澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

VisAidMath : Benchmarking Visual-Aided Mathematical Reasoning

Jingkun Ma[1], Runzhe Zhan[1], Yang Li[1],
Di Sun[2], Hou Pong Chan[3], Lidia S. Chao[1], Derek F. Wong[1*]

[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
[2]University of Macau
[3]DAMO Academy, Alibaba Group

Mathematical Reasoning

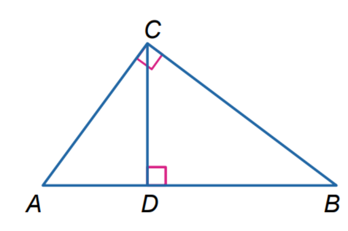
- Benchmark：Capabilities in logical thinking, arithmetic operation, mathematical knowledge.
- Task Paradigms: Text-only & Visual Context Reasoning

Question: *Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. Then, how many tennis balls does Roger have now?*

Answer: *Roger started with 5 balls. 2 cans of 3 tennis balls each are 6 tennis balls. 5 + 6 = 11. The answer is 11.*

Text-only Mathematical Reasoning

Question: In triangle ABC, AD = 3 and BD = 14. Find CD.
Choices: (A) 6.0 (B) 6.5 (C) 7.0 (D) 8.5
Visual Context:

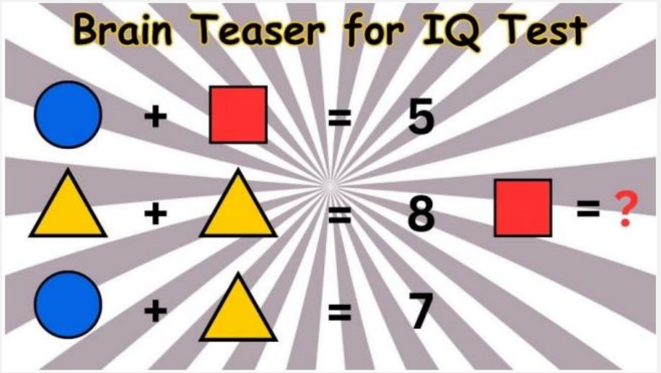


Answer: (B) 6.5

Mathematical Reasoning in Visual Context

Related Work *Multi-Modal Mathematical Reasoning*

- Representative Benchmark – MathVista: investigated multi-modal MPS by introducing visual context.
- Focus on evaluating reasoning steps in **textual dimension** to solve the problems



Question: Find the value of the square in the figure.

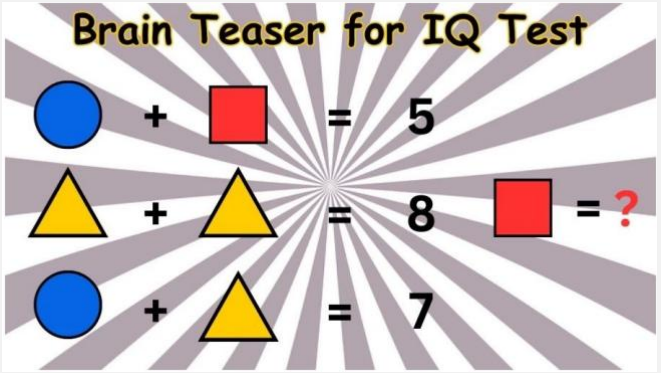
Solution: Circle + Square = 5, Triangle + Triangle = 8, Triangle = 4. Circle + Triangle = 7, Circle = 3. Therefore Square = 2 Answer: 2

Example of mathematical reasoning with visual context in MathVista

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao 337 Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical 338 reasoning of foundation models in visual contexts. In proceedings of the ICLR 2024

Related Work *Multi-Modal Mathematical Reasoning*

- Representative Benchmark – MathVista: investigated multi-modal MPS by introducing visual context.
- Focus on evaluating reasoning steps in **textual dimension** to solve the problems
- Observation: Multi-modal capabilities of MPS extend beyond comprehending input modalities (e.g. inference of information from other modalities)



Question: Find the value of the square in the figure.

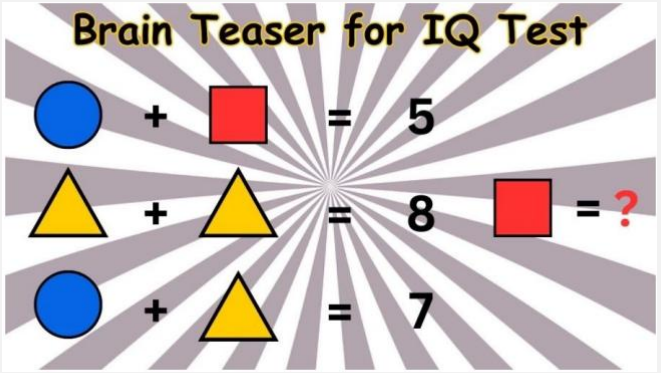
Solution: Circle + Square = 5, Triangle + Triangle = 8, Triangle = 4. Circle + Triangle = 7, Circle = 3. Therefore Square = 2 Answer: 2

Example of mathematical reasoning with visual context in MathVista

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao 337 Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical 338 reasoning of foundation models in visual contexts. In proceedings of the ICLR 2024

Related Work *Multi-Modal Mathematical Reasoning*

- Representative Benchmark – MathVista: investigated multi-modal MPS by introducing visual context.
- Focus on evaluating reasoning steps in **textual dimension** to solve the problems
- Observation: Multi-modal capabilities of MPS extend beyond comprehending input modalities (e.g. inference of information from other modalities)
- Problem: Cross-modality evaluation aspects are rarely taken into account in the evaluation



Question: Find the value of the square in the figure.

