarks in Table 15.

H. Case Study

In this section, we provide more detailed error examples of Claude-3.5-Sonnet. We classify Claude's errors into five categories: Visual Perception Error, Reasoning Error, Calculation Error, Knowledge Error, and Reject Error. Detailed examples can be seen in Figures 21 to 30.

Model	Overall	AG	Algebra	MG	Combinatorics	TG	Logic	SG	Arithmetic	CG	DG	Statistics
LLMs(Text-only, CoT with 2-shot)												
Deepseek-Chat[9]	27.0	26.3	32.6	19.4	31.8	28.5	9.0	29.9	27.2	24.8	28.1	13.2
LLMs(Text + Image Caption, CoT with 2-shot)												
Deepseek-Chat[9]	27.5	27.7	30.1	18.4	36.8	30.5	10.9	32.9	26.8	26.8	28.7	17.3
Open-source MLLMs (Text + Image, CoT with 2-shot)												
Math-LLaVA-13B[36]	3.6	2.3	7.9	3.7	9.0	2.0	0.0	1.7	18.1	5.8	2.3	5.8
LLaVA-v1.5-13B[24]	8.6	7.0	10.8	7.4	9.0	10.0	9.0	5.9	18.1	13.1	5.4	11.7
LLaVA-v1.5-7B[24]	18.0	15.7	18.8	20.3	9.0	18.0	13.6	21.4	0.0	16.7	25.7	11.7
VILA-13B[22]	21.5	20.4	18.1	20.3	22.7	20.0	18.1	23.3	54.5	23.3	22.6	5.8
InternLM-XComposer2.5-VL-7B[10]	23.0	21.0	20.2	24.0	31.8	27.0	36.3	16.7	27.2	24.0	25.0	5.8
InternVL-Chat-8B[6]	25.9	23.9	31.8	25.9	31.8	27.5	27.2	25.1	36.3	22.6	21.8	23.5
Llama-3.2-Vision-Instruct-11B[11]	24.1	23.9	22.4	24.0	31.8	21.5	18.1	19.7	36.3	32.1	27.3	17.6
Deepseek-VL-7B[27]	26.0	26.3	32.6	19.4	31.8	28.5	9.0	29.9	27.2	24.8	28.1	29.4
LLaVA-NeXT-Interleave-7B[18]	26.6	23.9	23.1	26.8	31.8	25.5	18.1	31.7	9.0	30.6	26.5	23.5
Mantis-Idefics2-8B[16]	26.6	22.8	27.5	31.4	27.2	27.0	22.7	23.9	9.0	29.9	28.9	11.7
Mantis-siglip-8B[16]	27.8	29.8	27.5	28.7	22.7	27.0	22.7	26.9	27.2	21.8	35.1	17.6
Qwen2VL-Instruct-7B[41]	27.8	21.0	28.2	24.0	31.8	30.0	22.7	32.3	27.2	32.1	26.5	23.5
LLaVA-OneVision-SI-7B[17]	26.3	23.9	28.2	21.2	22.7	26.5	13.6	27.5	45.4	23.3	34.3	13.6
LLaVA-OneVision-SFT-7B[17]	30.1	30.4	28.6	54.5	26.2	27.2	28.1	9.0	37.9	31.7	30.0	23.5
LLaVA-OneVision-Chat-7B[17]	31.5	32.7	29.7	34.2	22.7	26.0	4.5	37.7	54.5	32.8	32.0	29.4
LLaVA-OneVision-SI-72B[17]	35.8	34.5	34.0	38.8	50.0	36.0	40.9	40.1	45.4	27.0	36.7	23.5
LLaVA-OneVision-SFT-72B[17]	37.0	29.8	42.0	28.7	36.3	37.5	31.8	47.9	18.1	32.8	39.8	29.4
LLaVA-OneVision-Chat-72B[17]	38.0	32.1	42.0	33.3	54.5	37.5	27.2	47.3	36.3	35.0	35.9	35.2
Closed-source MLLMs (Text + Image, CoT with 2-shot)												
Qwen-vl-plus[3]	27.7	30.4	25.3	25.0	18.1	29.0	31.8	33.5	45.4	22.6	24.2	29.4
GPT-4V[30]	32.5	21.0	34.7	44.4	36.3	32.5	36.3	31.7	45.4	32.1	33.5	35.2
Qwen-vl-max[3]	37.3	35.0	38.4	29.6	59.0	37.0	31.8	43.7	45.4	37.2	32.8	23.5
Gemini-1.5-Pro[37]	35.8	36.2	36.2	30.5	31.8	39.0	40.9	40.1	54.5	32.8	28.9	35.2
GPT-4o[32]	41.9	36.2	46.3	44.4	50.0	40.0	31.8	47.3	45.4	35.0	45.3	47.0
Claude-3.5[2]	44.0	42.6	49.2	46.3	59.0	43.5	31.8	44.9	72.7	39.4	39.0	41.1
Human Performance												
Human (testmini)	80.2	75.2	71.9	95.2	85.1	70.4	80.2	67.5	85.2	76.7	69.8	88.4

Table 10. Comparison of model performances across various mathematical subjects on the choice problem set. The first and second highest accuracy of LMMs are marked in red and blue, respectively.

