MAVIS: MATHEMATICAL VISUAL INSTRUCTION TUNING WITH AN AUTOMATIC DATA ENGINE

Renrui Zhang[1*†], **Xinyu Wei**[3*], **Dongzhi Jiang**[1], **Ziyu Guo**[2], **Yichi Zhang**[3], **Chengzhuo Tong**[4]
Jiaming Liu[3], **Aojun Zhou**[1], **Shanghang Zhang**[3], **Peng Gao**[4], **Hongsheng Li**[1,5‡]

[1]CUHK MMLab & [2]MiuLar Lab [3]Peking University
[4]Shanghai AI Laboratory [5]CPII under InnoHK
`renruizhang@link.cuhk.edu.hk, allen_wei@stu.pku.edu.cn`

[*] Equal contribution [†] Project lead [‡] Corresponding author

A APPENDIX

A.1 RELATED WORK

Visual Instruction Tuning. The advancement of large language models (LLMs) (Brown et al., 2020; Jiang et al., 2024; Touvron et al., 2023b; Chiang et al., 2023) with instruction tuning has significantly enhanced zero-shot capabilities across a range of tasks. Drawing inspiration from this, LLaMA-Adapter series (Zhang et al., 2024a; Gao et al., 2023b; Han et al., 2023) propose a zero-initialized attention mechanism to align frozen vision encoders (Radford et al., 2021) with LLaMA (Touvron et al., 2023a) for multi-modal learning. LLaVA series (Liu et al., 2023b;a) employ a linear projector for vision-language alignment, establishing visual instruction tuning as a standard training approach in the multi-modal field. Flamingo (Alayrac et al., 2022) and OpenFlamingo (Awadalla et al., 2023) have honed visual representation by integrating a cross-attention resampler with vision encoders. SPHINX series (Gao et al., 2024; Lin et al., 2023; 2025) and MR-MLLM (Wang et al., 2024a) utilize a blend of visual encoders to make the LLM cognizant of various image aspects. InternVL series (Chen et al., 2024; Dong et al., 2024; Team, 2023) employ a large vision encoder and Q-Former (Li et al., 2022) to incorporate high-quality visual information through a multi-stage training methodology. LLaVA-NexT (Liu et al., 2024a; Li et al., 2024a;b) further introduces the 'AnyRes' technique to manage images at any given resolution, and LLaVA-NexT-Interleave (Li et al., 2024c) extends the scope widely to interleave multi-image settings. There are also recent efforts to apply visual instruction tuning to 3D (Guo et al., 2023; Xu et al., 2023; Guo* et al., 2024; Tang et al., 2025), video (Li et al., 2023; Fu et al., 2024), reasoning Guo et al. (2025); Jiang et al. (2025); Peng et al. (2024), and robotics Jia et al. (2024); Liu et al. (2024b) scenarios. Despite the impressive strides made in both model capability and training efficiency by multi-modal large language models (MLLMs) through visual instruction tuning, there is currently no MLLM specifically designed for mathematical problem-solving, nor a substantial dataset available for such purposes in the open-source community. In this paper, we mitigate the issue by proposing MAVIS with high-quality mathematical visual datasets and training paradigms.

Mathematics in Large Models. Recent research has predominantly concentrated on text-only mathematical problem-solving using LLMs. MAmmoTH (Yue et al., 2023; 2024) has compiled extensive collections of mathematical problems, training LLMs using the reasoning processes described in solutions. MetaMATH (Yu et al., 2023) has expanded upon this by rewriting existing problems to create a larger dataset. MathCoder (Wang et al., 2024b) and ToRA (Gou et al., 2023) introduced a tools agent approach, employing Python code and symbolic resolvers during the training phase, significantly outperforming traditional models that rely on text-only mathematical reasoning. However, in the multi-modal field, despite the introduction of several datasets such as Geometry3K (Lu et al., 2021), GeoQA (Chen et al., 2021b), UniGeo (Chen et al., 2022), UniMath (Liang et al., 2023), and GeomVerse (Kazemi et al., 2023), aiming at enhancing the performance of MLLMs in solving graphical mathematical problems, these datasets are quite limited in scale and domain. Based on these datasets, G-LLaVA (Gao et al., 2023a) has developed superior capabilities for understanding graphical geometries but struggles with mathematical problems in other domains. The comprehensive benchmark MathVerse (Zhang et al., 2024b) has also highlighted the existing MLLMs' unsatisfactory capacity for encoding visual diagrams in diverse mathematical domains. Therefore, there is a pressing need for the development of more robust encoders for mathematical images and the tuning of MLLMs with mathematical visual instructions, for which we propose MAVIS to address the challenges.

A.2 HUMAN EVALUATION OF MAVIS-INSTRUCT

To assess the dataset's coverage, validity, and quality, human verification is employed. The creation process of our MAVIS-Instruct dataset can be broadly categorized into two approaches:

- **GPT-generated:** This method leverages GPT-4 to generate new problems (including questions, rationales, and answers) based on existing problems with diagrams. While this approach produces fluent, human-like sentences, it may be influenced by the inherent capabilities and occasional instability of GPT-4V.
- **Data Engine:** As the main source of our mathematical visual data, this method utilizes the custom automatic data engine to generate new problems (including diagrams, questions, rationales, and answers), without relying on GPT models. It guarantees 100% correctness due to the use of rigorous templates, though it may occasionally exhibit rigid expressions.

Specifically, we evaluate four aspects(Diagram, Question, Rationale and Answer) of each problem using seven metrics. Each metric is scored on a scale of 1 to 3, where 1 denotes *poor*, 2 denotes *moderate*, and 3 denotes *good*. The human evaluation results are shown in Figure 1 and score statistics are shown in Figure 2. In addition, we also showcase some specific examples in Figure 3 and Figure 4. We analyze each aspect as follows:

- **Diagram:** The diagrams in GPT-generated problems are directly collected from existing sources with rigorous human filtering, ensuring high quality, resulting in scores close to 3. In contrast, for rule-based problems, the diagrams are drawn accurately using Python code driven by our data engine, which guarantees correctness. However, these diagrams may lack alignment with human aesthetic preferences, as indicated by 3% of them receiving an appearance score of 1.
- **Question:** Regarding the questions, both GPT-generated and rule-based problems display a high degree of accuracy in aligning with the diagram elements. This is attributed to the well-crafted prompts used with GPT-4 and the meticulous template design of the data engine. Nevertheless, rule-based questions may occasionally exhibit minor fluency issues, as they lack human refinement.
- **Rationale:** In terms of the rationales, most instances feature a precise and detailed chain-of-thought (CoT) reasoning process. However, in a few cases (3% receiving an accuracy score of 1), some GPT-generated rationales contain minor reasoning or calculation errors, which are inherent to GPT-4's limitations in problem-solving. These errors usually affect only one or two steps and do not compromise the overall logic. Conversely, the rule-based rationales are highly accurate due to the carefully designed data engine, although there is still room for improvement in language fluency.
- **Answer:** The answers in both methods achieve high correctness scores. For GPT-generated problems, we prompt GPT-4 to identify a known condition from the original problems as the answer. Similarly, for rule-based problems, we randomly select a known attribute from the generated diagrams to serve as the answer.

Overall, the randomly sampled instances show that our dataset exhibits good question quality and answer accuracy.

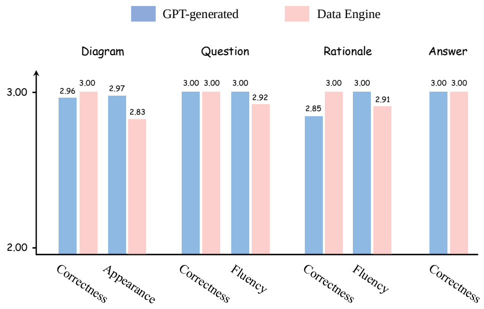


Figure 1: **Human Evaluation Results** on 200 randomly sampled problems in MAVIS-Instruct, 100 GPT-generated and 100 Data Engine. We set three levels (1, 2, and 3) for each metric, and report average scores.

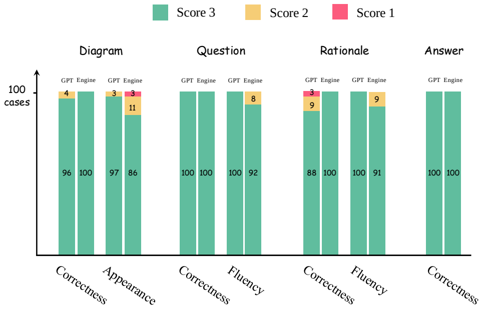
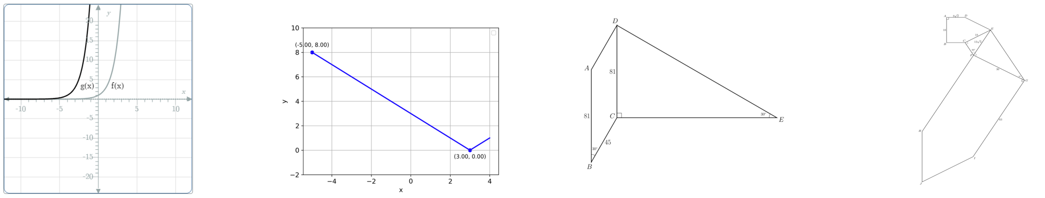


Figure 2: **Human Evaluation Statistics** on 200 randomly sampled problems in MAVIS-Instruct, 100 GPT-generated and 100 Data Engine. We count the numbers of three score levels (1, 2, and 3) for each metric.



GPT-generated Date Engine Date Engine Date Engine

Appearance Score: 3 Appearance Score: 3 Appearance Score: 3 Appearance Score: 1

Figure 3: **Diagram Examples in MAVIS-Instruct.** The first three diagrams showcase superior correctness and appearance, while a small portion of Data Engine generated diagrams (3%) are not aligned with human preference, e.g., the fourth diagram.

A.3 ABLATION STUDY

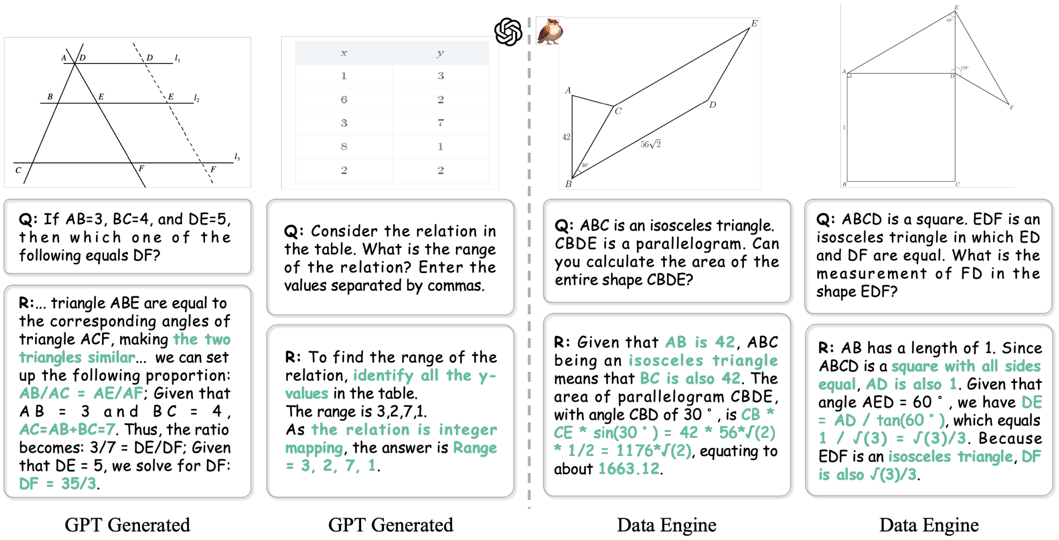
A.3.1 MAVIS-CAPTION

To validate the enhancement of Math-CLIP's **diagram perception capability**, we sampled 100 validation diagram-caption pairs and computed their cosine similarity using both CLIP and Math-CLIP. The results, as shown in Table 3, indicate that Math-CLIP encodes more discriminative diagram features. Additionally, the attention visualization in Figure **??**(a) of the main paper further demonstrates that Math-CLIP captures mathematical visual elements within diagrams more effectively, highlighting the efficacy of MAVIS-Caption.

