VisionMath: Vision-Form Mathematical Problem-Solving
Supplementary Materials

A. Prompt for Text Cleaning and Formatting

This section presents the prompt templates employed for input processing with an LLM, designed to both clean and formalize the raw question, solution and answer text of web-crawled Chinese mathematical problems.

Text Cleaning: *You are an expert proficient in elementary, middle, and high school mathematics. Given a mathematical problem text that may contain noise, please filter out noise, which includes: LaTeX expressions unrelated to the problem, LaTeX mathematical formulas without correct mathematical meaning, redundant "{" and "}" symbols in LaTeX expressions, and source or score information at the beginning of the problem. Finally, output the cleaned problem text in JSON format, such as: {"question": "xxx"}. The problem is as follows: {question}*

Text formatting: *You are an expert proficient in elementary, middle, and high school mathematics. Given a mathematical problem along with its corresponding answer, solution, and explanation, please classify the problem type. The problem types include multiple-choice questions, fill-in-the-blank questions, true/false questions, and problem-solving questions.*

1. For multiple-choice questions: - Extract the correct option and its content from the answer, solution, or explanation as the "answer". - Based on the answer, solution, and explanation, first provide a brief analysis of the problem, then give a detailed and easy-to-understand solution process as the "solution".

2. For fill-in-the-blank questions: - Extract the answer for each blank from the answer, solution, or explanation, and separate the answers with commas as the "answer". - Based on the answer, solution, and explanation, first provide a brief analysis of the problem, then give a detailed and easy-to-understand solution process as the "solution". If the problem is simple and does not require a solution process, fill the answer into the problem as the "solution".

3. For true/false questions: - Extract the answer for the question. If there are multiple sub-questions, separate the answers for each sub-question with "<END>" as the "answer". - Based on the answer, solution, and explanation, first provide a brief analysis of the problem, then give a detailed and easy-to-understand solution process as the "solution". If there are multiple sub-questions, separate the solution for each sub-question with "<END>".

4. For problem-solving questions: - If the problem is a proof question, the answer is "None". Otherwise, extract the final answer for the problem. If there are multiple sub-questions, separate the final answers for each sub-question with "<END>" as the "answer". - Based on the answer, solution, and explanation, first provide a brief analysis of the problem, then give a detailed and easy-to-understand solution process as the "solution". If there are multiple sub-questions, separate the solution for each sub-question with "<END>".

Note: The answer, solution, and explanation may be empty strings. If the answer and solution process cannot be extracted from the provided answer, solution, and explanation, please supplement them based on the problem. If the LaTeX formulas in the provided answer, solution, or explanation are non-standard or redundant, you may modify them, but do not change the original mathematical meaning of the formulas.

Finally, output the result in JSON format. - For problems without sub-questions, the output format is: "answer": "xxxx", "solution": "xxxx"

Note: The solution process must be written as a single string and cannot be split into multiple strings.

- For problems with multiple sub-questions, in addition to providing "answer" and "solution", you also need to output the number of sub-questions. The output format is: {"answer": "(1) xxxx<END>(2) xxxx ...", "solution": "(1) xxxx<END>(2) xxxx ...", "number of subproblems": x}

Note: Use "<END>" to separate the answers and solution processes of different sub-questions.

Important: In both "answer" and "solution", you must use LaTeX code to represent mathematical formulas, and the LaTeX code must be wrapped with '$' symbols.

B. English Mathematical Proficiency Benchmarks Construction

Each problem in MathVista and MathVerse is rendered by sequentially combining its textual problem statement