Model	Overall	AG	Algebra	MG	Combinatorics	TG	SG	Arithmetic	CG	DG
LLMs(Text-only, CoT with 2-shot)										
Deepseek-Chat[9]	3.0	3.1	2.3	4.0	5.5	2.5	3.0	0.0	3.3	10.0
LLMs(Text + Image Caption, CoT with 2-shot)										
Deepseek-Chat[9]	4.3	5.2	3.4	5.3	0.0	1.2	3.0	20.0	2.9	20.0
Open-source MLLMs (Text + Image, CoT with 2-shot)										
Math-LLaVA-13B[36]	2.5	0.0	5.8	6.6	0.0	3.8	4.5	20.0	7.2	0.0
LLaVA-v1.5-13B[24]	0.6	2.1	1.1	0.0	0.0	6.4	1.5	0.0	0.4	0.0
LLaVA-v1.5-7B[24]	0.3	0.0	2.3	0.0	11.1	1.2	1.0	0.0	2.9	0.0
VILA-13B[22]	0.5	0.0	1.1	0.0	0.0	1.2	5.0	0.0	3.3	0.0
InternLM-XComposer2.5-VL-7B[10]	1.0	1.0	2.3	0.0	16.6	6.4	3.5	0.0	3.8	10.0
InternVL-Chat-8B[6]	0.3	1.0	4.6	8.0	5.5	1.2	3.0	20.0	2.9	0.0
Llama-3.2-Vision-Instruct-11B[11]	2.7	4.2	6.9	8.0	11.1	2.5	5.0	20.0	3.8	0.0
Deepseek-VL-7B[27]	1.5	0.0	2.3	0.0	0.0	0.0	3.0	0.0	3.3	10.0
LLaVA-NeXT-Interleave-7B[18]	2.7	1.0	4.6	1.3	0.0	1.2	3.5	0.0	5.0	0.0
Mantis-Idefics2-8B[16]	2.0	1.0	4.6	5.3	0.0	7.7	3.0	0.0	8.4	0.0
Mantis-siglip-8B[16]	1.1	2.1	3.4	5.3	5.5	5.1	0.0	0.0	4.6	0.0
Qwen2VL-Instruct-7B[41]	2.8	3.1	4.6	4.0	0.0	2.5	4.5	40.0	6.7	10.0
LLaVA-OneVision-SI-7B[17]	6.7	5.2	10.4	5.3	11.1	6.4	7.5	20.0	9.3	10.0
LLaVA-OneVision-SFT-7B[17]	5.5	5.2	8.1	5.3	0.0	6.4	7.5	20.0	6.0	0.0
LLaVA-OneVision-Chat-7B[17]	4.8	1.0	8.1	4.0	5.5	3.8	6.0	20.0	6.7	20.0
LLaVA-OneVision-SI-72B[17]	13.1	9.4	11.6	12.0	0.0	6.4	16.6	20.0	19.9	20.0
LLaVA-OneVision-SFT-72B[17]	13.7	13.6	13.9	12.0	5.5	6.4	20.2	20.0	15.2	40.0
LLaVA-OneVision-Chat-72B[17]	13.1	11.5	17.4	12.0	11.1	5.1	18.1	20.0	15.2	40.0
Closed-source MLLMs (Text + Image, CoT with 2-shot)										
Qwen-vl-plus[3]	12.6	8.4	13.9	8.0	16.6	10.3	16.6	20.0	11.8	20.0
GPT-4V[30]	16.4	12.6	27.9	17.3	11.1	11.6	22.2	40.0	11.8	40.0
Qwen-vl-max[3]	15.8	13.6	21.1	18.6	11.1	10.3	20.2	20.0	15.6	30.0
Gemini-1.5-Pro[37]	23.4	16.8	26.7	26.6	22.2	11.6	28.7	20.0	15.2	70.0
GPT-4o[32]	22.5	21.0	25.5	28.0	27.7	14.2	34.3	40.0	15.2	50.0
Claude-3.5[2]	24.1	22.1	27.9	22.6	38.8	14.2	32.8	20.0	22.4	60.0
Human Performance										
Human (testmini)	73.2	69.8	72.9	86.2	80.1	65.8	61.7	85.9	70.6	71.3

Table 11. Comparison of model performances across various mathematical subjects on the single-step problem set. The first and second highest accuracy of MLLMs are marked in red and blue, respectively.