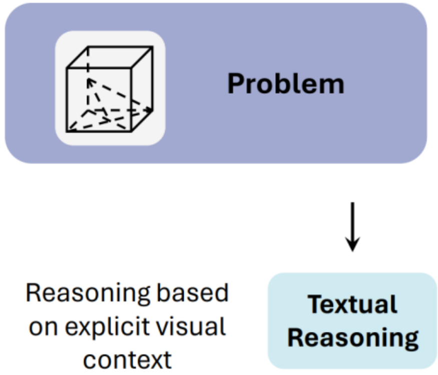
Solution: Circle + Square = 5, Triangle + Triangle = 8, Triangle = 4. Circle + Triangle = 7, Circle = 3. Therefore Square = 2 Answer: 2

Example of mathematical reasoning with visual context in MathVista

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao 337 Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical 338 reasoning of foundation models in visual contexts. In proceedings of the ICLR 2024

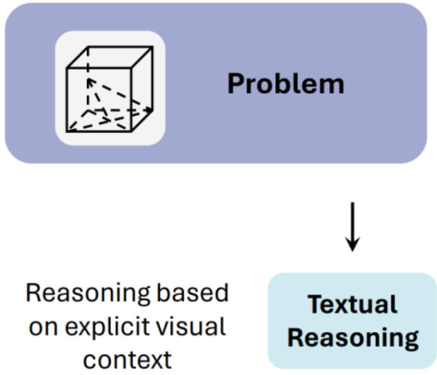
Background

- Problem: Cross-modality evaluation aspects are rarely taken into account in the evaluation
- Cause: Visual elements are viewed as static context only, providing fixed information.



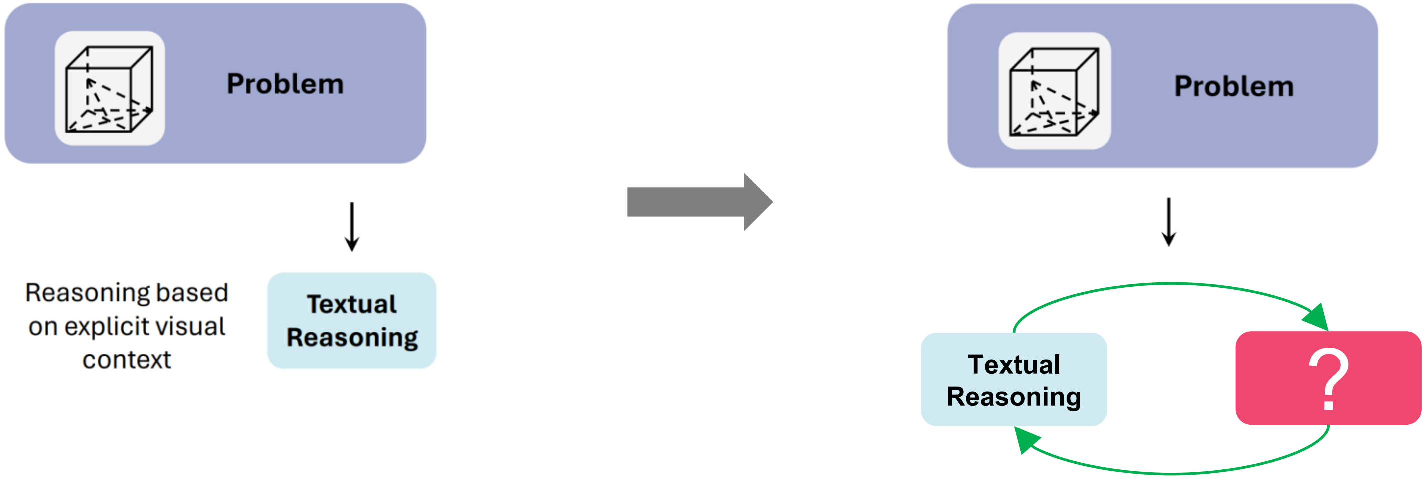
Background

- Problem: Cross-modality evaluation aspects are rarely taken into account in the evaluation
- Cause: Visual elements are viewed as static context only, providing fixed information.
- Limitation:
  - Decision space is pruned
  - Hard to measure the interactive reasoning between different modalities.