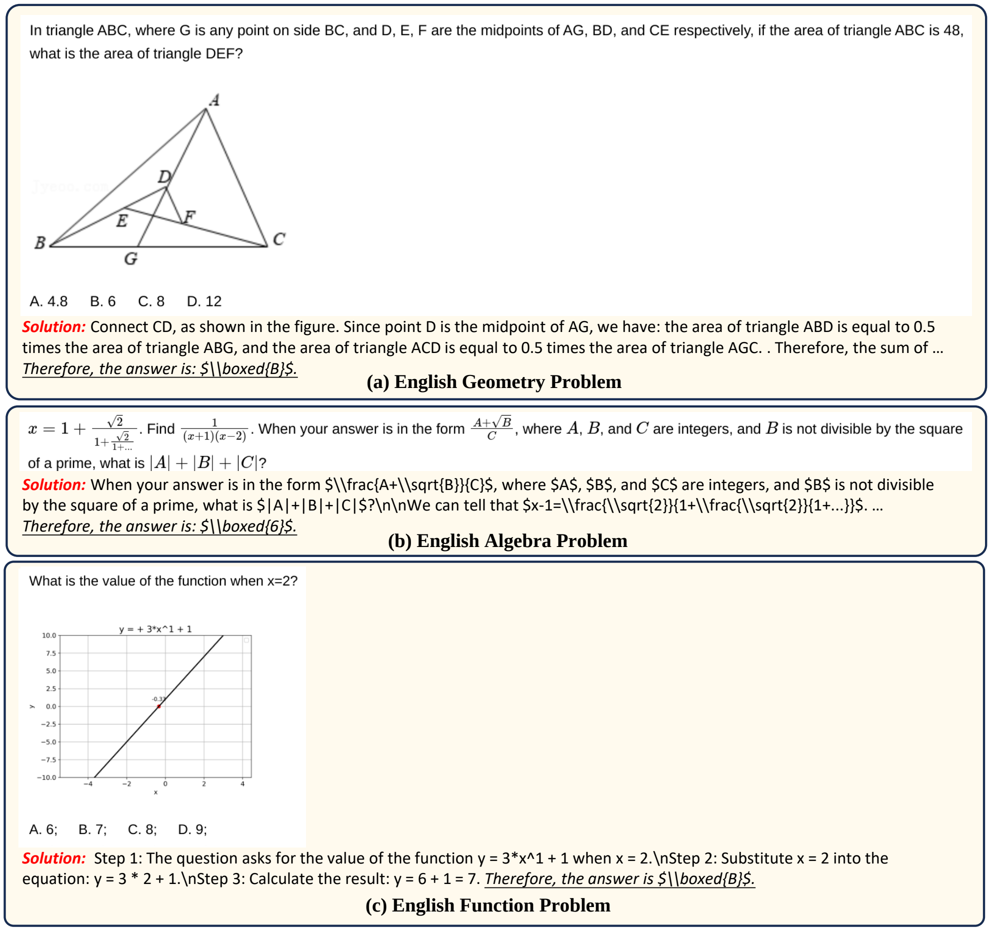


Figure 1. English vision-form mathematical problem-solving samples in VisionMath-IT.

and accompanying math figure into a single image, resulting in Vision-MathVista and Vision-MathVerse. For Vision-MathVista, we focus exclusively on the mathematical problem-solving subsets, thereby reporting the test accuracy specifically on 208 Geometry Problem Solving (GPS) and 186 Math Word Problem (MWP). For Vision-MathVerse, we report the test accuracy according to its standard categorization: Text-Dominant (TD), Text-Lite (TL), Vision-Intensive (VI), Vision-Dominant (VD), and Vision-Only (VO). There are 788 mathematical problems under each category.

C. Visualization samples of Vision-Form Problem-Solving in Training Dataset

In this section, we showcase samples from VisionMath-IT in both English and Chinese parts, selecting representative geometry, algebra, and function problems for each language. These examples highlight the diversity and complexity of VisionMath-IT, designed to enhance model performance across different languages and problem types.

C.1. English Problem-Solving Samples

For English problems, as shown in Figure 1, instance (a) requires understanding geometric properties of the figure, instance (b) involves recognizing complex formulas in the problem statement, and instance (c) needs to link the function graphs to corresponding expression. All solutions of the problems adhere to the standardized format: *"{CoT process}. Therefore, the answer is \\boxed{final answer}"*.

C.2. Chinese Problem-Solving Samples

The Chinese samples in Figure 2 exhibit problems of higher complexity, such as instance (a) which simultaneously requires geometric understanding and precise OCR parsing of mathematical formulas. Furthermore, many Chinese problems comprise multiple interconnected subproblems requiring simultaneous resolution. All solutions of the problems also maintain consistent formatting, ensuring structural uniformity across the dataset.