To validate the role of MAVIS-Caption in second-stage training, we present both quantitative and qualitative results for diagram captioning on the same 100 validation pairs in the first column of Table 4. The use of MAVIS-Caption significantly enhances the **diagram understanding capability**. This shows that MAVIS-Caption helps the LLM generate accurate captions from diagrams, improving its ability to comprehend each visual token from Math-CLIP and align visual elements with textual descriptions. We also evaluated MAVIS's performance on MathVerse without second-stage training, as shown in the second column of Table 4. Without MAVIS-Caption training, the CoT reasoning quality of MAVIS-7B is somewhat compromised. This suggests that training the model in diagram captioning improves its **mathematical expression capability**, enabling it to produce language expressions that align with mathematical concepts. This foundational skill supports the generation of subsequent CoT reasoning steps.

A.3.2 MAVIS-INSTRUCT

Redundant Text When curating questions for MAVIS-Instruct, we minimize the redundant content within the question texts, which refers to the directly observable content in the diagram, e.g., the presence of shapes or intersection points of functions. Such information is repetitive to visual



x	y
1	3
6	2
3	7
8	1
2	2

Q: If AB=3, BC=4, and DE=5, then which one of the following equals DF?

R:... triangle ABE are equal to the corresponding angles of triangle ACF, making the two triangles similar... we can set up the following proportion: AB/AC = AE/AF; Given that AB = 3 and BC = 4, AC=AB+BC=7. Thus, the ratio becomes: 3/7 = DE/DF; Given that DE = 5, we solve for DF: DF = 35/3.

GPT Generated

Q: Consider the relation in the table. What is the range of the relation? Enter the values separated by commas.

R: To find the range of the relation, identify all the y-values in the table.
The range is 3,2,7,1.
As the relation is integer mapping, the answer is Range = 3, 2, 7, 1.

GPT Generated

Q: ABC is an isosceles triangle. CBDE is a parallelogram. Can you calculate the area of the entire shape CBDE?

R: Given that AB is 42, ABC being an isosceles triangle means that BC is also 42. The area of parallelogram CBDE, with angle CBD of 30 ° , is CB * CE * sin(30 °) = 42 * 56*√(2) * 1/2 = 1176*√(2), equating to about 1663.12.

Data Engine

Q: ABCD is a square. EDF is an isosceles triangle in which ED and DF are equal. What is the measurement of FD in the shape EDF?

R: AB has a length of 1. Since ABCD is a square with all sides equal, AD is also 1. Given that angle AED = 60 ° , we have DE = AD / tan(60 °), which equals 1 / √(3) = √(3)/3. Because EDF is an isosceles triangle, DF is also √(3)/3.

Data Engine

Figure 4: **Accurate Rationale Examples in MAVIS-Instruct.** Most GPT-generated and Data Engine-generated rationales ensure correctness.

Table 3: **Diagram Perception Enhancement by Math-CLIP**, using MAVIS-Caption in the first stage. We calculate the average cosine similarity among 100 validation diagram-caption pairs.

Vision Encoder	Matched Pair ↑	Unmatched Pair ↓
CLIP	0.22	0.24
Math-CLIP	**0.83**	**0.17**

Table 4: **Diagram Understanding Enhancement** and **Mathematical Expression Enhancement** in LLM using MAVIS-Caption in the second Stage. We compare the METEOR and CIDEr scores for diagram captioning on 100 validation samples, as well as the accuracy and CoT evaluation results on MathVerse, both with and without the MAVIS-Caption training.

Training Data	Diagram-Caption Pairs		MathVerse	
	METEOR	CIDEr	Acc (%)	CoT-E (%)
w MAVIS-Caption	23.7	161.3	28.4	35.2
w/o MAVIS-Caption	14.0	69.4	25.6	32.8

Table 1: **Statistics of MAVIS-Caption.**

Statistic	**Number**
Total Captions	
- Total number	588K
- Average length (words)	62.85
- Average length (characters)	339.68
- Vocabulary size	418
Plane Geometry	
- Total number	299K (50.9%)
- Average length (words)	69.77
- Average length (characters)	385.85
- Vocabulary size	195
Analytic Geometry	
- Total number	77K (13.1%)
- Average length (words)	39.64
- Average length (characters)	210.10
- Vocabulary size	158
Function	
- Total number	212K (36.0%)
- Average length (words)	61.48
- Average length (characters)	321.46
- Vocabulary size	149

Table 2: **Subject Distribution of MAVIS-Instruct.**

Statistic	**Number**
Total questions	834K
- Multiple-choice questions	615K (62.4%)
- Free-form questions	218K (37.6%)
Data Engine Generated Problems	582K
- Geometry questions	466K (80.0%)
- Function questions	116K (20.0%)
Data Engine Captions Annotated by GPT-4	51K
- Geometry questions	30K (58.8%)
- Function questions	21K (41.2%)
Manual Collection Augmented by GPT-4	83K
- Geometry questions	72K (86.5%)
- Function questions	11K (13.5%)
Existing Datasets Augmented by GPT-4	118K
- Geometry questions	118K (100.0%)
- Function questions	0 (0%)
Number of unique images	611K (73.3%)
Number of unique questions	804K (96.5%)
Number of unique answers	675K (81.0%)
Average question length	44.60
Average answer length	62.82

components, and may assist MLLMs in bypassing the process of diagram interpretation, thereby harming their related skills. By mostly avoiding redundant texts in MAVIS-Instruct, our data enforces MLLMs to learn stronger **diagram interpretation capabilities**. In Table 5, we add redundant texts (diagram captions) to the Data Engine Generated Problems for training, leading to expected performance drop.

CoT Rationales For each instance in MAVIS-Instruct, we incorporate detailed rationales for problem-solving, either generated by GPT-4 or our rule-based data engine. In Table 6, we remove all intermediate rationales of each problem in MAVIS-Instruct, and train the model to directly output the final answer. As shown, both the CoT evaluation and accuracy scores are degraded. This demonstrates the significance of our rationale annotations, which effectively improves the CoT **reasoning capabilities** of MLLMs.

Table 5: **Diagram Interpretation Enhancement for MLLM**, using MAVIS-Instruct in the third stage. We compare the results by adding redundant texts (diagram captions) to the Data Engine Generated Problems within MAVIS-Instruct.

MAVIS-Instruct	MathVerse	GeoQA	FunctionQA
w/o Redundant Texts	**28.4**	**68.3**	**50.0**
w Redundant Texts	26.5	66.5	48.4

Table 6: **Reasoning Capability Enhancement for MLLM**, using MAVIS-Instruct in the third stage.

Training Data	MathVerse	
	Acc	CoT-E
w Rationales	**28.4**	**35.2**
w/o Rationales	25.2	26.6

A.3.3 COMPARED TO GENERAL VISUAL INSTRUCTION DATA

Since Mammoth-2 is a highly capable LLM for mathematical tasks, one possible question is whether simply integrating a vision encoder into Mammoth-2 and training it with conventional visual instruction tuning data would suffice for effectively solving visual-based mathematical problems. To compare MAVIS data with other visual instruction tuning datasets and investigate the specific benefits of MAVIS data in Mammoth-2 (7B), we conduct an ablation study. We utilize the data from LLaVA-NeXT (558K for pre-training and 760K for fine-tuning) and compare it with our MAVIS data (558K MAVIS-Caption for pre-training and 834K MAVIS-Instruct for fine-tuning). Performance is evaluated using the accuracy metric on MathVerse, excluding the DPO training stage for fairness.

Based on the results presented in Table 7, we make the following observations:

Table 7: Ablation study results for comparison between MAVIS Data and other visual instruction tuning data. The first row in the table represents the original LLaVA-NeXT-8B.

Visual Encoder	**LLM**	**Pre-training**	**Fine-tuning**	**MathVerse Acc (%)**
CLIP	LLaMA-3 (8B)	LLaVA data	LLaVA data	15.6
CLIP	Mammoth-2 (7B)	LLaVA data	LLaVA data	18.3
CLIP	Mammoth-2 (7B)	LLaVA data	**MAVIS-Instruct**	25.7
CLIP	Mammoth-2 (7B)	**MAVIS-Caption**	**MAVIS-Instruct**	26.4
Math-CLIP	Mammoth-2 (7B)	**MAVIS-Caption**	**MAVIS-Instruct**	27.5

1. **Mammoth-2 vs. LLaMA-3:** Mammoth-2 achieves a +2.7 improvement in accuracy compared to LLaMA-3, highlighting its prior knowledge and inherent capability in mathematical problem solving.
2. **Impact of MAVIS-Instruct:** Fine-tuning with MAVIS-Instruct significantly enhances performance by +7.4, underscoring the substantial advantage of our dataset for mathematical reasoning tasks compared to general visual instruction datasets.
3. **MAVIS-Caption and Math-CLIP:** Using MAVIS-Caption for pre-training and employing the Math-CLIP encoder further boosts performance, leading to enhanced mathematical visual perception and reasoning capabilities. Overall, our MAVIS data contributes a +9.2 improvement in accuracy over Mammoth-2 trained with LLaVA data.

A.3.4 PERFORMANCE ACROSS DIFFERENT SUBJECTS

Although MAVIS-Instruct contains a substantial number of high-quality solid geometry problems that were manually curated, our data engine only generates plane geometry and function problems. Therefore, we aim to evaluate the performance of the MAVIS model across different mathematical domains, specifically plane geometry, functions, and solid geometry. We provide the detailed subject scores of MAVIS-7B on MathVerse, comparing the CoT evaluation score (note that the subject-level accuracy scores are not publicly released) with other models on the official leaderboard.

Table 8: Performance comparison across different models on Plane Geometry, Solid Geometry, and Functions of MathVerse evaluation tasks.

Model	**All (CoT-Eval)**	**Plane Geometry**	**Solid Geometry**	**Functions**
LLaVA-NeXT	17.2	15.9	19.6	23.1
ShareGPT4V	17.4	16.9	15.0	20.2
SPHINX-MoE	22.8	24.5	15.8	19.5
InternLM-XC2	25.9	26.2	20.1	23.7
MAVIS-7B	**35.2**	**37.1**	**28.9**	**31.0**

The results shown in Table 8 demonstrate that our model achieves leading performance across all three subjects. Notably, its proficiency in plane geometry and functions can be attributed to the training with our meticulously curated MAVIS dataset. Additionally, for solid geometry, which shares similarities with plane geometry in both visual appearance and reasoning process, we believe that our model effectively generalizes its learned knowledge and reasoning capabilities, leading to enhanced performance in this domain as well.

A.3.5 SYNTHETIC DATA VS REAL DATA

In MAVIS-Instruct, we integrate both synthetic problems generated by the data engine (633K, 76%) and real-world problems augmented with GPT (201K, 24%). The synthetic data is composed of both geometry and function problems, while the real-world data primarily focuses on geometry. We conduct an ablation study to assess the contributions of these data components, excluding the DPO training stage to ensure fairness.

Table 9: Ablation study of synthetic and real data contributions to MAVIS-7B's performance.