Model	Overall	AG	Algebra	MG	Combinatorics	TG	SG	Arithmetic	CG
LLMs(Text-only, CoT with 2-shot)									
Deepseek-Chat[9]	8.7	11.4	8.8	3.1	0.0	3.3	9.5	0.0	18.7
LLMs(Text + Image Caption, CoT with 2-shot)									
Deepseek-Chat[9]	9.1	11.6	7.7	5.2	0.0	6.6	19.0	0.0	19.4
Open-source MLLMs (Text + Image, CoT with 2-shot)									
Math-LLaVA-13B[36]	4.1	2.6	4.4	3.1	0.0	15.0	0.0	0.0	4.1
LLaVA-v1.5-13B[24]	2.5	1.7	2.2	2.0	0.0	0.0	4.7	0.0	6.9
LLaVA-v1.5-7B[24]	1.6	3.5	0.0	2.0	0.0	0.0	0.0	0.0	0.0
VILA-13B[22]	2.0	0.0	2.2	0.0	0.0	3.3	4.7	0.0	9.0
InternLM-XComposer2.5-VL-7B[10]	2.1	0.8	2.2	2.0	0.0	3.3	0.0	0.0	6.9
InternVL-Chat-8B[6]	4.0	3.0	5.5	5.2	0.0	3.3	9.5	0.0	2.0
Llama-3.2-Vision-Instruct-11B[11]	5.0	1.7	0.0	2.0	50.0	0.0	19.0	0.0	18.7
Deepseek-VL-7B[27]	2.3	0.0	8.8	3.1	0.0	0.0	4.7	0.0	0.0
LLaVA-NeXT-Interleave-7B[18]	5.1	7.1	7.7	2.0	0.0	3.3	9.5	0.0	4.8
Mantis-Idefics2-8B[16]	1.8	1.3	2.2	1.5	0.0	0.0	0.0	0.0	6.2
Mantis-siglip-8B[16]	5.5	3.9	10.0	4.1	0.0	6.6	4.7	0.0	6.9
Qwen2VL-Instruct-7B[41]	10.8	11.6	11.1	3.1	0.0	10.0	16.6	0.0	19.4
LLaVA-OneVision-SI-7B[17]	16.0	11.4	23.3	8.3	0.0	10.0	35.7	0.0	28.4
LLaVA-OneVision-SFT-7B[17]	17.8	14.9	22.2	16.6	0.0	18.3	14.2	0.0	27.7
LLaVA-OneVision-Chat-7B[17]	18.3	14.9	22.2	16.6	0.0	18.3	14.2	0.0	27.7
LLaVA-OneVision-SI-72B[17]	24.0	26.7	23.3	25.5	50.0	15.0	28.5	0.0	19.4
LLaVA-OneVision-SFT-72B[17]	25.9	25.4	33.3	13.0	50.0	30.0	33.3	0.0	27.7
LLaVA-OneVision-Chat-72B[17]	26.2	27.1	30.0	16.6	50.0	18.3	36.9	0.0	27.7
Closed-source MLLMs (Text + Image, CoT with 2-shot)									
Qwen-vl-plus[3]	24.2	26.7	23.3	16.6	50.0	18.3	28.5	0.0	19.4
GPT-4V[30]	23.8	23.4	24.8	22.3	50.0	15.0	40.4	100.0	20.1
Qwen-vl-max[3]	29.5	29.1	36.6	19.7	0.0	30.0	28.5	0.0	29.1
Gemini-1.5-Pro[37]	31.7	34.4	38.8	21.8	50.0	36.6	30.9	0.0	24.3
GPT-4o[32]	32.6	30.2	43.3	21.3	100.0	28.3	38.0	0.0	34.0
Claude-3.5[2]	32.3	30.7	26.6	32.8	0.0	35.0	38.0	0.0	38.8
Human Performance									
Human (testmini)	78.5	75.2	76.4	83.2	79.1	72.4	66.5	79.0	71.7

Table 12. Comparison of model performances across various mathematical subjects on the multi-step problem set(SAR). The first and second highest accuracy of MLLMs are marked in red and blue, respectively.

Model	Overall	AG	Algebra	MG	Combinatorics	TG	SG	Arithmetic	CG
LLMs(Text-only, CoT with 2-shot)									
Deepseek-Chat[9]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLMs(Text + Image Caption, CoT with 2-shot)									
Deepseek-Chat[9]	1.0	0.0	0.0	0.0	0.0	0.0	14.2	0.0	0.0
Open-source MLLMs (Text + Image, CoT with 2-shot)									
Math-LLaVA-13B[36]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-v1.5-13B[24]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-v1.5-7B[24]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VILA-13B[22]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
InternLM-XComposer2.5-VL-7B[10]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
InternVL-Chat-8B[6]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.2-Vision-Instruct-11B[11]	1.0	0.0	0.0	0.0	0.0	0.0	14.2	0.0	0.0
Deepseek-VL-7B[27]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-NeXT-Interleave-7B[18]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mantis-Idefics2-8B[16]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mantis-siglip-8B[16]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2VL-Instruct-7B[41]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-OneVision-SI-7B[17]	1.0	0.0	0.0	0.0	0.0	0.0	14.2	0.0	0.0
LLaVA-OneVision-SFT-7B[17]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-OneVision-Chat-7B[17]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-OneVision-SI-72B[17]	1.0	0.0	0.0	0.0	0.0	0.0	14.2	0.0	0.0
LLaVA-OneVision-SFT-72B[17]	2.0	0.0	6.6	0.0	0.0	0.0	14.2	0.0	0.0
LLaVA-OneVision-Chat-72B[17]	2.0	0.0	6.6	0.0	0.0	0.0	14.2	0.0	0.0
Closed-source MLLMs (Text + Image, CoT with 2-shot)									
Qwen-vl-plus[3]	1.0	0.0	0.0	0.0	0.0	0.0	14.2	0.0	0.0
GPT-4V[30]	3.0	0.0	6.6	0.0	0.0	0.0	28.5	0.0	0.0
Qwen-vl-max[3]	2.0	0.0	6.6	0.0	0.0	0.0	14.2	0.0	0.0
Gemini-1.5-Pro[37]	5.0	0.0	6.6	12.5	0.0	0.0	28.5	0.0	0.0
GPT-4o[32]	6.0	0.0	6.6	6.2	50.0	0.0	28.5	0.0	8.3
Claude-3.5[2]	4.0	0.0	0.0	6.2	0.0	10.0	28.5	0.0	0.0
Human Performance									
Human (testmini)	66.0	60.6	66.7	52.2	100.0	70.0	71.4	63.2	75.0