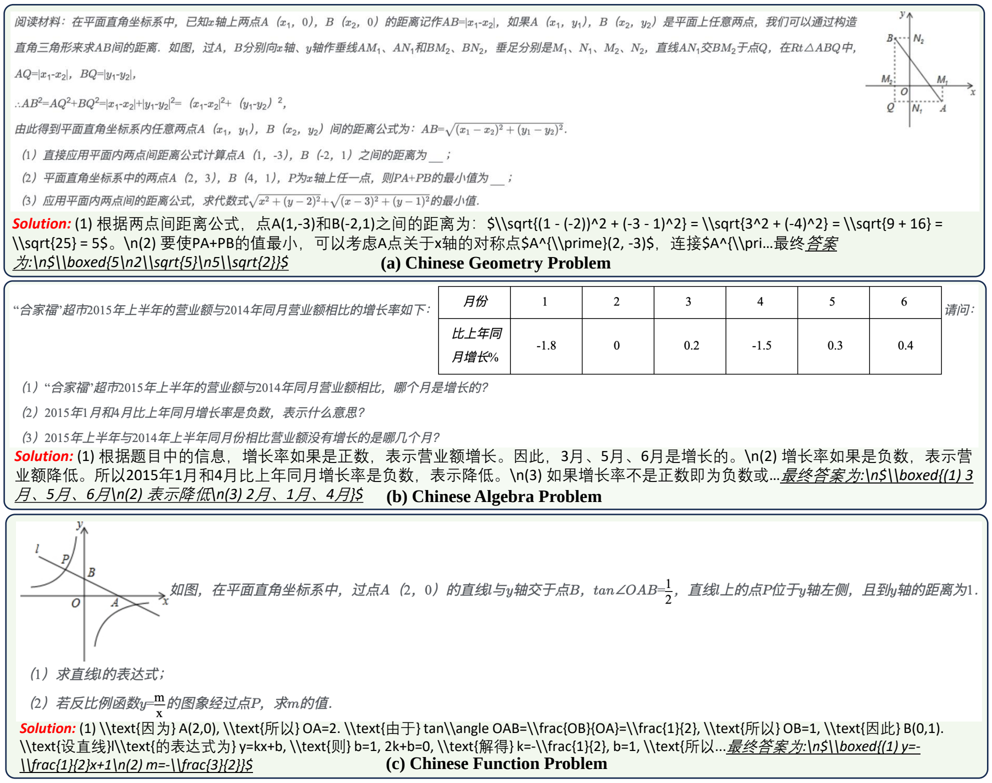


Figure 2. Chinese vision-form mathematical problem-solving samples in VisionMath-IT.

D. More Qualitative Examples of Vision-Form Problem-Solving

Figures 3-6 present comparative evaluations of representative MLLMs, *i.e*. open-source Qwen2-VL [1], closed-source GPT-4o [2] and our VisionMath, across English and Chinese mathematical problem-solving instances: Figure 3 (English plane geometry problem), Figure 4 (English solid geometry problem), Figure 5 (English function problem), and Figure 6 (Chinese function problem). The qualitative analysis reveals systematic errors in existing approaches: In Figure 3's plane geometry problems, both open-source Qwen2-VL and closed-source GPT-4o exhibit flawed geometric reasoning, Qwen2-VL incorrectly deduces *"... Since OA = OB, we have angle OAB = angle OBA = 36° ... "* while GPT-4o erroneously states *"... In triangle OBC, angle OBC is 90 degrees, and angle AOB is 36 degrees .. "*. Figure 4 demonstrates algebraic computation errors like GPT-4o's *"... Total volume ≈ 162 × 3.14159 ≈ 509.65 ..."*. Figure 5 exposes foundational mathematical misconceptions, as seen in Qwen2-VL's flawed definition *"... The amplitude of the function is the distance from the maximum value to the minimum value ..."*. Figure 6 reveals perceptual failures in Chinese function analysis, where both models incorrectly derive *"... T = 2 × (7 - 1) = 12 ..." and "... The full period T is: T=2×6=12 ..."* through invalid reasoning paths. In contrast, our approach, with enhanced perception of problem statements, structured understanding of math figures, and strengthened grasp of their properties, helps the model avoid these errors and produce correct answers.

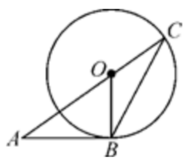
References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3

[2] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

Problem Image

As shown in the figure, AB is tangent to circle O at point B, and the extended line of AO intersects circle O at point C. Connect BC, if angle A = 36.0, then angle C is equal to ()



A.36° B.54° C.60° D.27°

Qwen2-VL Answer:
Since AB is tangent to circle O at point B, we have angle ABO = 90°. Therefore, angle AOB = 90° - 36° = 54°. *Since OA = OB, we have angle OAB = angle OBA = 36°. Therefore, angle C = 180° - angle AOB - angle OAB = 180° - 54° - 36° = 90°.* Therefore, angle C = 90° - angle AOB = 90° - 54° = 36°. Therefore, angle C = 36°. *Therefore, the answer is A.*