Synthetic Data	Real-world Data	MathVerse Acc (%)	GeoQA	FunctionQA	MMMU-Math
✓	–	22.6	44.2	37.1	34.6
–	✓	24.3	66.4	25.8	29.8
✓	✓	27.5	66.7	40.3	39.2

The results shown in Table 9 indicate that the two data sources exhibit complementary characteristics, both playing a crucial role in achieving the final performance. Specifically, synthetic data significantly enhances the results on FunctionQA and MMMU-Math, as these benchmarks include a substantial proportion of function-related problems. Conversely, real-world data has a greater impact on GeoQA, given its stronger alignment with the geometry-focused nature of this benchmark.

A.3.6 DATA SCALING

A good instruction tuning dataset should exhibit the characteristic of data scaling: as the dataset size increases, the model trained on it should demonstrate progressively better performance. To verify that MAVIS-Instruct possesses this property, we conduct an ablation study on the 834K MAVIS-Instruct dataset by randomly sampling 25%, 50%, and 75% of the data for instruction tuning, excluding the DPO stage. We then evaluate the models using the accuracy metric on MathVerse. The results, as shown in Table 10, indicate that the performance of MAVIS-7B consistently improves as the data scale increases. This demonstrates the promising potential of our dataset to further enhance mathematical reasoning capabilities with larger-scale utilization.

Table 10: Performance of MAVIS-7B at different data proportions.

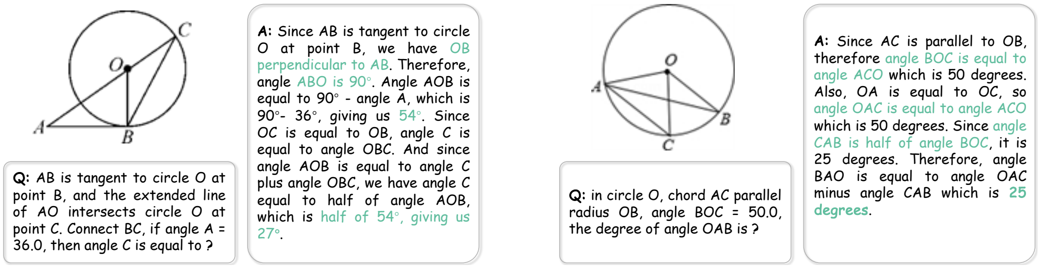
25%	50%	75%	100%
23.3	25.7	26.9	27.5

Table 11: Comparison of different training settings.

LLMs	Caption CIDEr	MathVerse Acc (%)
Frozen	79.6	26.2
Unfrozen	146.2	28.1
LoRA-based	161.3	28.4

A.3.7 GENERALIZATION ABILITY

Although our Data Engine considers as many problem types as possible, it is inherently challenging for a manually designed engine to cover all possible types of mathematical visual diagrams. To alleviate this, we incorporate real-world data by manual collection and GPT augmentation, which well enhances the generalization capablitity of MAVIS-7B. In Figure 5, we selected examples involving inscribed triangles (they cannot be generated via our data engine) from publicly available math benchmarks, and tested MAVIS-7B on these examples, where our model performs accurate reasoning.



Q: 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: Since AB is tangent to circle O at point B, we have OB perpendicular to AB. Therefore, angle ABO is 90°. Angle AOB is equal to 90° - angle A, which is 90°- 36°, giving us 54°. Since OC is equal to OB, angle C is equal to angle OBC. And since angle AOB is equal to angle C plus angle OBC, we have angle C equal to half of angle AOB, which is half of 54°, giving us 27°.

Q: in circle O, chord AC parallel radius OB, angle BOC = 50.0, the degree of angle OAB is ?

A: Since AC is parallel to OB, therefore angle BOC is equal to angle ACO which is 50 degrees. Also, OA is equal to OC, so angle OAC is equal to angle ACO which is 50 degrees. Since angle CAB is half of angle BOC, it is 25 degrees. Therefore, angle BAO is equal to angle OAC minus angle CAB which is **25 degrees**.

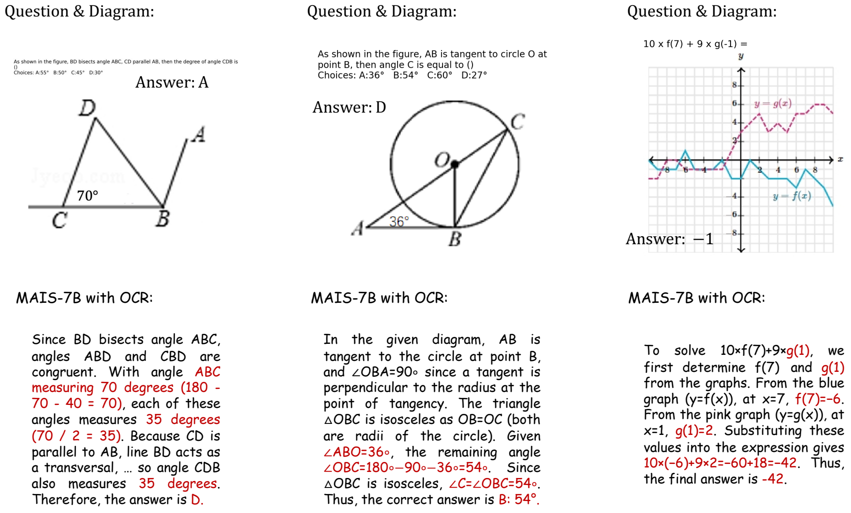
Figure 5: Examples for MAVIS-7B solving math problems with circumcircles of triangles.

A.3.8 ENHANCING VISION-ONLY PERFORMANCE

To evaluate the impact of OCR datasets on MAVIS-7B's performance, we conducted an ablation study focusing on vision-only and vision-dominant problems in MathVerse. These problems require the model to interpret question texts rendered directly in diagrams, thus relying heavily on OCR capabilities. MAVIS-7B, however, was initially not trained with OCR-specific datasets, limiting its performance in these tasks.

In contrast, generalist models like LLaVA-NeXT include extensive OCR datasets such as OCRVQA, DocVQA, and SynDog-EN, which significantly enhance their OCR capabilities. To bridge this gap, we incorporated OCR datasets (OCRVQA and DocVQA) in our third-stage instruction tuning to improve MAVIS-7B's OCR performance.

The results, as shown in Table 12, indicate a notable improvement in vision-dominant and vision-only problems for MAVIS-7B after the inclusion of OCR datasets, highlighting the potential of better OCR integration for further boosting its performance. In Figure 6, we also showcase some failure cases of our MAVIS-7B with OCR training on vision-only problems. Although the vision-only results are improved via the OCR instruction dataset, the model still suffers from limited perception capabilities of questions and visual elements within the diagram. This indicates that the OCR capability is still the bottleneck of vision-only performance. We leave this as a future work to further enhance the OCR capabilities of MAVIS for mathematical visual elements.



Question & Diagram:

As shown in the figure, BD bisects angle ABC, CD parallel AB, then the degree of angle CDB is ()
Choices: A:55° B:50° C:45° D:30°

Answer: A

MAIS-7B with OCR:

Since BD bisects angle ABC, angles ABD and CBD are congruent. With angle ABC measuring 70 degrees (180 - 70 - 40 = 70), each of these angles measures 35 degrees (70 / 2 = 35). Because CD is parallel to AB, line BD acts as a transversal, … so angle CDB also measures 35 degrees. Therefore, the answer is D.

Question & Diagram:

As shown in the figure, AB is tangent to circle O at point B, then angle C is equal to ()
Choices: A:36° B:54° C:60° D:27°

Answer: D

MAIS-7B with OCR:

In the given diagram, AB is tangent to the circle at point B, and ∠OBA=90∘ since a tangent is perpendicular to the radius at the point of tangency. The triangle △OBC is isosceles as OB=OC (both are radii of the circle). Given ∠ABO=36∘, the remaining angle ∠OBC=180∘−90∘−36∘=54∘. Since △OBC is isosceles, ∠C=∠OBC=54∘. Thus, the correct answer is B: 54°.

Question & Diagram:

10 x f(7) + 9 x g(-1) =

Answer: −1

MAIS-7B with OCR:

To solve 10×f(7)+9×g(1), we first determine f(7) and g(1) from the graphs. From the blue graph (y=f(x)), at x=7, f(7)=−6. From the pink graph (y=g(x)), at x=1, g(1)=2. Substituting these values into the expression gives 10×(−6)+9×2=−60+18=−42. Thus, the final answer is -42.

Figure 6: Failure cases of MAVIS-7B with OCR training on vision-only problems.

Table 12: Impact of OCR data on solving vision-only problems.

Model	LLM Size	All	Text Dominant	Text Lite	Vision Intensive	Vision Dominant	Vision Only
LLaVA-NeXT	8B	15.6	19.4	15.2	16.8	15.2	11.3
LLaVA-NeXT	110B	24.5	31.7	24.1	21.0	22.1	20.7
MAVIS-7B	7B	28.4	**41.6**	**29.5**	**27.9**	24.7	18.3
MAVIS-7B w/ OCR	7B	**28.9**	40.8	29.2	27.4	**26.2**	**21.1**

A.3.9 BASE LLM

We investigate different LLMs for the MAVIS model. As shown in Table 13, MAVIS is not very sensitive to LLM choices, and still surpasses previous models with the same LLM.

A.3.10 DIFFERENT TRAINING SETTINGS

Our training strategy is similar to LLaVA, but with key differences in the pre-training stage: we train both the projection layer and the LoRA-based LLM, whereas LLaVA only trains the projection

Table 13: **Performance Comparison using Different LLMs.** We compare the accuracy and CoT evaluation results on MathVerse.

Method	Base LLM	MathVerse	
		Acc	CoT-E
SPHINX-Plus	LLaMA2-13B	12.2	14.0
ShareGPT4V	Vicuna-13B	13.1	17.4
InternLM-XC2.	InternLM2-7B	16.5	25.9
MAVIS	LLaMA2-13B	24.5	30.7
	Vicuna-13B	24.8	30.6
	InternLM2-7B	28.0	33.8
	MAmmoTH2	**28.4**	**35.2**

layer. This design choice stems from the fundamental differences between general visual tasks and mathematical tasks:

1. For general visual tasks (e.g., LLaVA), training MLLMs typically requires the LLM to generate daily natural language responses, such as descriptive captions or instruction-following outputs. These outputs often rely on pre-existing knowledge within the pre-trained LLM. As a result, in LLaVA, there is no need to unfreeze the LLM to learn new types of outputs.
2. In contrast, for mathematical domains, LLMs need to generate math-specific responses, such as geometric descriptions, functional explanations, formulas, and theorems. These outputs often involve domain-specific knowledge not inherent in pre-trained LLMs. Given this, we incorporate learnable LoRA layers to infuse new knowledge into the LLM, enhancing its capability to produce high-quality mathematical expressions. Concurrently, we aim to prevent the LLM from overfitting to diagram captioning tasks during alignment. Therefore, using LoRA-based tuning allows us to preserve the LLM's generalizable pre-trained language knowledge while injecting specialized math-specific capabilities.

To further investigate the impact of different training settings on model performance, we conduct an ablation study comparing various LLM training settings during the alignment stage. We evaluate two tasks: the CIDEr score for diagram captioning on 100 validation samples (following the same setting as in Table 6 of the Appendix) and the accuracy score on MathVerse. The results, as shown in Table 11, indicate that the LoRA-based approach performs best, enabling MLLMs to generate high-quality mathematical captions while preserving pre-trained knowledge for improved problem-solving capabilities.