Background

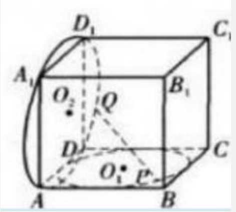
- Problem: Cross-modality evaluation aspects are rarely taken into account in the evaluation
- Cause: Visual elements are viewed as static context only, providing fixed information.
- Limitation:
  - Decision space is pruned
  - Hard to measure the interactive reasoning between different modalities.
- **What visual elements can be created to effectively aid mathematical problem-solving process?**



Background

- **What visual elements can be created to effectively aid problem-solving process?**
- Visual Context: Visual elements are viewed as static context only, providing fixed information.

Visual Context:

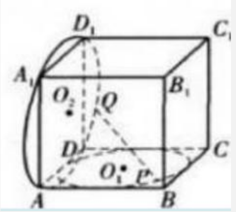


Question: *As shown in the figure, the prisms of the square ABCD - A1B1C1D1 have the lengths 1… Find the range of the length of P Q.*

Background

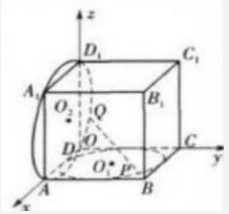
- **What visual elements can be created to effectively aid problem-solving process?**
- Visual Context: Visual elements are viewed as static context only, providing fixed information.
- **Visual-aids**:
  - Visual elements created in visual space.
  - Reveal critical hidden conditions and alleviate problem-solving difficulty.

Visual Context:



Question: *As shown in the figure, the prisms of the square ABCD - A1B1C1D1 have the lengths 1… Find the range of the length of P Q.*

Visual Aids:

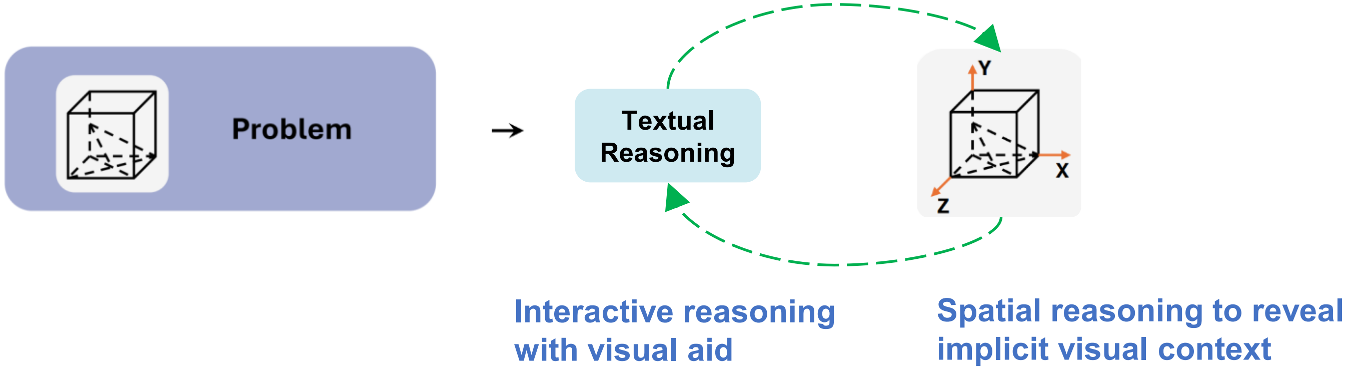


Answer: *By analysis and calculation using three-dimensional coordinate system, … PQ = …*

Example of mathematical problem with rectangular three-dimensional coordinate system as visual-aid

Motivation

- **Key Idea:** Benchmarking mathematical problems solved by creating visual-aids -> cross-modality inference evaluation

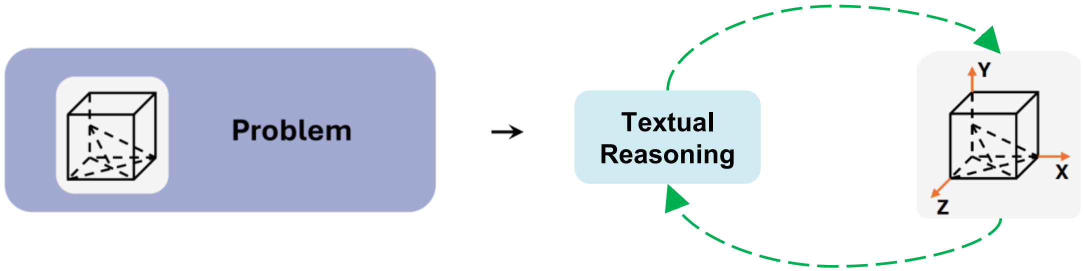

Problem
→
Textual Reasoning
Y
X
Z
Interactive reasoning with visual aid
Spatial reasoning to reveal implicit visual context

Motivation

- **Key Idea:** Benchmarking mathematical problems solved by creating visual-aids -> cross-modality inference evaluation

➢ Are created based on **comprehending input** modalities

➢ Are generated to reveal implicit elements and effectively enlarge MLLMs decision space