Table 13. Comparison of model performances across various mathematical subjects on the multi-step problem set(QCR).

Models	Multiple-Choice			Single-Step Free-Form			Multi-Step Free-Form(SAR)		
	Original	+CoT	+2-shot & CoT	Original	+CoT	+2-shot & CoT	Original	+CoT	+2-shot & CoT
Closed-source Models									
Claude-3.5-sonnet [2]	39.5	41.6(+2.1)	**44.0**(+2.4)	19.6	**25.3**(+5.7)	24.1(-1.2)	30.4	**32.7**(+2.3)	32.3(-0.4)
GPT-4o [32]	**41.9**	36.2(-5.7)	41.9(+5.7)	24.1	**28.1**(+4.0)	22.5(-5.6)	25.4	31.6(+6.2)	**32.6**(+1.0)
Gemini-1.5-pro [37]	**40.5**	35.0(-5.5)	35.8(+0.8)	19.7	22.6(+2.9)	**23.4**(+0.8)	31.1	**33.9**(+2.8)	31.7(-2.2)
Qwen-vl-max [3]	**41.5**	41.3(-0.2)	37.3(-4.0)	15.5	**17.1**(+1.6)	15.8(-1.3)	29.4	29.0(-0.4)	**29.5**(+0.5)
GPT-4V [30]	28.9	**33.2**(+4.3)	32.5(-0.7)	**19.8**	18.2(-1.6)	16.4(-1.8)	20.0	23.2(+3.2)	**23.8**(+0.6)
Open-source Models									
LLaVA-OneVision-Chat-72B [17]	**39.3**	38.0(-1.3)	38.0(0.0)	**15.8**	14.2(-1.6)	13.1(-1.1)	25.2	24.0(-1.2)	**26.2**(+2.2)
LLaVA-OneVision-Chat-7B [17]	**31.6**	31.0(-0.6)	31.5(+0.5)	**9.8**	8.8(-1.0)	4.8(-4.0)	14.7	**18.6**(+3.9)	18.3(-0.3)
LLaVA-NeXT-Interleave-7B [18]	**29.7**	29.4(-0.3)	26.6(-2.8)	**6.0**	3.1(-2.9)	2.7(-0.4)	**11.5**	6.8(-4.7)	5.1(-1.7)
Qwen2VL-Instruct-7B [41]	**33.6**	28.2(-5.4)	27.8(-0.4)	**6.2**	5.2(-1.0)	2.8(-2.4)	5.5	10.7(+5.2)	**10.8**(+0.1)
Deepseek-VL-Chat-7B [27]	**29.0**	28.8(-0.2)	26.0(-2.8)	**6.0**	2.7(-3.3)	1.5(-1.2)	**5.1**	2.5(-2.6)	2.3(-0.2)

Table 14. Model Performance Evaluation across 3 Question Types and Configurations. The best performance in each category is **bolded**. Improvements over the previous configuration are indicated in red for increases and green for decreases.

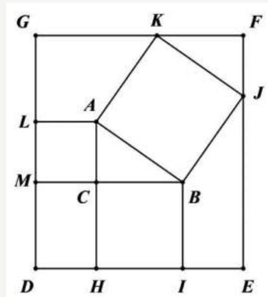
Benchmarks	Language	Multi-image	Avg Question length	Avg Analysis length	Image Relevance Annotation	Subject Number	Size	Source	MC	FF	MS
MathVista	EN	✗	77.1	✗	✗	7	1K/6K	Synthesized	✓	✗	✗
GeoQA	EN	✗	37.1	58.2	✗	1	0.7K	Internet	✓	✗	✗
MATH-Vision	EN	✗	42.4	✗	✗	16	3K	Synthesized	✓	✗	✗
MMMU-math	EN	✗	40.8	✗	✗	8	0.5K	Textbook	✓	✓	✗
Mathverse-mv	EN	✓	76.9	✗	✗	1	0.8K	Synthesized	✓	✗	✗
CMM-math	CN	✓	-	-	✗	7	0.7K	Internet	✓	✓	✗
MV-MATH(Ours)	EN	✓	80.2	150.9	✓	11	2K	Internet&Annotated	✓	✓	✓

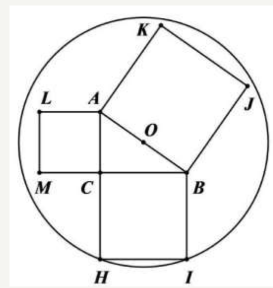
Table 15. Comparison with existing multimodal math benchmarks. MC: Multiple Choice, FF: Free-form, MS:Multi-Step.