A.3.11 ENHANCING A PRE-TRAINED MLLM

To investigate whether our curated data and training techniques can improve the mathematical performance of a pre-trained large model (LLaVA-NeXT), we conducted an ablation study. Specifically, we progressively employed MAVIS-Instruct for instruction tuning, followed by DPO alignment on top of LLaVA-NeXT-8B, with both training stages performed for one epoch using a learning rate of 1×10^{-5}. The results, as shown in Table 14, demonstrate that these two continual training stages significantly enhance LLaVA-NeXT's ability to solve mathematical problems, with notable improvements across all evaluation categories.

Table 14: Performance improvement of LLaVA-NeXT-8B with MAVIS-Instruct and DPO alignment.

Model	**LLM Size**	**All**	**Text Dominant**	**Text Lite**	**Vision Intensive**	**Vision Dominant**	**Vision Only**
LLaVA-NeXT	8B	15.6	19.4	15.2	16.8	15.2	11.3
+ MAVIS-Instruct	8B	22.8	32.3	25.3	24.6	18.3	14.2
+ DPO	8B	**24.0**	**33.7**	**26.9**	**25.4**	**19.1**	**15.1**

A.4 DETAILS OF AUTOMATIC DATA ENGINE

A.4.1 DIAGRAM GENERATION

In this section, we detail the implementation specifics of the process for generating diagrams related to plane geometry, analytic geometry, and function domains.

Plane Geometry Diagram. Inspired by previous multi-hop reasoning methods (Kazemi et al., 2023; Wei et al., 2022; Nye et al., 2021), we employ an iterative generation method over logical theories to generate plane geometric images along with corresponding captions and question-answering pairs, whose complexity can be controlled across multiple axes. Specifically, we first define a set of fundamental geometric shapes in Figure 7.

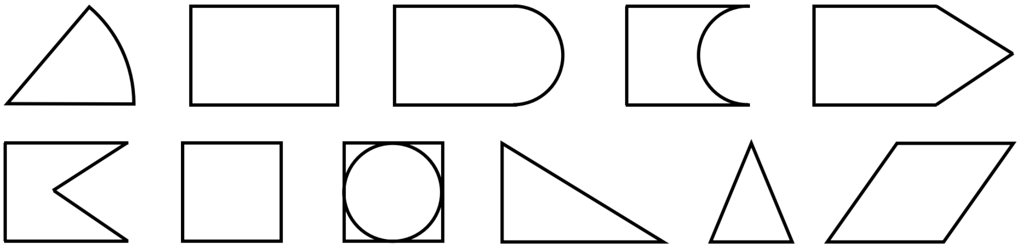


Figure 7: The set of fundamental shapes in plane geometry diagrams, whose straight edges can be extended into other basic shapes.

Within each shape, new basic shapes can be generated by extending a particular edge. For each basic shape, we initially define a meta reasoning process:

$$O_{n-1}, C^{i}_{m_{n-1}} \xrightarrow{\mathrm{E}^{i}_{m_{n-1}}} O_n, i \in [1, z], \tag{1}$$

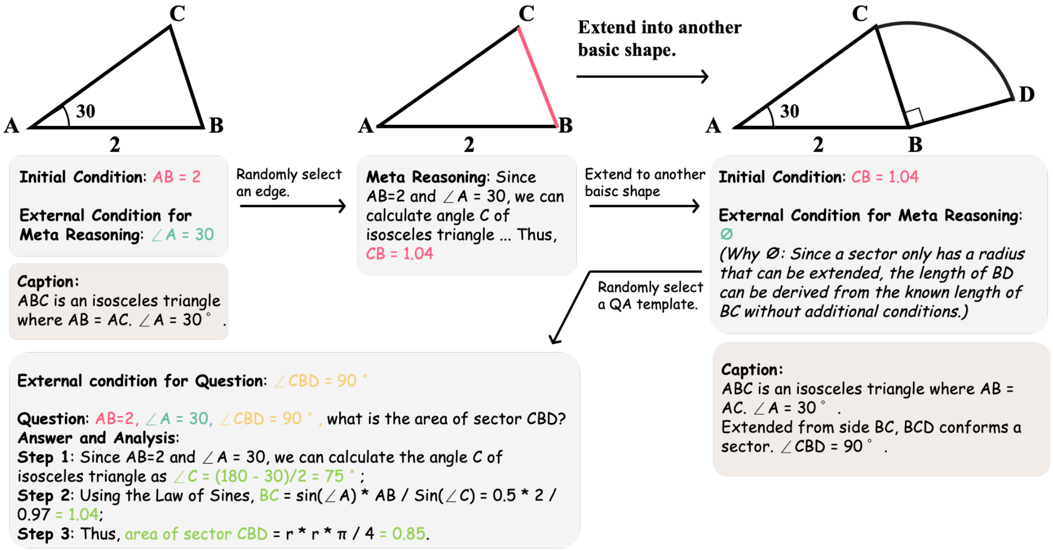
where O represents the initial side length of the shape, C_m denotes the additional conditions required to complete meta reasoning, and E_m provides a detailed explanation of the meta reasoning process. For example, when considering an isosceles triangle as the $(n-1)^{th}$ shape in a sequence, the vertex angle is still required as C_m to reason about base side length, and then to expand to the nth shape, with E_m serving as the explanation of this process. The variable z indicates that there are z sets of possible meta reasoning for the shape, n indicates the length of the generating sequence, which is also the number of hops of reasoning required to answer the question. The initial side, extend side, and additional conditions for meta-reasoning of each basic shape can be referred to in Figure 7. In the final shape, question-answering pairs pertinent to this shape can be generated as

$$O_n, C^{j}_{q_n}, Q^{j}_{n} \xrightarrow{\mathrm{E}^{j}_{q_n}} A^{j}_{n}, j \in [1, m], \tag{2}$$

where C_q represents the additional conditions required to solve the problem, while Q and A denote the question and answer, respectively. E_q refers to the detailed explanation of the solving process. The variable m indicates that there are m pairs of question-answering and corresponding detailed explanations within the shape. By applying meta reasoning to the $n-1$th shape, the initial side length of the nth shape can be deduced. Therefore, for a complex composite figure consisting of n shapes, the overall question-answering pair can be defined as follows:

$$O_1, \sum_{k=1}^{n-1} C_{m_k}, C^{j}_{q_n}, Q^{j}_{n} \xrightarrow{\mathrm{E}^{j}_{q_n}} A^{j}_{n}. \tag{3}$$

Each shape defines a sufficient number of conditions, explanations, and answers to ensure the diversity of the generated question-answering pairs. Based on the aforementioned rules, controlling the length of the generation sequence can regulate the number of reasoning steps, and controlling the type of questions can manage the knowledge required for solving the problems. Thus, we can generate questions of varying difficulty levels, which can also be illustrated in Figure 8a.


C
A
30
B
2
C
A
B
2
Extend into another basic shape.
C
D
A
30
B
2
B
Randomly select an edge.
Extend to another baisc shape
Randomly select a QA template.

Initial Condition: AB = 2

External Condition for Meta Reasoning: ∠A = 30

Caption:
ABC is an isosceles triangle where AB = AC. ∠A = 30°.

Meta Reasoning: Since AB=2 and ∠A = 30, we can calculate angle C of isosceles triangle ... Thus, CB = 1.04

Initial Condition: CB = 1.04

External Condition for Meta Reasoning: ∅

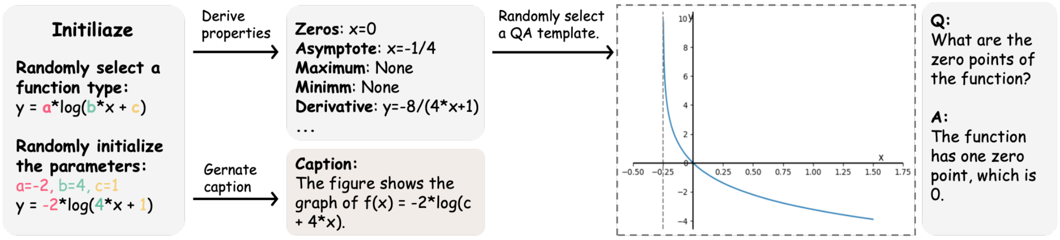
(Why ∅: Since a sector only has a radius that can be extended, the length of BD can be derived from the known length of BC without additional conditions.)

Caption:
ABC is an isosceles triangle where AB = AC. ∠A = 30°.
Extended from side BC, BCD conforms a sector. ∠CBD = 90°.

External condition for Question: ∠CBD = 90°

Question: AB=2, ∠A = 30, ∠CBD = 90°, what is the area of sector CBD?
Answer and Analysis:
Step 1: Since AB=2 and ∠A = 30, we can calculate the angle C of isosceles triangle as ∠C = (180 - 30)/2 = 75°;
Step 2: Using the Law of Sines, BC = sin(∠A) * AB / Sin(∠C) = 0.5 * 2 / 0.97 = 1.04;
Step 3: Thus, area of sector CBD = r * r * π / 4 = 0.85.

(a) A single process for generating plane geometry diagrams and corresponding question-answering pairs as well as image captions. In this example, the generation sequence length is specified as 2. Initial side length is painted in pink, C_m is painted in green, while C_q is painted in yellow. Whenever a new basic shape is generated, its caption is appended to the previous caption.



Initiliaze

Randomly select a function type:
y = a*log(b*x + c)

Randomly initialize the parameters:
a=-2, b=4, c=1
y = -2*log(4*x + 1)

Zeros: x=0
Asymptote: x=-1/4
Maximum: None
Minimm: None
Derivative: y=-8/(4*x+1)
...

Caption:
The figure shows the graph of f(x) = -2*log(c + 4*x).

Q:
What are the zero points of the function?

A:
The function has one zero point, which is 0.

(b) A single process is used for generating function diagrams along with the corresponding question-answer pairs and image captions. Once the functional expression is determined, all its properties can be directly computed, and the function plot can be generated accordingly. The caption for the function diagram simply states the functional expression.

Figure 8: The pipeline of our data engine, consisting of (a) the generation of plane geometry diagrams and (b) the generation of function diagrams.

Analytic Geometry Diagram. The image generation method for analytic geometry is relatively straightforward. First, we randomly select a range within the coordinate system: the minimum value of x is chosen as an integer between $[-12, -8]$, and the maximum value of x is chosen as an integer between $[8, 12]$; the range for y is the same as for x. Then, we define the following basic shapes: point, line segment, line, circle, ellipse, rectangle, square, polygon, and sector. During the generation process, we select a number between 1 and 4 as the number of shapes to generate. The generation rule is that **nonlinear shapes** other than points, line segments, and lines **must not overlap**.

Function Diagram. The generation of function graphs is also straightforward as shown in Figure 8b. We define the following basic functions, each with a set of parameters that can be randomly selected:

Sine Function	$y = A \cdot \sin(f \cdot x + \phi)$, where the amplitude A is a random integer between 1 and 3, the frequency f is either 1 or 2, and the phase ϕ is a random integer between 0 and 2π.
Cosine Function	$y = A \cdot \cos(f \cdot x + \phi)$, where the amplitude A is a random integer between 1 and 3, the frequency f is either 1 or 2, and the phase ϕ is a random integer between 0 and 2π.

Tangent Function	$y = A \cdot \tan(f \cdot x + \phi)$, where the amplitude A is a random integer between 1 and 3, the frequency f is either 1 or 2, and the phase ϕ is a random integer between 0 and 2π.
Polynomial Function	$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, where the degree n is a random integer between 1 and 4. The coefficients a_i are randomly selected integers ranging from -3 to 3.
piece-wise Function	piece-wise polynomial functions are divided into 2 or 3 segments, with each segment's parameters identical to those of a polynomial function.
Logarithmic Function	$y = a \cdot \log_b(c \cdot x + d)$, where the coefficient a is randomly chosen from $\{-3, -2, -1, 1, 2, 3\}$, the base b is randomly chosen from $\{2, 10, \lfloor e \rfloor\}$, the coefficient c is a random integer between 1 and 3, and the coefficient d is a random integer between 1 and 6, ensuring that $c \cdot x + d$ is positive.
Absolute Function	$y = \lvert a \cdot x + b \rvert$, where a and b are random integer between -5 and 5.