Interactive reasoning with visual aid

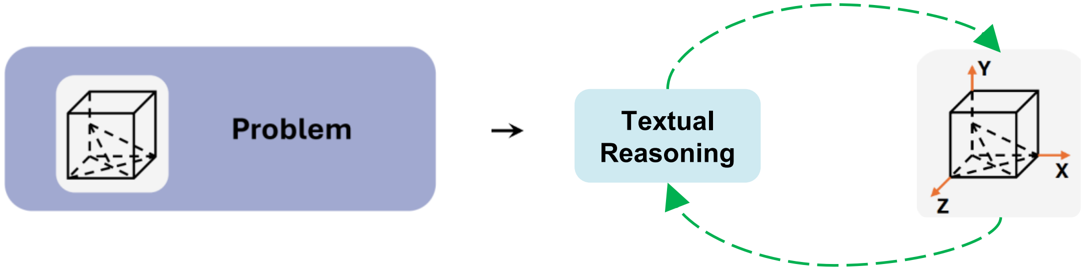
Spatial reasoning to reveal implicit visual context

Motivation

- **Key Idea:** Benchmarking mathematical problems solved by creating visual-aids -> cross-modality inference evaluation

➢ Are created based on **comprehending input** modalities → Spatial Imagination

➢ Are generated to reveal implicit elements and effectively enlarge MLLMs decision space → Cross-modality Spatial Reasoning

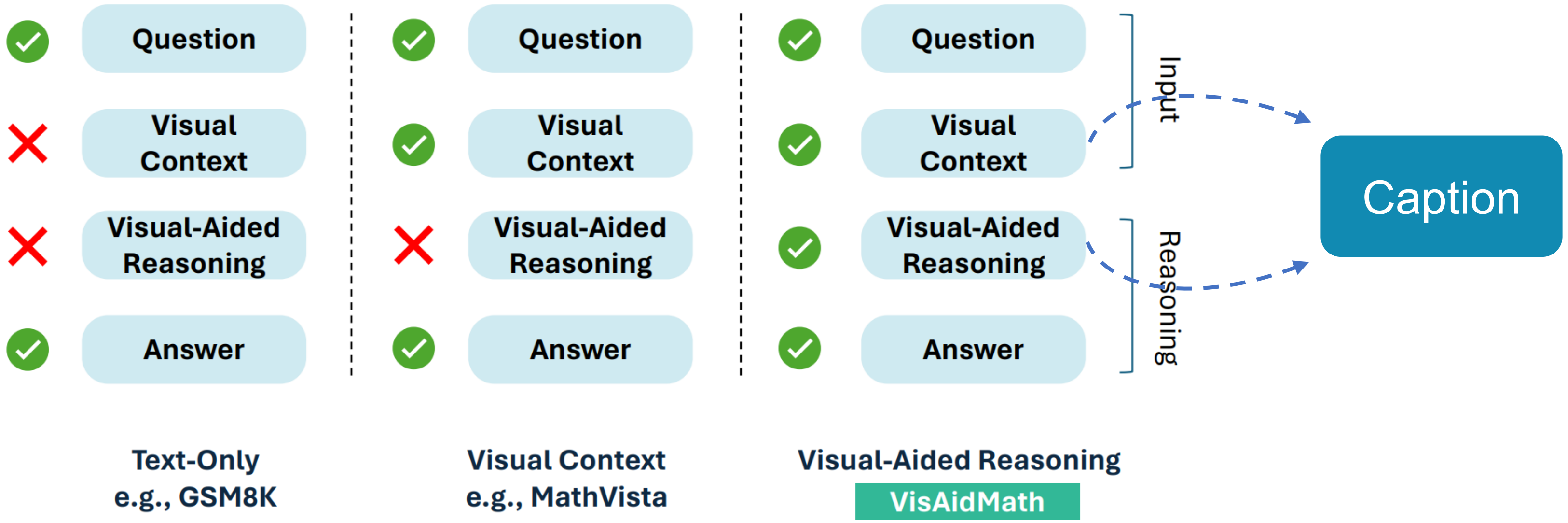


Interactive reasoning with visual aid

Spatial reasoning to reveal implicit visual context

Dataset *Principles*

- **Visual-aids** is included as **essential** data elements within each question, while the **visual context** is optional
- Additionally annotate precise **captions** for both the visual context and the visual aids
  - Observation: Extremely poor performance on visual-aids image generation task



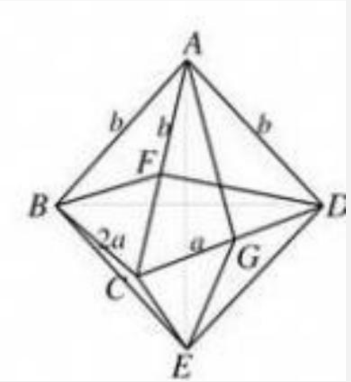
	Text-Only e.g., GSM8K	Visual Context e.g., MathVista	Visual-Aided Reasoning VisAidMath
Question	✓	✓	✓
Visual Context	✗	✓	✓
Visual-Aided Reasoning	✗	✗	✓
Answer	✓	✓	✓