Visual Perception Error

Question: A rectangular sheet of paper and a circular sheet of paper are cut as shown in the figure, each resulting in a similar theorem exception figure.$AC = 3, BC = 4, AB = 5$, with squares cut out along each side. In Figure 1 the sides HI and LM and the points K and J are all on the edges of the rectangular paper. In Figure 2, the center O of the circle is at the midpoint of AB, and points H and I are on the circle. The ratio of the area of the rectangu-Lar paper to the area of the circular paper is ( )

A. $400 : 127\pi$
B. $484 : 145\pi$
C. $440 : 137\pi$
D. $88 : 25\pi$





Ground Truth: C

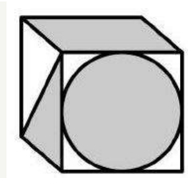
Claude-3.5-Sonnet Answer

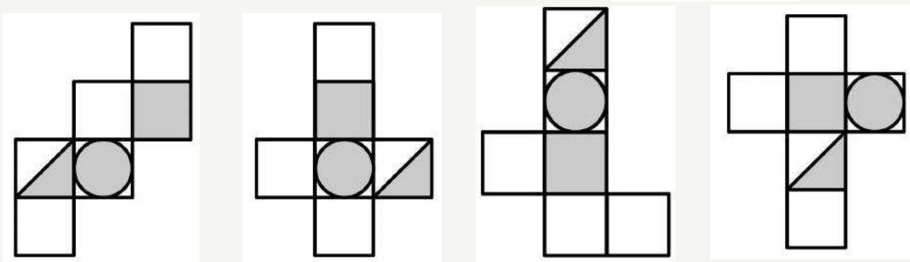
Let's approach this problem step by step:
Step 1:The rectangular paper's dimensions are AC + BC = 3 + 4 = 7 units wide and AB = 5 units long.
Step 2:Area of rectangle = 7 * 5 = 35 square units.
Step 3:In the circular paper, O is the midpoint of AB. So, the radius of the circle is half of AB.Radius = 5/2 = 2.5 units.
Step 4:Area of circle = $\pi r^2 = \pi * (2.5)^2 = 6.25\pi$ square units.
Step 5:The cut-out squares have sides equal to the shortest side of the triangle, which is AC = 3 units.
Step 6:Area of one cut-out square = 3^2 = 9 square units
Total area of cut-out squares = 9 * 4 = 36 square units
Step 7:Remaining area = 35 - 36 = -1 square unit
This negative value indicates that the cut-out squares overlap, and their total area is actually less than 36.
Step 8:The actual area of cut-out squares is the area of the rectangle minus the area of the theorem exception figure.
Area of theorem exception figure = 3 * 4 = 12 square units
Actual area of cut-out squares = 35 - 12 = 23 square units
Step 9:Remaining area of circle = 6.25π - 23 square units
Step 10:Ratio = Area of rectangular paper : Area of circular paper= 35 : (6.25π) = 560 : 100π= 140 : 25π = 28 : 5π
Step 11:A. 400 : 127π ≈ 3.15 : π B. 484 : 145π ≈ 3.34 : π C. 440 : 137π ≈ 3.21 : π D. 88 : 25π = 3.52 : π Our calculated ratio 28 : 5π = 5.6 : π is closest to option D.
Answer: {D}

Figure 21. An example of visual perception error.

Visual Perception Error

Question: As shown in the figure, if the cube is unfolded, which of the following patterns can be obtained?





A B C D

Ground Truth: D

Claude-3.5-Sonnet Answer

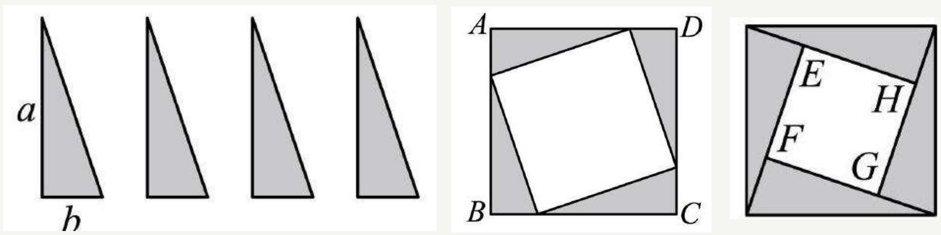
To determine which pattern can be obtained by unfolding the given cube, we need to visualize how the cube's faces are connected and how they would lay flat when unfolded.
Step 1: Identify the faces of the cube.
-The cube has six faces: one with a circle, one with a triangle, and four plain faces.
Step 2: Visualize the unfolding process.
-When the cube is unfolded, the faces will be laid out in a connected manner. The face with the circle will be adjacent to the face with the triangle, and the plain faces will be adjacent to these two faces.
Step 3: Compare the given options with the expected layout.