We first determine the domain range to be displayed on the function graph. For trigonometric functions, the domain is set to $[-\pi, \pi]$. For piece-wise polynomial functions, the minimum value of x is a random integer between $[-12, -8]$, and the maximum value of x is a random integer between $[8, 12]$. For other functions, the minimum and maximum values of x are random integers within the ranges of $[-6, -3]$ and $[3, 6]$, respectively. During the plotting process, we calculate the local maxima, minima, and zeros of the function by iterating through the domain. We then render the x-coordinates of these extrema and zeros on the x-axis of the function graph.

A.4.2 MAVIS-CAPTION

In this section, we detail how the captions corresponding to images in the MAVIS-Caption Dataset are generated with our automatic data engine.

Plane Geometry Caption. Based on the generation process described in Section A.4.1, when generating each shape, a caption is randomly selected from a set of captions for that shape and some connecting words are randomly added. We also randomly select some edges or angles and state their measurements in the caption. After generating the raw caption, we use GPT-3.5 to refine it, enhancing its linguistic structure and semantic diversity. An example is shown in Figure **??**.

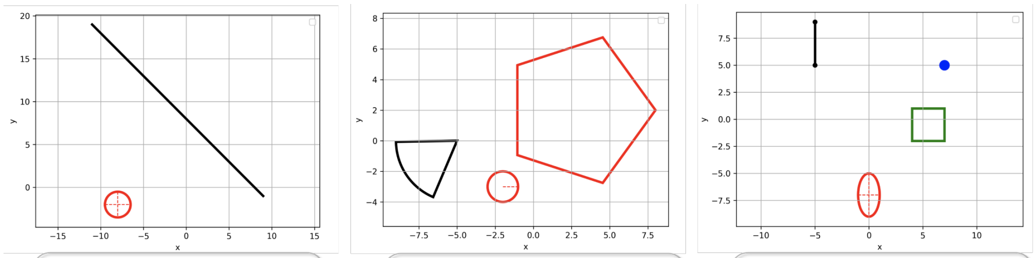
Function Caption. According to the function graph generation process described in Section A.4.1, we record the function's zeros and extrema. Additionally, we also record the function's expression and asymptotes. These attributes are incorporated into a randomly selected caption template to form the function graph's caption. Some examples are provided in Figure 10.

Analytic Geometry Caption. For each shape, we maintain a set of caption templates that describe the shape's type, coordinate position, and other attributes. In the generation process described in Section A.4.1, we select a template and randomly add some diverse connecting words to form a complete caption. Examples of some captions are shown in Figure 9.

A.4.3 MAVIS-INSTRUCT

Manual Collection Augmented by GPT-4. To complement the dataset with real-world problem-solving scenarios, we hire 8 human experts to manually collect visual math problems from various public sources[1,2,3], spanning plane geometry, analytic geometry, and function. For problems, we try to obtain their content as complete as possible, including questions, diagrams, answers, and rationales if available. The collection process consists of the following steps:

1. **Problem Collection:** We gathered problems from three public sources as comprehensively as possible, including questions, diagrams, answers, category information, and rationales where available. The problems are primarily at the high-school level, covering plane geometry and functions (including analytic geometry).

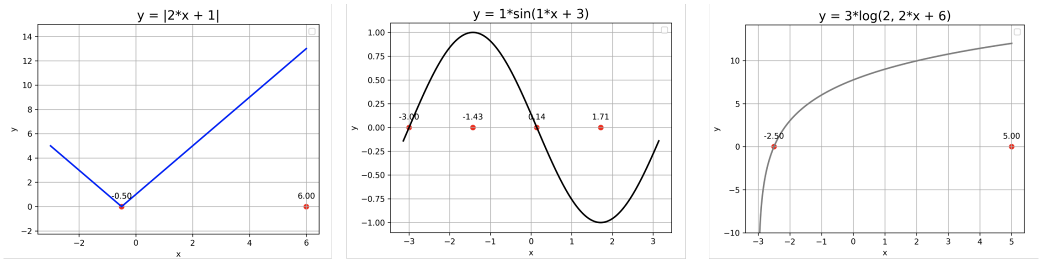


At the outset, There is an ellipse with a height of 3 and a width of 3, centered at (-8, -2). Beyond this, Crossing at (0, 8) on the y-axis and (8.0, 0) on the x-axis, a straight line exists.

At the outset, Located at (-2, -3), a circle encompasses a radius of 1. Later, A sector originates from a center at (-5, 0) and extends a radius of 4. Following this, A pentagon with points at (-1.05, -0.94), (4.55, -2.76), (8.0, 2.0), (4.55, 6.76), (-1.05, 4.94) is established.

To start with, A geometric figure, an ellipse, at (0, -7), measures 4 in height and 2 in width. There exists a point at (7, 5). Also, With its left bottom corner at (4, -2), a rectangle extends a height of 3 and a width of 3. Moreover, A line segment is established between (-5, 9) and (-5, 5).

Figure 9: Examples of analytical geometry diagram caption.



The expression for the function is y = |2*x + 1|. The zeros for the x value range [-3.0, 6.0] are found to be -0.5. The function spans a range of values from 0 at -0.5 to 13.0 at 6.0 across the interval [-3.0, 6.0]. No asymptotes are present in the functions.

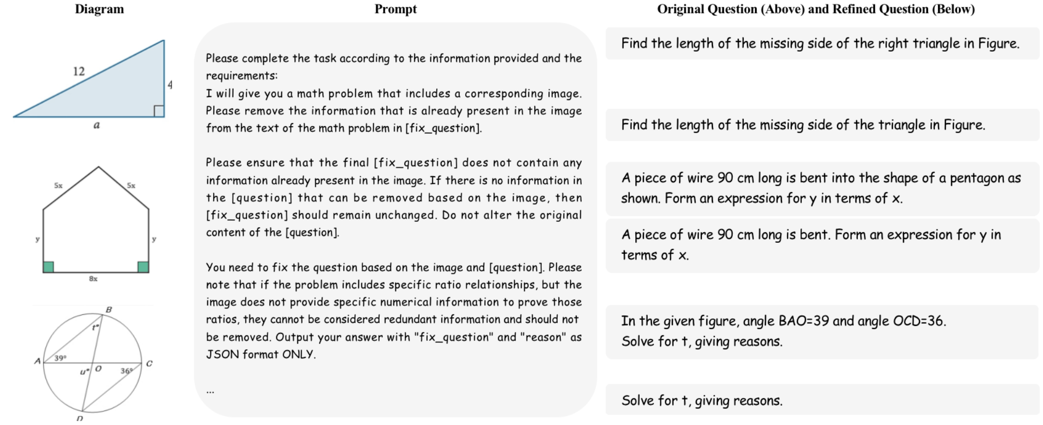
The equation of the function is y = 1*sin(1*x + 3). Observing the x range of [-\u03c0, \u03c0], the zero points determined are -3.0, 0.14. The function spans a range of values from -1.0 at 1.71 to 1.0 at -1.43 across the interval [-π, π]. The functions lack asymptotes.

The function is given by y = 3*log(2, 2*x + 6). With an x range of [-3.0, 5.0], we find the zero points at -2.5. Within the given range [-3.0, 5.0], the maximum value of the function is 12.0 at 5.0, and there is no minimum value point in this range. An asymptote of the function is observed at: x = -3.00.

Figure 10: Function diagram captions.

2. **Data Verification:** Based on their initial categories (subject, subfield, and difficulty level), the problems were organized into distinct groups. Six expert annotators were tasked with meticulously verifying the correctness and completeness of each problem. They refined the detailed chain-of-thought (CoT) rationales and *ensured that there was no overlap with evaluation data by visually inspecting the diagrams*. This rigorous verification process resulted in a total of 4K verified problems.
3. **Text-lite Construction:** To optimize the problems for training mathematical visual capabilities, the 4K problems were processed using GPT-4V with a customized prompt (as shown in Figure 15). This step involved removing redundant information from the question text to create concise, text-lite problems, specifically tailored to our training objectives.

Then, we first feed all the related information into GPT-4V to eliminate the redundant information within text questions, constructing the text-lite version of problems by the prompt in Figure 11. Then, we design three types of prompts for GPT-4 to augment 15 multiple-choice questions (including 10 multiple-choice and 5 binary-choice, i.e., 'True' or 'False') and 5 free-form questions, respectively, as shown in Figure 12. We do not adopt GPT-4V here, since GPT-4V itself would misunderstand diagrams for low-quality data augmentation. The newly generated problems contain detailed CoT rationales and diverse question forms.



Please complete the task according to the information provided and the requirements:
I will give you a math problem that includes a corresponding image. Please remove the information that is already present in the image from the text of the math problem in [fix_question].

Please ensure that the final [fix_question] does not contain any information already present in the image. If there is no information in the [question] that can be removed based on the image, then [fix_question] should remain unchanged. Do not alter the original content of the [question].

You need to fix the question based on the image and [question]. Please note that if the problem includes specific ratio relationships, but the image does not provide specific numerical information to prove those ratios, they cannot be considered redundant information and should not be removed. Output your answer with "fix_question" and "reason" as JSON format ONLY.

...

Find the length of the missing side of the right triangle in Figure.

Find the length of the missing side of the triangle in Figure.

A piece of wire 90 cm long is bent into the shape of a pentagon as shown. Form an expression for y in terms of x.

A piece of wire 90 cm long is bent. Form an expression for y in terms of x.

In the given figure, angle BAO=39 and angle OCD=36.
Solve for t, giving reasons.

Solve for t, giving reasons.

Figure 11: Manually collect visual math problems text-lite version.

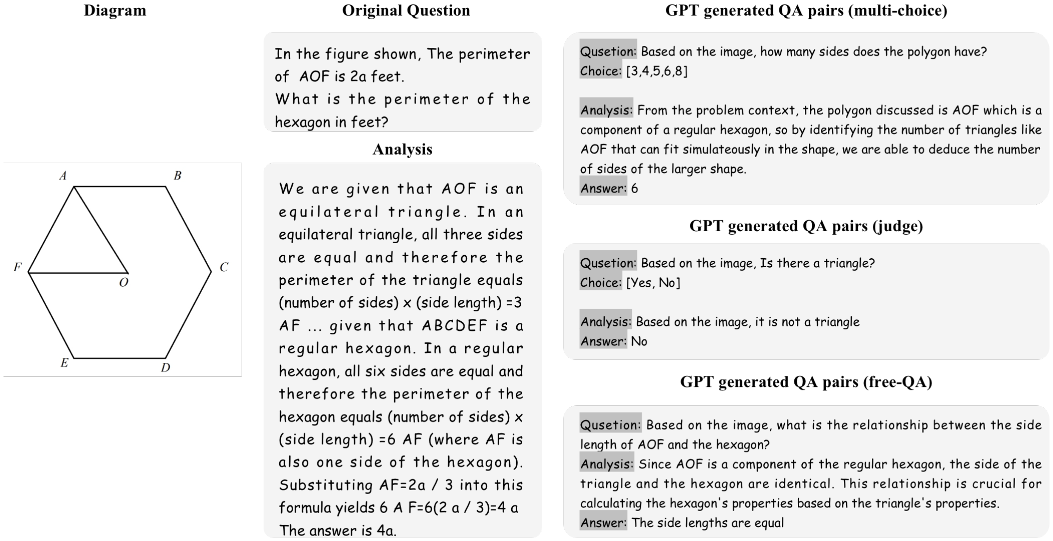


Diagram
Original Question
Analysis
GPT generated QA pairs (multi-choice)
GPT generated QA pairs (judge)
GPT generated QA pairs (free-QA)
A
B
F
O
C
E
D

In the figure shown, The perimeter of AOF is 2a feet.
What is the perimeter of the hexagon in feet?