Categories

- Mathematical Branches
- Complexity Level
- Visual-Aids Type

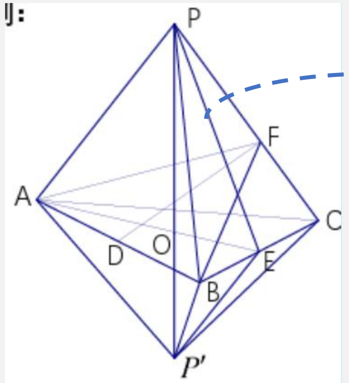
Question: *Given that two congruent triangular pheons are glued together to obtain a hexahedron with all the dihedral angles equal, and that the shortest prong of the hexahedron is 2, the distance between the two farthest vertices is?*

Visual Context:



Solid Geometry

Visual-Aids:

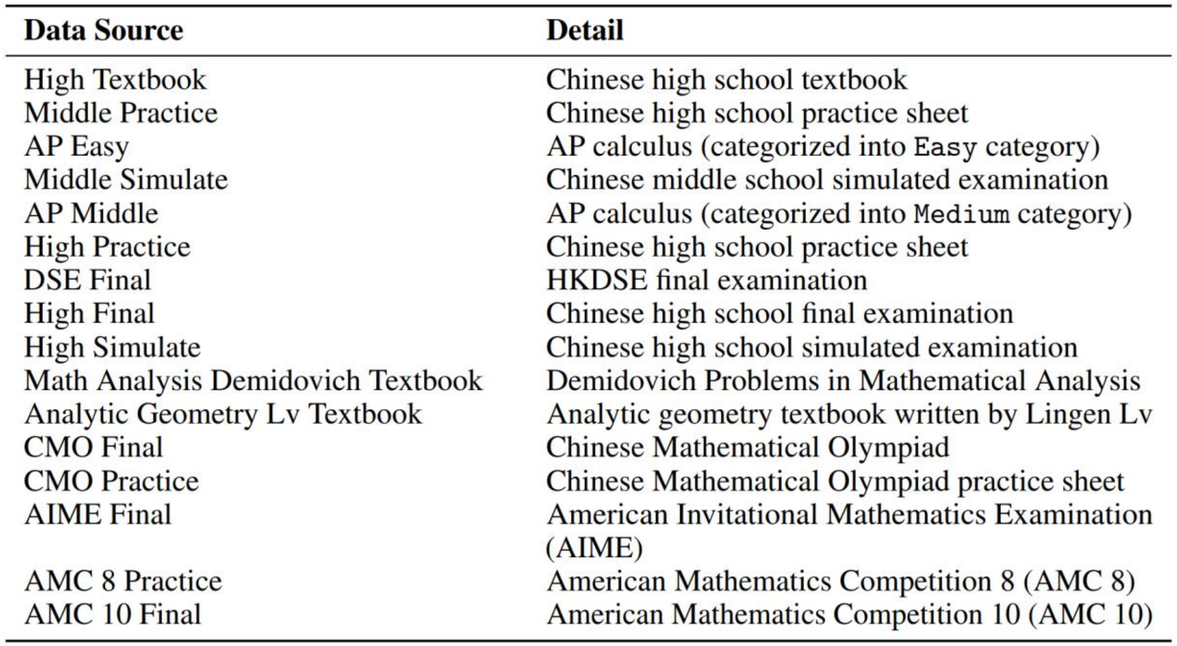


Auxiliary Line

Chinese Mathematical Olympiad

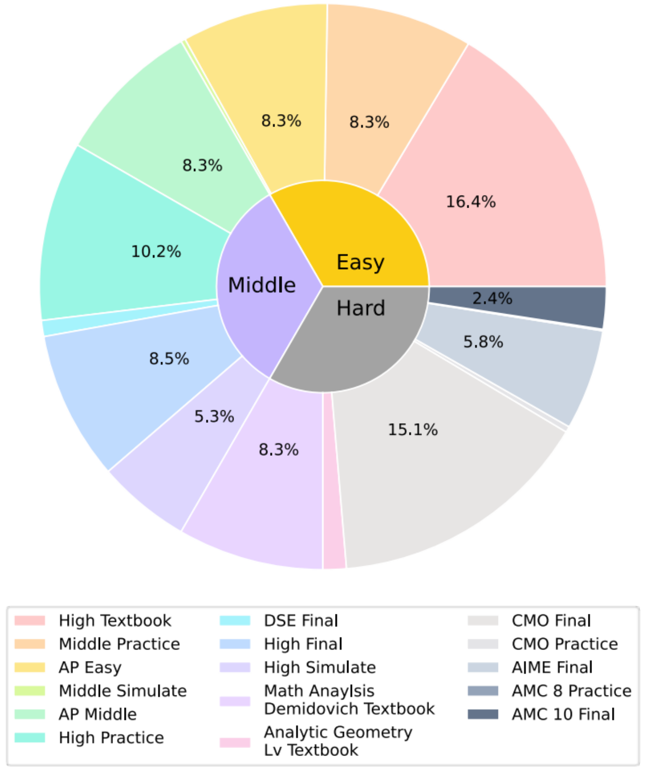
Categories *Complexity*

- **Chinese community** offers a **larger** pool of mathematical problems with visual aids across various complexity level.
- Categorize data samples based on data sources
  - ➢ Easy: e.g., High School Entrance Examination,
  - ➢ Medium: e.g., College Entrance Examination,
  - ➢ High: e.g., Mathematical Olympiad.