GPT-4o Answer:
1. Recognize that AB is tangent to the circle at point B, which means that angle OBA is 90 degrees because the radius OB is perpendicular to the tangent at the point of tangency.
2. Since AO is a line segment from the center of the circle to the point A, and it is extended to intersect the circle at point C, we can consider triangle OBC.
3. In triangle OBC, angle OBC is 90 degrees (as OB is the radius and AB is the tangent).
4. Given that angle A = 36 degrees, we need to find angle C.
5. Since AB is tangent to the circle at B, and AO is extended to intersect the circle at C, angle AOC is a straight line, making it 180 degrees.
6. *In triangle OBC, angle OBC is 90 degrees, and angle AOB is 36 degrees* (since angle A is given as 36 degrees).
7. The sum of angles in triangle OBC is 180 degrees. Therefore, angle C can be found by subtracting the sum of the other two angles from 180 degrees.
8. Angle C = 180 degrees - angle OBC - angle AOB = 180 degrees - 90 degrees - 36 degrees = 54 degrees

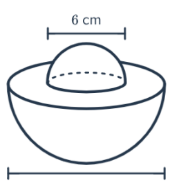
Therefore, the correct option is B.

VisionMath Answer:
Since AB is tangent to circle O at point B, we have angle ABO = 90 °. Also, since angle A = 36 °, we can find angle BOA = 54 °. *Since OB = OC, we have angle C = angle OBC. Now, angle BOA = angle C + angle OBC = 2angle C.* Therefore, angle C = 27 °. So, *the correct answer is $\\boxed{D}$.*

Figure 3. Qualitative comparison of representative MLLMs on the English vision-form plane geometry problem.

Problem Image

The composite solid consists of two hemispheres stuck together. The diameter of large hemisphere is 12 centimeters and small hemisphere is 6 centimeters. Find the volume of the solid. Round your answer to two decimal places.



Qwen2-VL Answer:
The volume of the large hemisphere is *1/2 × 4/3 × π × 6^3 = 452.39 cm^3*, and the volume of the small hemisphere is *1/2 × 4/3 × π × 3^3 = 42.41 cm^3.* Therefore, *the volume of the composite solid is 452.39 + 42.41 = 494.80 cm^3.*

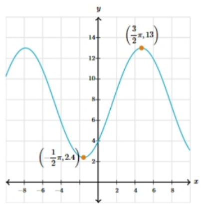
GPT-4o Answer:
To find the volume of the composite solid, we need to calculate the volume of each hemisphere and then add them together.
1. *Calculate the volume of the large hemisphere:*
   *- Volume of a hemisphere = $(2/3)\pi r^3$ = $(2/3)\pi(6)^3$ = $(2/3)\pi(216)$ = (2/3) * 216π = 144π*
2. *Calculate the volume of the small hemisphere:*
   *- Volume of the small hemisphere = $(2/3)\pi(3)^3$ = $(2/3)\pi(27)$ = (2/3) * 27π = 18π*
3. Add the volumes of the two hemispheres: - Total volume = = 144π + 18π = 162π
4. Convert the total volume to a numerical value: *- Total volume ≈ 162 * 3.14159 ≈ 509.65*

VisionMath Answer:
Step 1: Find the radius of the large hemisphere. The diameter is 12 cm, so the radius is 12 cm / 2 = 6 cm.
Step 2: Find the radius of the small hemisphere. The diameter is 6 cm, so the radius is 6 cm / 2 = 3 cm.
Step 3: Calculate the volume of the large hemisphere. *For the large hemisphere, the volume is = $(2/3)\pi(6)^3$ = $(2/3)\pi(216)$ = 144π cm^3.*
Step 4: Calculate the volume of the small hemisphere, *the volume of the small hemisphere is $(2/3)\pi(3)^3$ = $(2/3)\pi(27)$ = 18π cm^3.*
Step 5: Add the volumes of the two hemispheres to find the total volume of the composite solid. Total volume = 144π cm^3 + 18π cm^3 = 162π cm^3.
Step 6: Calculate the numerical value and round to two decimal places. 162π cm^3 ≈ 508.94 cm^3.
Therefore, *the answer is $\\boxed{508.94 cm³}$.*

Figure 4. Qualitative comparison of representative MLLMs on the English vision-form solid geometry problem.

Problem Image

Below is the graph of a trigonometric function. It has a minimum point at $\left(-\frac{1}{2}\pi, 2.4\right)$ and a maximum point at $\left(\frac{3}{2}\pi, 13\right)$. What is the amplitude of the function?