We are given that AOF is an equilateral triangle. In an equilateral triangle, all three sides are equal and therefore the perimeter of the triangle equals (number of sides) x (side length) =3 AF ... given that ABCDEF is a regular hexagon. In a regular hexagon, all six sides are equal and therefore the perimeter of the hexagon equals (number of sides) x (side length) =6 AF (where AF is also one side of the hexagon). Substituting AF=2a / 3 into this formula yields 6 A F=6(2 a / 3)=4 a
The answer is 4a.

Qusetion: Based on the image, how many sides does the polygon have?
Choice: [3,4,5,6,8]

Analysis: From the problem context, the polygon discussed is AOF which is a component of a regular hexagon, so by identifying the number of triangles like AOF that can fit simulateously in the shape, we are able to deduce the number of sides of the larger shape.
Answer: 6

Qusetion: Based on the image, Is there a triangle?
Choice: [Yes, No]

Analysis: Based on the image, it is not a triangle
Answer: No

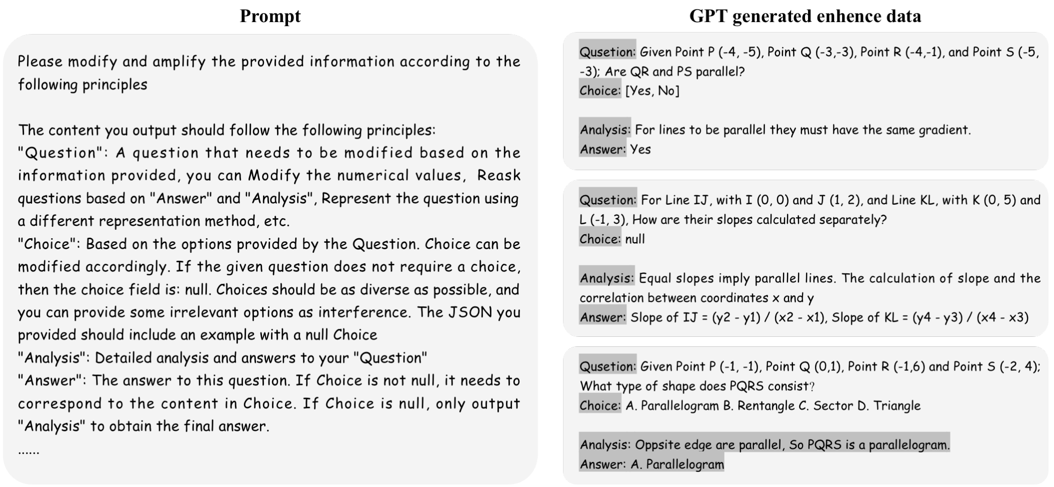
Qusetion: Based on the image, what is the relationship between the side length of AOF and the hexagon?
Analysis: Since AOF is a component of the regular hexagon, the side of the triangle and the hexagon are identical. This relationship is crucial for calculating the hexagon's properties based on the triangle's properties.
Answer: The side lengths are equal

Figure 12: We design different types of prompts for GPT-4 to augment 15 multiple-choice questions and 5 free-form questions, respectively.

Existing Datasets Augmented by GPT-4. Previous efforts have been made to provide some small-scale, plane geometry datasets, e.g., GeoQA (Chen et al., 2021c), GeoQA+ (Chen et al., 2021a), and Geometry3K (Lu et al., 2021). Although they are limited in data scale for tuning MLLMs and include no rationales, we can also regard them as a seed dataset and adopt GPT-4 to augment larger-scale training data. We do not utilize GPT-4V here for the same reason aforementioned. In detail, we design 3 types of question generation approaches using different prompts, as shown in Figure 13. For Geometry3K, as the question texts are normally brief and contain marginal descriptive information, posing challenges for GPT-4 to understand the diagram, we only augment them to generate binary-choice questions, i.e., 'Ture' or 'False'. For GeoQA+, we can leverage the sufficient redundant information within their texts to generate more diverse and accurate multi-choice and free-form questions. Likewise, GPT-4 can produce CoT rationales for each problem.

---

[1] https://homework.study.com
[2] https://www.ixl.com/math
[3] https://mathspace.co/us



Prompt

Please modify and amplify the provided information according to the following principles

The content you output should follow the following principles:
"Question": A question that needs to be modified based on the information provided, you can Modify the numerical values, Reask questions based on "Answer" and "Analysis", Represent the question using a different representation method, etc.
"Choice": Based on the options provided by the Question. Choice can be modified accordingly. If the given question does not require a choice, then the choice field is: null. Choices should be as diverse as possible, and you can provide some irrelevant options as interference. The JSON you provided should include an example with a null Choice
"Analysis": Detailed analysis and answers to your "Question"
"Answer": The answer to this question. If Choice is not null, it needs to correspond to the content in Choice. If Choice is null, only output "Analysis" to obtain the final answer.
......

GPT generated enhence data

Qusetion: Given Point P (-4, -5), Point Q (-3,-3), Point R (-4,-1), and Point S (-5, -3); Are QR and PS parallel?
Choice: [Yes, No]

Analysis: For lines to be parallel they must have the same gradient.
Answer: Yes

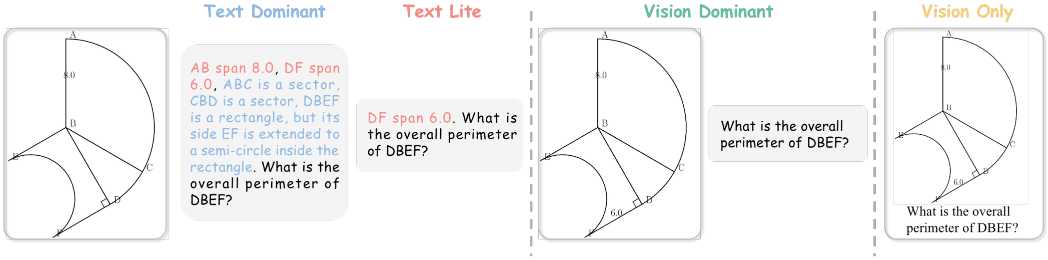
Qusetion: For Line IJ, with I (0, 0) and J (1, 2), and Line KL, with K (0, 5) and L (-1, 3), How are their slopes calculated separately?
Choice: null

Analysis: Equal slopes imply parallel lines. The calculation of slope and the correlation between coordinates x and y
Answer: Slope of IJ = (y2 - y1) / (x2 - x1), Slope of KL = (y4 - y3) / (x4 - x3)

Qusetion: Given Point P (-1, -1), Point Q (0,1), Point R (-1,6) and Point S (-2, 4); What type of shape does PQRS consist?
Choice: A. Parallelogram B. Rentangle C. Sector D. Triangle

Analysis: Oppsite edge are parallel, So PQRS is a parallelogram.
Answer: A. Parallelogram

Figure 13: We design 3 types of question generation approaches using different prompts to augment existing visual mathematical dataset.



AB span 8.0, DF span 6.0, ABC is a sector, CBD is a sector, DBEF is a rectangle, but its side EF is extended to a semi-circle inside the rectangle. What is the overall perimeter of DBEF?

DF span 6.0. What is the overall perimeter of DBEF?

What is the overall perimeter of DBEF?

What is the overall perimeter of DBEF?

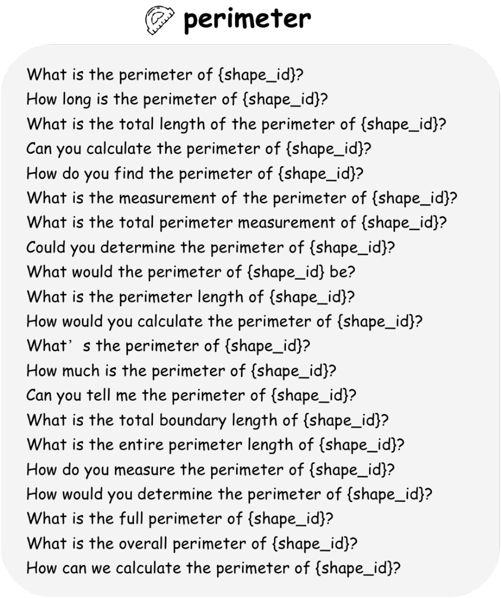
Figure 14: The Text Dominant, Text Lite, Vision Dominant, and Vision Only versions of the same question. Text Dominant and Text Lite use the same image. In the text, the necessary conditions for solving the problem are highlighted in red, while redundant descriptive conditions are highlighted in blue. In the Vision Only version, the question is rendered in the image, with no textual format.

Data Engine Captions Annotated by GPT-4. Given the delicately designed data engine for automatic diagram-caption creation, we can utilize the generated large-scale pairs to annotate question-answering data using GPT-4V. Different from the previous two sources that augment questions based on questions, we utilize the GPT-4V model here for caution data with two reasons: first, the detailed caption from our data engine can well guide GPT-4V for relatively higher-quality visual embedding; second, the visual input serves as guidance to provide additional spatial information for broad question forms. As shown in Figure 21 and Figure 22, we adopt different prompts for function and plane geometry problems, ensuring that the generated question-answering data is of high quality for instruction tuning.

Data Engine Generated Problems:

PLANE GEOMETRY. Based on the generation process described in Section A.4.1, we pose questions about the final shape in the generation sequence. We designed 6 types of questions: finding the perimeter, finding the area, finding the base length, finding the angle, finding the arc length, and finding the extended edge length. Each type of question has a set of templates that can be randomly selected, as shown in Figure 15-20. As for the answer and analysis, each shape has a set of templates for different types of questions to choose from, as shown in Section A.4.1.

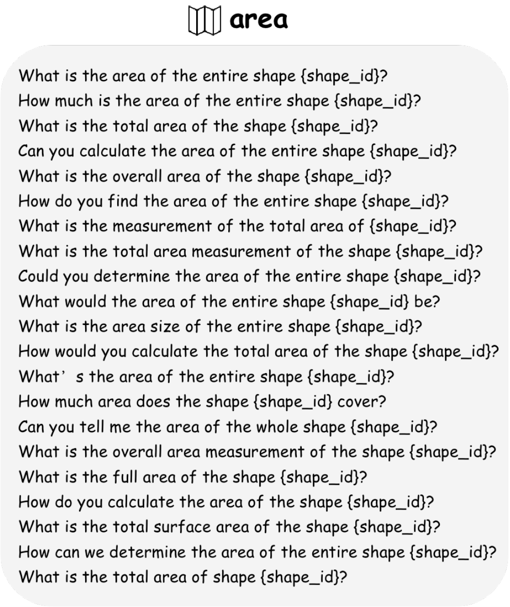
To further enhance the model's understanding of different forms of questions and better utilize the diverse modal information in the text and images, we divided the plain geometry questions generated by the Data Engine into four versions referring to MathVerse (Zhang et al., 2024b): Text Dominant, Text Lite, Vision Dominant, and Vision Only.



perimeter

What is the perimeter of {shape_id}?
How long is the perimeter of {shape_id}?
What is the total length of the perimeter of {shape_id}?
Can you calculate the perimeter of {shape_id}?
How do you find the perimeter of {shape_id}?
What is the measurement of the perimeter of {shape_id}?
What is the total perimeter measurement of {shape_id}?
Could you determine the perimeter of {shape_id}?
What would the perimeter of {shape_id} be?
What is the perimeter length of {shape_id}?
How would you calculate the perimeter of {shape_id}?
What’ s the perimeter of {shape_id}?
How much is the perimeter of {shape_id}?
Can you tell me the perimeter of {shape_id}?
What is the total boundary length of {shape_id}?
What is the entire perimeter length of {shape_id}?
How do you measure the perimeter of {shape_id}?
How would you determine the perimeter of {shape_id}?
What is the full perimeter of {shape_id}?
What is the overall perimeter of {shape_id}?
How can we calculate the perimeter of {shape_id}?