Data Source	Detail
High Textbook	Chinese high school textbook
Middle Practice	Chinese high school practice sheet
AP Easy	AP calculus (categorized into `Easy` category)
Middle Simulate	Chinese middle school simulated examination
AP Middle	AP calculus (categorized into `Medium` category)
High Practice	Chinese high school practice sheet
DSE Final	HKDSE final examination
High Final	Chinese high school final examination
High Simulate	Chinese high school simulated examination
Math Analysis Demidovich Textbook	Demidovich Problems in Mathematical Analysis
Analytic Geometry Lv Textbook	Analytic geometry textbook written by Lingen Lv
CMO Final	Chinese Mathematical Olympiad
CMO Practice	Chinese Mathematical Olympiad practice sheet
AIME Final	American Invitational Mathematics Examination (AIME)
AMC 8 Practice	American Mathematics Competition 8 (AMC 8)
AMC 10 Final	American Mathematics Competition 10 (AMC 10)

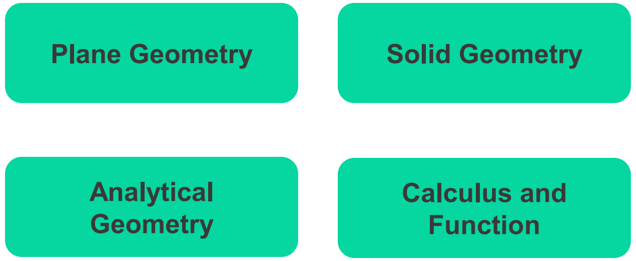
Detail of data sources



Distribution of data sources and difficulty levels.

Categories *Math Branch & Visual-aids Type*

- Ensure **diversity** and **balance**: Manually collected and annotated a range of categories within the benchmark
- Mathematical Branch: Different **theorem** and **logic thinking**
- Visual-aids Type: Different **spatial reasoning path**
  - ➢ **Multiple types** of Visual-aids can be created within **a data sample**