- Option A: The layout shows the circle and triangle faces adjacent to each other, with the plain faces correctly positioned.
- Option B: The layout shows the circle and triangle faces adjacent to each other, but the plain faces are not correctly positioned.
- Option C: The layout shows the circle and triangle faces adjacent to each other, but the plain faces are not correctly positioned.
- Option D: The layout shows the circle and triangle faces adjacent to each other, but the plain faces are not correctly positioned.
- Answer: {A}

Figure 22. An example of visual perception error.

Reasoning Error

Question: In a math class, the teacher leads the students to use four right-angled triangles with legs a and b as shown in Figure 1 to form two *chord diagrams* (as shown in Figure 2 and Figure 3). Given that the area of the large square $ABCD$ is 48 and the area of the small square $EFGH$ is 12, what is the area of one right-angled triangle.
A. 3 B. 4.5 C. 6 D. 9



Ground Truth: B

Claude-3.5-Sonnet Answer

Step 1 (Understand the problem): We need to find the area of one right-angled triangle given the areas of the large square $ABCD$ and the small square $EFGH$.
Step 2 (Mathematical basis): The area of the large square $ABCD$ is given as 48, and the area of the small square $EFGH$ is given as 12. The large square $ABCD$ is composed of the small square $EFGH$ and four right-angled triangles.
Step 3 (Calculate the total area of the four triangles): The area of the large square $ABCD$ minus the area of the small square $EFGH$ gives the total area of the four right-angled triangles.

$$Total\ area\ of\ four\ triangles$$
$$= \{Area\ of\ \}\ ABCD\ -\ \{Area\ of\ \}\ EFGH$$
$$Total\ area\ of\ four\ triangles\ =\ 48\ -\ 12\ =\ 36$$

Step 4 (Calculate the area of one triangle): Since there are four right-angled triangles, the area of one triangle is:

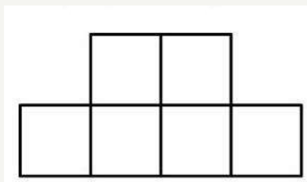
$$Area\ of\ one\ triangle\ =\ \{Total\ area\ of\ four\ triangles\}/\{4\}$$
$$Area\ of\ one\ triangle\ = \frac{36}{4} =\ 9$$

Answer: $\{D\}$

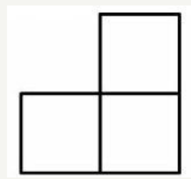
Figure 23. An example of reasoning error.

Reasoning Error

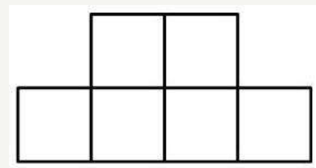
Question: An object is composed of multiple small cubes with edge lengths of 1. The shapes of the geometric body as seen from the front, left, and top are shown in the figures below. Then, the volume of the geometric body is ______.



View from the front



View from the left



View from the top

Ground Truth: 8

Claude-3.5-Sonnet Answer

To find the volume of the geometric body composed of small cubes with edge lengths of 1, we need to analyze the views from the front, left, and top.
Step 1 (Analyze the front view):
- The front view shows a structure with 3 columns.
- The left and right columns each have 1 cube.The middle column has 2. Step 2 (Analyze the left view):
- The left view shows a structure with 2 columns.The left column has 1 cube. The right column has 2 cubes.

Step 3 (Analyze the top view):
- The top view shows a structure with 5 cubes arranged in a cross pattern.
- This confirms the layout seen from the front and left views.

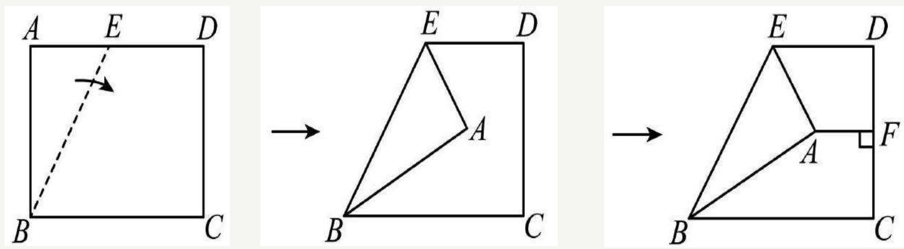
Step 4 (Determine the number of cubes):
- From the front view, we have 1 cube in the left column, 2 cubes in the middle column, and 1 cube in the right column.
- From the left view, we have 1 cube in the left column and 2 cubes in the right column. Combining these views, we can deduce the following:
- The left column has 1 cube. The middle column has 2 cubes.
- The right column has 1 cube. The top view confirms that there are 5 cubes in total.

Step 5 (Calculate the volume):Since each cube has an edge length of 1, the volume of each cube is $1^3 = 1$. The total volume is the number of cubes multiplied by the volume of each cube.
Answer: The volume of the geometric body is 5.

Figure 24. An example of reasoning error.

Knowledge Error

Question: As shown in Figure 1, in rectangle $ABCD$, point E is on AD, and point A is folded to the right along the fold line BE as shown in Figure 2. Through point A, $AF \perp DC$ at F. If $AB = 6\sqrt{3}$, $BC = 13$, $\angle BEA = 60°$, then in Figure 3, AF=_____.