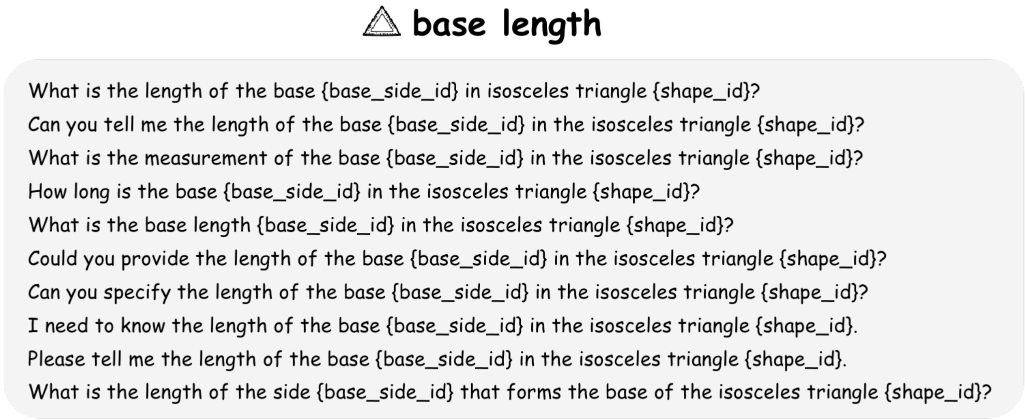
Figure 15: Perimeter problem templates.



area

What is the area of the entire shape {shape_id}?
How much is the area of the entire shape {shape_id}?
What is the total area of the shape {shape_id}?
Can you calculate the area of the entire shape {shape_id}?
What is the overall area of the shape {shape_id}?
How do you find the area of the entire shape {shape_id}?
What is the measurement of the total area of {shape_id}?
What is the total area measurement of the shape {shape_id}?
Could you determine the area of the entire shape {shape_id}?
What would the area of the entire shape {shape_id} be?
What is the area size of the entire shape {shape_id}?
How would you calculate the total area of the shape {shape_id}?
What’ s the area of the entire shape {shape_id}?
How much area does the shape {shape_id} cover?
Can you tell me the area of the whole shape {shape_id}?
What is the overall area measurement of the shape {shape_id}?
What is the full area of the shape {shape_id}?
How do you calculate the area of the shape {shape_id}?
What is the total surface area of the shape {shape_id}?
How can we determine the area of the entire shape {shape_id}?
What is the total area of shape {shape_id}?

Figure 16: Area problem templates.



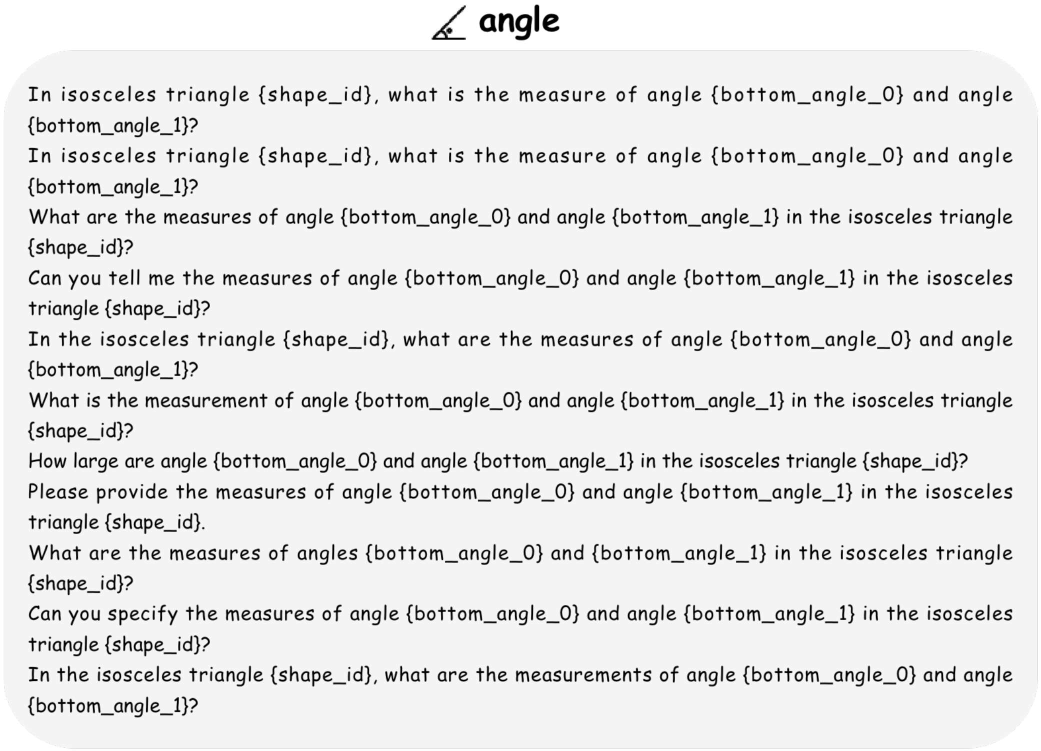
base length

What is the length of the base {base_side_id} in isosceles triangle {shape_id}?
Can you tell me the length of the base {base_side_id} in the isosceles triangle {shape_id}?
What is the measurement of the base {base_side_id} in the isosceles triangle {shape_id}?
How long is the base {base_side_id} in the isosceles triangle {shape_id}?
What is the base length {base_side_id} in the isosceles triangle {shape_id}?
Could you provide the length of the base {base_side_id} in the isosceles triangle {shape_id}?
Can you specify the length of the base {base_side_id} in the isosceles triangle {shape_id}?
I need to know the length of the base {base_side_id} in the isosceles triangle {shape_id}.
Please tell me the length of the base {base_side_id} in the isosceles triangle {shape_id}.
What is the length of the side {base_side_id} that forms the base of the isosceles triangle {shape_id}?

Figure 17: Base length problem templates.

Text Dominant	We marked all the conditions required for solving the problem in the diagram and also described these conditions in the text, along with some redundant descriptive text.
Text Lite	All the conditions required for solving the problem are randomly divided into two parts: one part is marked in the diagram, and the other part is described in the text. In other words, the conditions in the diagram and the conditions in the text do not overlap.
Vision Dominant	All the conditions required for solving the problem are marked in the diagram, while the text only contains the question without any conditions.
Vision Only	Not only are all the conditions required for solving the problem marked in the diagram, but the question is also rendered in the diagram, leaving the text portion empty.

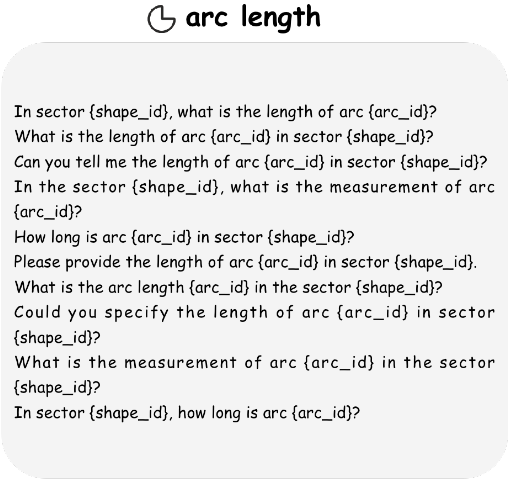
The differences among the four versions of the same question are illustrated in Figure 14. Each basic shape will retain a set of redundant conditions. During the shape generation process, there is a 50% probability of including these redundant conditions.



angle

In isosceles triangle {shape_id}, what is the measure of angle {bottom_angle_0} and angle {bottom_angle_1}?
In isosceles triangle {shape_id}, what is the measure of angle {bottom_angle_0} and angle {bottom_angle_1}?
What are the measures of angle {bottom_angle_0} and angle {bottom_angle_1} in the isosceles triangle {shape_id}?
Can you tell me the measures of angle {bottom_angle_0} and angle {bottom_angle_1} in the isosceles triangle {shape_id}?
In the isosceles triangle {shape_id}, what are the measures of angle {bottom_angle_0} and angle {bottom_angle_1}?
What is the measurement of angle {bottom_angle_0} and angle {bottom_angle_1} in the isosceles triangle {shape_id}?
How large are angle {bottom_angle_0} and angle {bottom_angle_1} in the isosceles triangle {shape_id}?
Please provide the measures of angle {bottom_angle_0} and angle {bottom_angle_1} in the isosceles triangle {shape_id}.
What are the measures of angles {bottom_angle_0} and {bottom_angle_1} in the isosceles triangle {shape_id}?
Can you specify the measures of angle {bottom_angle_0} and angle {bottom_angle_1} in the isosceles triangle {shape_id}?
In the isosceles triangle {shape_id}, what are the measurements of angle {bottom_angle_0} and angle {bottom_angle_1}?

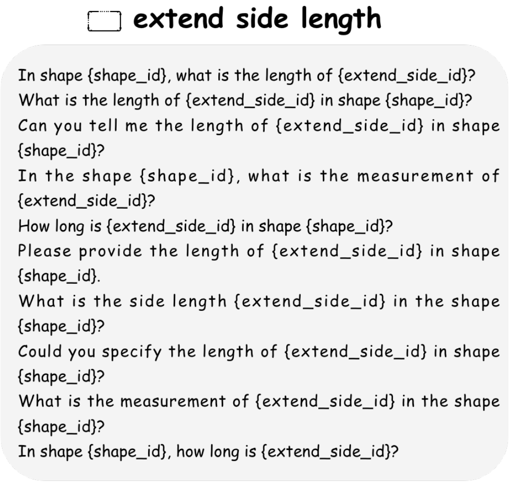
Figure 18: Angle problem templates.



arc length

In sector {shape_id}, what is the length of arc {arc_id}?
What is the length of arc {arc_id} in sector {shape_id}?
Can you tell me the length of arc {arc_id} in sector {shape_id}?
In the sector {shape_id}, what is the measurement of arc {arc_id}?
How long is arc {arc_id} in sector {shape_id}?
Please provide the length of arc {arc_id} in sector {shape_id}.
What is the arc length {arc_id} in the sector {shape_id}?
Could you specify the length of arc {arc_id} in sector {shape_id}?
What is the measurement of arc {arc_id} in the sector {shape_id}?
In sector {shape_id}, how long is arc {arc_id}?

Figure 19: Arc length problem templates.