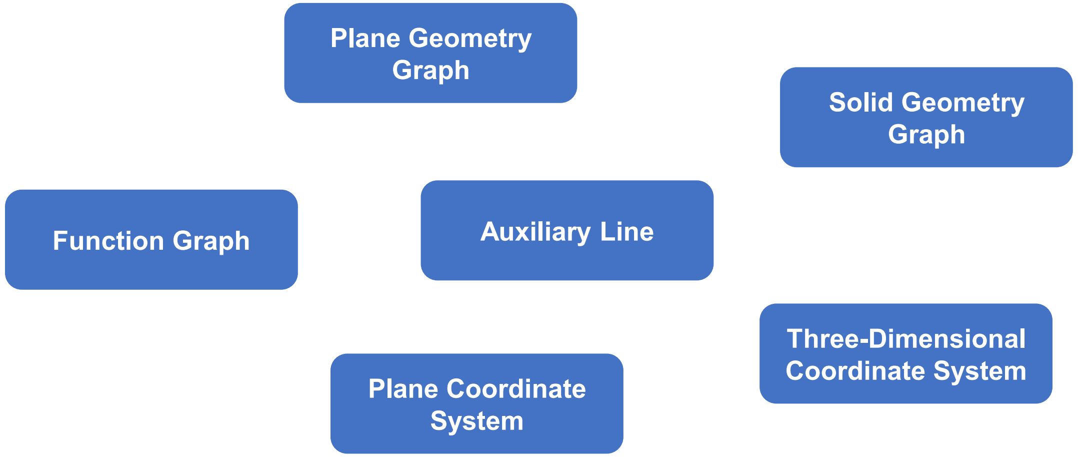
Plane Geometry

Solid Geometry

Analytical Geometry

Calculus and Function

Mathematical Branches



Plane Geometry Graph

Solid Geometry Graph

Function Graph

Auxiliary Line

Three-Dimensional Coordinate System

Plane Coordinate System

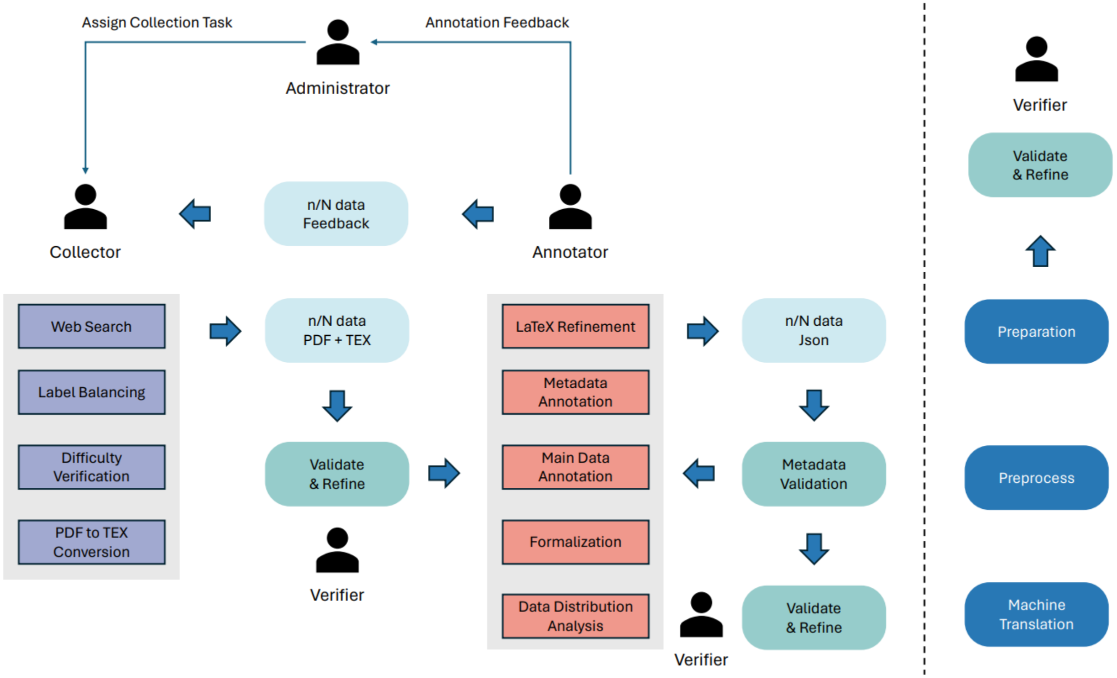
Visual-aids Type

Construction Pipeline

- Challenge:
  - Collect and filter **qualified** mathematical problems
  - Ensure data **diversity** and **balance**