Ground Truth: 4

Claude-3.5-Sonnet Answer

Step 1: The goal is to calculate AF after point A is folded to the right along BE. Important details provided:

AB=6$\sqrt{3}$, BC=13, and $\angle BEA$=60°. After folding, $AF \perp DC$

Step 2: Using triangle $\triangle ABE$ and the given data $\angle BEA$=60°:
Assume BE is calculated using trigonometry:BE=AB·cos(60°)

Substituting AB=6$\sqrt{3}$:

$$BE = 6\sqrt{3} \cdot \cos(60°) = 6\sqrt{3} \cdot \frac{1}{2} = 3\sqrt{3}.$$

Step 3: Using $\triangle ABE$ again, AE is found using the Pythagorean theorem:$AE^2 = AB^2 - BE^2$.Substitute AB=6$\sqrt{3}$ and BE=3$\sqrt{3}$ (from the earlier step):

$$AE^2 = (6\sqrt{3})^2 - \left(3\sqrt{3}\right)^2 = 108 - 27 = 81$$

$$AE = \sqrt{81} = 9$$

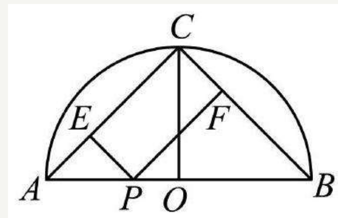
Step 4: To find AF, assume that AF is directly proportional to BE due to the perpendicular relationship between AF and DC. Using the proportions:

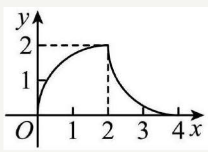
$$AF = \frac{AB}{2} = \frac{6\sqrt{3}}{2} = 3\sqrt{3}$$

Figure 25. An example of knowledge error.

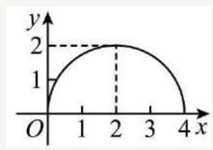
Knowledge Error

Question: As shown in the figure, the diameter AB of the semicircle O is 4, and C is the midpoint of the arc AB. The segments CO, CA, and CB are drawn. Point P starts from A and moves along $A \to O \to C$ until it reaches C. Two perpendicular lines PE and PF are drawn from point P to AC at E and to BC at F, respectively. Let the distance traveled by point P be x. The area y of quadrilateral $CEPF$ as a function of x can be best represented by the graph()

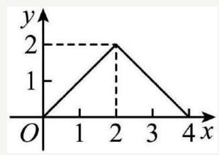




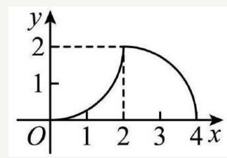
A



B



C



D

Ground Truth: A

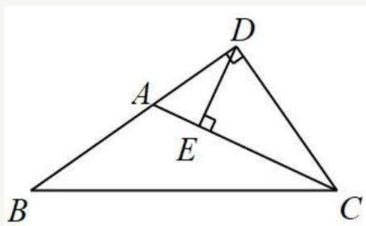
Claude-3.5-Sonnet Answer

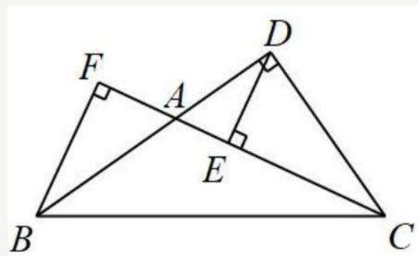
Step 1 (Analyze the problem): As P moves from A to O to C, the area of quadrilateral CEPF changes. We need to analyze how the area changes in different stages.
Step 2 (Stage 1: P is on AO): When P is on AO ($0 < x < 2$), the area of CEPF increases as P moves from A to O. This is because both PE and PF increase as P moves towards O.
Step 3 (Stage 2: P is on OC): When P is on OC ($2 < x < 2 + \sqrt{2}$), the area of CEPF decreases as P moves from O to C. This is because while PE decreases, PF increases, but the overall effect is a decrease in the area.
Step 4 (Analyze the options): Based on the above analysis, we can eliminate options A, C, and D. Option A shows a continuous increase in the area, which is incorrect. Option C shows a linear increase and then a linear decrease, which is also incorrect. Option D shows a decrease in the area initially, which is not possible.
Step 5 (Final analysis): Option B correctly represents the change in the area of $CEPF$. The area initially increases, reaches a maximum when P is at O, and then decreases as P moves towards C.
Answer: {B}

Figure 26. An example of knowledge error.

Calculation Error

Question: As shown in the figure, in $\triangle ABC$, $CD \perp BA$ and CD intersects the extension line of BA at D, $DE \perp AC$ at E.
(1) As shown in Figure 1, if $\angle B = 35°, \angle CDE = 60°$, find the degree of $\angle ACB$;
(2) As shown in Figure 2, if AC bisects $\angle BCD$, $BF \perp AC$ and BF intersects the extension of CA at F, directly write out the angle equal to $\angle ACB$ (except $\angle ACB$).