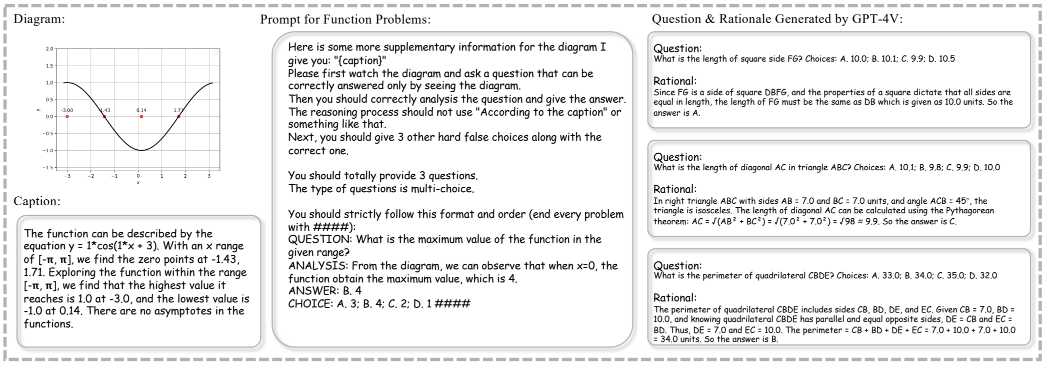


extend side length

In shape {shape_id}, what is the length of {extend_side_id}?
What is the length of {extend_side_id} in shape {shape_id}?
Can you tell me the length of {extend_side_id} in shape {shape_id}?
In the shape {shape_id}, what is the measurement of {extend_side_id}?
How long is {extend_side_id} in shape {shape_id}?
Please provide the length of {extend_side_id} in shape {shape_id}.
What is the side length {extend_side_id} in the shape {shape_id}?
Could you specify the length of {extend_side_id} in shape {shape_id}?
What is the measurement of {extend_side_id} in the shape {shape_id}?
In shape {shape_id}, how long is {extend_side_id}?

Figure 20: Extend side length problem templates.

FUNCTION. All functions will be examined with two types of questions: finding the derivative and finding the extrema. After obtaining the derivative, we calculate whether the derivative has zeros within the given domain. The presence of zeros determines the method for calculating the extrema.



Caption:

The function can be described by the equation y = 1*cos(1*x + 3). With an x range of [-π, π], we find the zero points at -1.43, 1.71. Exploring the function within the range [-π, π], we find that the highest value it reaches is 1.0 at -3.0, and the lowest value is -1.0 at 0.14. There are no asymptotes in the functions.

Prompt for Function Problems:

Here is some more supplementary information for the diagram I give you: "{caption}"
Please first watch the diagram and ask a question that can be correctly answered only by seeing the diagram.
Then you should correctly analysis the question and give the answer.
The reasoning process should not use "According to the caption" or something like that.
Next, you should give 3 other hard false choices along with the correct one.

You should totally provide 3 questions.
The type of questions is multi-choice.

You should strictly follow this format and order (end every problem with ####):
QUESTION: What is the maximum value of the function in the given range?
ANALYSIS: From the diagram, we can observe that when x=0, the function obtain the maximum value, which is 4.
ANSWER: B. 4
CHOICE: A. 3; B. 4; C. 2; D. 1 ####

Question & Rationale Generated by GPT-4V:

Question:
What is the length of square side FG? Choices: A. 10.0; B. 10.1; C. 9.9; D. 10.5

Rational:
Since FG is a side of square DBFG, and the properties of a square dictate that all sides are equal in length, the length of FG must be the same as DB which is given as 10.0 units. So the answer is A.

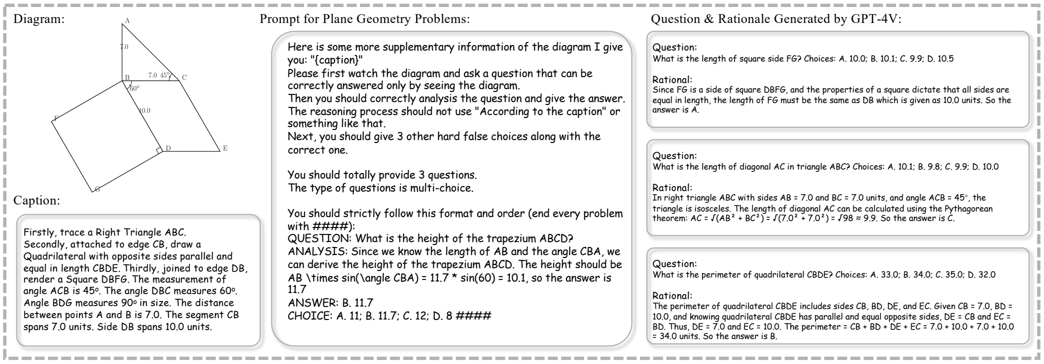
Question:
What is the length of diagonal AC in triangle ABC? Choices: A. 10.1; B. 9.8; C. 9.9; D. 10.0

Rational:
In right triangle ABC with sides AB = 7.0 and BC = 7.0 units, and angle ACB = 45°, the triangle is isosceles. The length of diagonal AC can be calculated using the Pythagorean theorem: AC = √(AB² + BC²) = √(7.0² + 7.0²) = √98 ≈ 9.9. So the answer is C.

Question:
What is the perimeter of quadrilateral CBDE? Choices: A. 33.0; B. 34.0; C. 35.0; D. 32.0

Rational:
The perimeter of quadrilateral CBDE includes sides CB, BD, DE, and EC. Given CB = 7.0, BD = 10.0, and knowing quadrilateral CBDE has parallel and equal opposite sides, DE = CB and EC = BD. Thus, DE = 7.0 and EC = 10.0. The perimeter = CB + BD + DE + EC = 7.0 + 10.0 + 7.0 + 10.0 = 34.0 units. So the answer is B.

Figure 21: The function prompt for GPT-4V and the generated questions and rationals.



Caption:

Firstly, trace a Right Triangle ABC.
Secondly, attached to edge CB, draw a Quadrilateral with opposite sides parallel and equal in length CBDE. Thirdly, joined to edge DB, render a Square DBFG. The measurement of angle ACB is 45°. The angle DBC measures 60°. Angle BDG measures 90° in size. The distance between points A and B is 7.0. The segment CB spans 7.0 units. Side DB spans 10.0 units.

Prompt for Plane Geometry Problems:

Here is some more supplementary information of the diagram I give you: "{caption}"
Please first watch the diagram and ask a question that can be correctly answered only by seeing the diagram.
Then you should correctly analysis the question and give the answer.
The reasoning process should not use "According to the caption" or something like that.
Next, you should give 3 other hard false choices along with the correct one.

You should totally provide 3 questions.
The type of questions is multi-choice.

You should strictly follow this format and order (end every problem with ####):
QUESTION: What is the height of the trapezium ABCD?
ANALYSIS: Since we know the length of AB and the angle CBA, we can derive the height of the trapezium ABCD. The height should be AB \times sin(\angle CBA) = 11.7 * sin(60) = 10.1, so the answer is 11.7
ANSWER: B. 11.7
CHOICE: A. 11; B. 11.7; C. 12; D. 8 ####

Question & Rationale Generated by GPT-4V:

Question:
What is the length of square side FG? Choices: A. 10.0; B. 10.1; C. 9.9; D. 10.5

Rational:
Since FG is a side of square DBFG, and the properties of a square dictate that all sides are equal in length, the length of FG must be the same as DB which is given as 10.0 units. So the answer is A.

Question:
What is the length of diagonal AC in triangle ABC? Choices: A. 10.1; B. 9.8; C. 9.9; D. 10.0

Rational:
In right triangle ABC with sides AB = 7.0 and BC = 7.0 units, and angle ACB = 45°, the triangle is isosceles. The length of diagonal AC can be calculated using the Pythagorean theorem: AC = √(AB² + BC²) = √(7.0² + 7.0²) = √98 ≈ 9.9. So the answer is C.

Question:
What is the perimeter of quadrilateral CBDE? Choices: A. 33.0; B. 34.0; C. 35.0; D. 32.0

Rational:
The perimeter of quadrilateral CBDE includes sides CB, BD, DE, and EC. Given CB = 7.0, BD = 10.0, and knowing quadrilateral CBDE has parallel and equal opposite sides, DE = CB and EC = BD. Thus, DE = 7.0 and EC = 10.0. The perimeter = CB + BD + DE + EC = 7.0 + 10.0 + 7.0 + 10.0 = 34.0 units. So the answer is B.

Figure 22: The geometry prompt for GPT-4V and the generated questions and rationals.

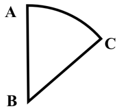
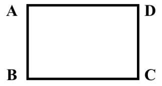
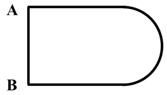
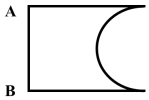
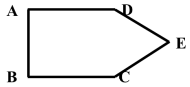
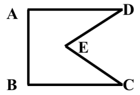
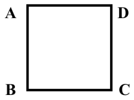
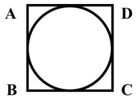
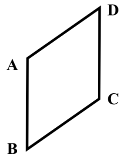
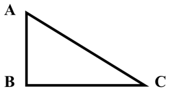
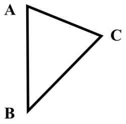
	Initial Side	Extend Side	Additional Condition for Meta Reasoning
	AB	BC	∅
	AB	CD	∅
	AB	∅	
	AB	∅	
	AB	CE DE	CDE is a equilateral triangle
	AB	∅	
	AB	AD CD BC	∅
	AB	AD CD BC	∅
	AB	CD	∅
	AB	AC BC	∠C = x degree
	AB	AC BC	∠B = x degree

Figure 23: The initial side, extend side, and additional conditions for meta-reasoning of each basic shape. Some special shapes are not extended and only appear in the last position of the generation sequence, thus their extend side is ∅.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, pp. 1877–1901, 2020.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.46. URL `https://aclanthology.org/2021.findings-acl.46`.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *ArXiv*, abs/2105.14517, 2021b. URL `https://api.semanticscholar.org/CorpusID:235253782`.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021c.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *ArXiv*, abs/2212.02746, 2022.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. `https://lmsys.org/blog/2023-03-30-vicuna/`, March 2023.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023a.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023b.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, Weizhu Chen, et al. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*, 2023.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

Ziyu Guo*, Renrui Zhang*#, Xiangyang Zhu, Chengzhuo Tong, Peng Gao, Chunyuan Li, and Pheng-Ann Heng. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. *arXiv preprint arXiv:2408.16768*, 2024.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilue Wang, Longzan Luo, Lily Lee, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, et al. Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation. *arXiv preprint arXiv:2411.18623*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *Arxiv 2401.04088*, 2024.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024a. URL `https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/`.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next: Tackling multi-image, video, and 3d in large multimodal models, June 2024b. URL `https://llava-vl.github.io/blog/2024-06-16-llava-next-interleave/`.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024c. URL `https://arxiv.org/abs/2407.07895`.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. Unimath: A foundational and multimodal mathematical reasoner. In *EMNLP*, 2023.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want, 2025. URL https://arxiv.org/abs/2403.20271.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Lily Lee, Kaichen Zhou, Pengju An, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. *NeurIPS 2024*, 2024b.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Tianshuo Peng, Mingsheng Li, Hongbin Zhou, Renqiu Xia, Renrui Zhang, Lei Bai, Song Mao, Bin Wang, Conghui He, Aojun Zhou, et al. Chimera: Improving generalist model with domain-specific experts. *arXiv preprint arXiv:2412.05983*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Yiwen Tang, Zoey Guo, Zhuhao Wang, Ray Zhang, Qizhi Chen, Junli Liu, Delin Qu, Zhigang Wang, Dong Wang, Xuelong Li, et al. Exploring the potential of encoder-free architectures in 3d lmms. *arXiv preprint arXiv:2502.09620*, 2025.

InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Guanqun Wang, Xinyu Wei, Jiaming Liu, Ray Zhang, Yichi Zhang, Kevin Zhang, Maurice Chong, and Shanghang Zhang. Mr-mllm: Mutual reinforcement of multimodal comprehension and vision perception, 2024a. URL https://arxiv.org/abs/2406.15768.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in LLMs for enhanced mathematical reasoning. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=z8TW0ttBPp.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. *arXiv preprint arXiv:2405.03548*, 2024.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024a. URL `https://openreview.net/forum?id=d4UiXAHN2W`.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024b.