Qwen2-VL Answer:

The amplitude of the function is the distance from the maximum value to the minimum value. From the graph, we can see that the maximum value is 13 and the minimum value is 2.4. Therefore, *the amplitude is 13 - 2.4 = 10.6.*

GPT-4o Answer:

1. Identify the midline value: The midline value is given as 1.2.
2. Identify the minimum value: The minimum value is given as -3.4.
3. Calculate the distance between the midline and the minimum point:
 \[\text{Distance} = \mid \text{Midline value} - \text{Minimum value} \mid\]
 \[\text{Distance} = \mid 1.2 - (-3.4) \mid\]
\[\text{Distance} = \mid 4.6 \mid \]
The amplitude of the function is half of this distance: \[\text{Amplitude} = \frac{4.6}{2} = 2.3\]
Therefore, *the amplitude of the function is 2.3.*

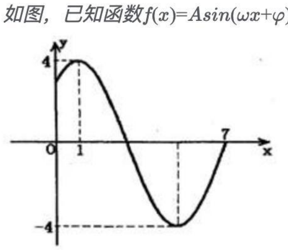
VisionMath Answer:

Step 1: Identify the minimum and maximum points of the function from the graph. *The minimum point is \\((- \\frac{1}{2}\\pi, 2.4)\\) and the maximum point is \\((\\frac{3}{2}\\pi, 13)\\).*
Step 2: Calculate the vertical distance between the minimum and maximum points. This distance represents the difference between the maximum and minimum values of the function. The vertical distance is 13 - 2.4 = 10.6.
Step 3: Determine the amplitude of the function. *The amplitude is half the vertical distance between the maximum and minimum points.* Therefore, the amplitude is 10.6 / 2 = 5.3.
Therefore, *the answer is $\\boxed{5.3}$.*

Figure 5. Qualitative comparison of representative MLLMs on the English vision-form function problem.

Problem Image

如图，已知函数$f(x)=A\sin(\omega x+\varphi)(A>0,\omega>0,0<\varphi<\pi)$,的部分图象如图所示，则函数$f(x)$的解析式为___________ .



Qwen2-VL Answer:

由可知，A=4，周期*T=2×(7-1)=12*，所以ω=π/6，又因为f(1)=4，所以4sin(π/6+ϕ)=4，所以ϕ=π/3，所以*f(x)=4sin(π/6x+π/3)*。

GPT-4o Answer:

1. **Amplitude \(A \)**: From the graph, the maximum value of \(f(x) \) is 4 and the minimum value is -4. Therefore, the amplitude \(A \) is: \[A = 4\]
2. **Period \(T \) and Angular Frequency \(\omega \)**: The period \(T \) is the distance between two consecutive peaks or troughs. From the graph, the distance between the peak at \(x = 1 \) and the next trough at \(x = 7 \) is 6 units. Since this distance represents half of the period (from peak to trough), *the full period \(T \) is: \[T = 2 \times 6 = 12\].* The angular frequency \(\omega \) is related to the period by the formula: \[\omega = \frac{2\pi}{T} = \frac{2\pi}{12} = \frac{\pi}{6}\]
3. **Phase Shift \(\varphi \)**: The phase shift \(\varphi \) can be determined by noting the horizontal shift of the sine function. The sine function \(\sin(\omega x + \varphi) \) reaches its maximum value at \(\omega x + \varphi = \frac{\pi}{2} \).
...
Therefore, the analytical expression of the function *\(f(x) \) is:\[f(x) = 4 \sin\left(\frac{\pi}{6} x + \frac{\pi}{3}\right)\]*

VisionMath Answer:

1. **振幅A:**从图像中可以看出，函数的峰值为4，简单为-4，因此振幅$A = \\frac{4 - (-4)}{2} = 4$。
2. **周期T：**图像中，*相邻两个波峰之间的距离为8因此，周期$T = 8$。*
3. **角频率 ω:** 周期 T 与角频率 ω 的关系为 $T = \\frac{2\\pi}{\\omega}$，所以 $\\omega = \\frac{2\\pi}{T} = \\frac{2\\pi}{8} = \\frac{\\pi}{4}$。
4. **相位 φ:** 从图像中可以看出，当 x = 1 时，函数值为 4，即 $f(1) = 4$。将已知条件代入函数表达式 $f(x) = 4sin(\\frac{\\pi}{4}x + \\phi)$，得到 $4 = 4sin(\\frac{\\pi}{4} + \\phi)$。由于 $0 < \\phi < \\pi$，所以 $\\frac{\\pi}{4} + \\phi = \\frac{\\pi}{2}$，解得$\\phi = \\frac{\\pi}{4}$。
最终答案为:\n*$\\boxed{$f(x) = 4sin(\\frac{\\pi}{4}x + \\frac{\\pi}{4})$}$*

Figure 6. Qualitative comparison of representative MLLMs on the Chinese vision-form function problem.