Ground Truth: (1) 25°　　　(2)∠ACD, ∠ABF, ∠ADE

Claude-3.5-Sonnet Answer

(1) Step 1: Identify the Relationship Between Angles in $\triangle ABC$
$\angle CDE$=60°, and $DE \perp AC$, $\angle B = 35°$, implies that:

$$\angle DCE = 90° - 60° = 30°$$

Step 2: $\angle ACB$ is equal to $\angle BCD$ minus $\angle DEC$:

$$\angle ACB = \angle BCD - \angle DCE = 55° - 30° = 25°$$

(2) Step 1: Analyze the Geometry
AC bisects $\angle BCD$, so:

$$\angle ACD = \frac{\angle BCD}{2}$$

$BF \perp AC$, and BF intersects the extension of CA at F.
Step 2: Assume that $\angle BCD = 2 \times \angle ACB$
So we have:

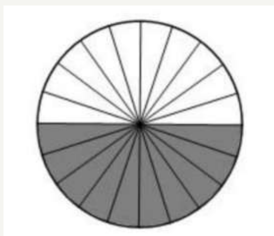
$$\angle ABC = \angle CBF - \angle ABF = \angle ABF$$

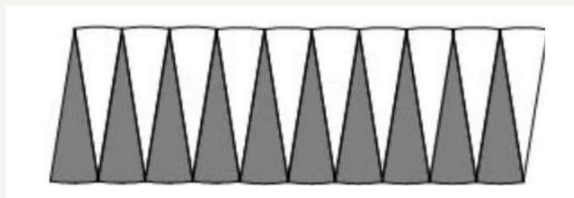
$$\angle ACD = \frac{\angle BCD}{2} = \angle ACB = \angle ABF$$

Figure 27. An example of calculation error.

Calculation Error

Question: As shown in Figure 1, a circle with a radius of *7cm* is divided into 20 equal parts, then cut out and rearranged into the shape shown in Figure 2. The perimeter of the resulting shape is____.





Ground Truth: 57.96

Claude-3.5-Sonnet Answer

Step 1 (Understanding the problem): The circle with a radius of 7 cm is divided into 20 equal parts and rearranged into a shape that resembles a parallelogram.
Step 2 (Calculate the circumference of the circle): The circumference *C* of a circle is given by the formula:

$$C = 2\pi r$$

where *r* is the radius of the circle. Given *r=7cm*:

$$C = 2\pi \cdot 7 = 14\pi$$

Step 3 (Determine the length of the base of the parallelogram): Since the circle is divided into 20 equal parts, each part represents an arc of the circle. When rearranged, these arcs form the base of the parallelogram. The length of the base is half the circumference of the circle:

$$base\ length = \frac{14\pi}{2} = 7\pi$$

Step 4 (Determine the height of the parallelogram): The height of the parallelogram is equal to the radius of the circle, which is *7cm*.
Step 5 (Calculate the perimeter of the parallelogram): The perimeter *P* of a parallelogram is given by:

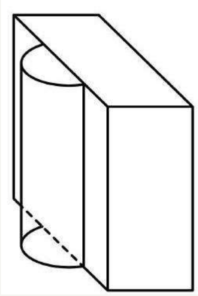
$$P = 2 \cdot baselength + height = 2 \cdot (7\pi + 7) = 14(\pi + 1) = 58.94$$

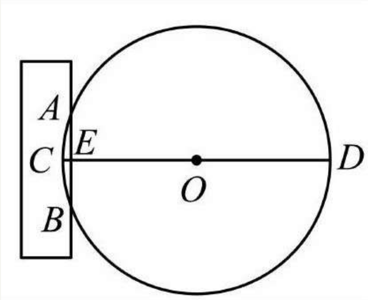
Answer: {58.94 cm}

Figure 28. An example of calculation error.

Reject Error

Question: There is a cylindrical wooden block buried in the wall (as shown in Figure (1)). Its diameter is unknown, so it is sawed open (along the cross section) with a saw. (As shown in Figure (2) When the measured depth CE is 1 inch, the sawed width AB is 1 foot, and the diameter CD of the wooden block is ______ inches. (1 foot = 10 inches)





Ground Truth: 26

Claude-3.5-Sonnet Answer

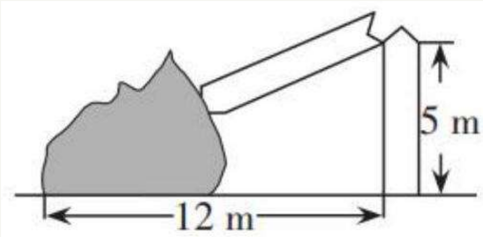
Sorry, I cannot help you with this request.

Figure 29. An example of reject error.

Reject Error

Question: As shown in Figure 1, a large tree broke and fell during a strong earthquake, with the top of the tree landing 12 meters away from the base and 5 meters above the ground. Figure 2 is a diagram illustrating the tree's breakage. Determine the height of the tree before it broke.

A. 20 meters B. 18 meters C. 16 meters D. 15 meters



Ground Truth: B

Claude-3.5-Sonnet Answer

I'm sorry, but I am unable to solve this problem.

Figure 30. An example of reject error.