➢ Multi-round Verification

➢ Batch Collection with Feedback



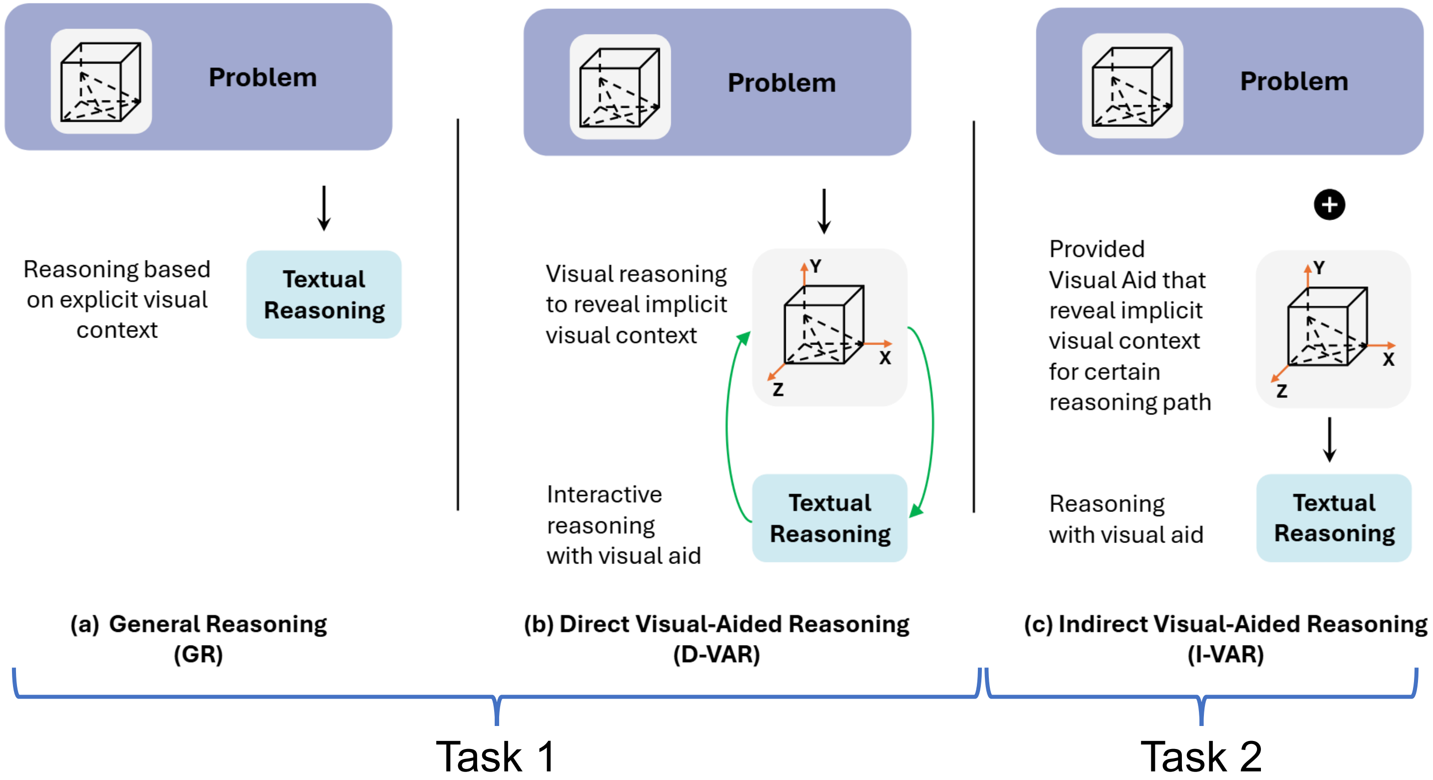
Task Definition

- **Definition**: **Generate** or **leverage visual aids** alongside mathematical reasoning to achieve the correct answers

➢ Task 1: General Reasoning (GR)

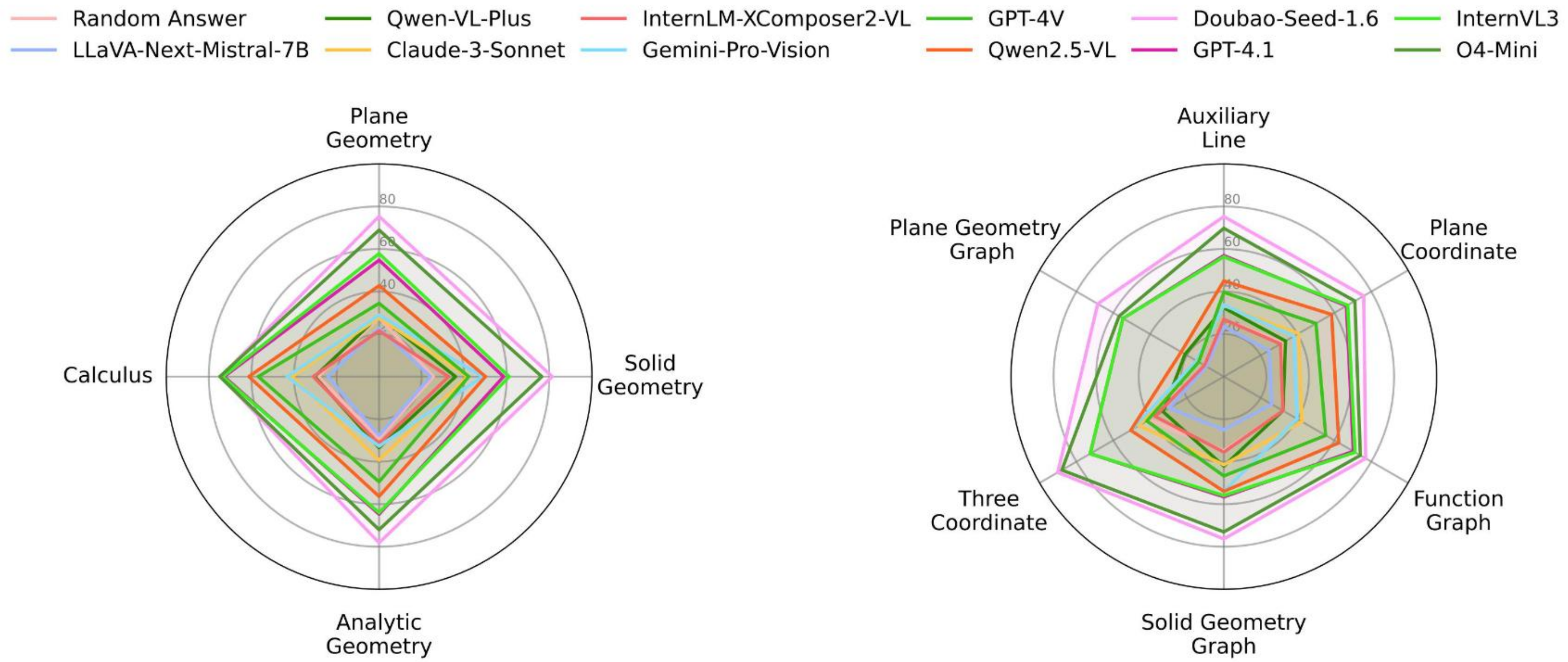
➢ Task 2: Direct Visual-aided Reasoning (D-VAR)

➢ Task 3: In-direct Visual-aided Reasoning (I-VAR)


Problem
Reasoning based on explicit visual context
Textual Reasoning
Problem
Visual reasoning to reveal implicit visual context
Y
X
Z
Interactive reasoning with visual aid
Textual Reasoning
Problem
Provided Visual Aid that reveal implicit visual context for certain reasoning path
Y
X
Z
Reasoning with visual aid
Textual Reasoning
(a) General Reasoning (GR)
(b) Direct Visual-Aided Reasoning (D-VAR)
(c) Indirect Visual-Aided Reasoning (I-VAR)
Task 1
Task 2

Categorization Comparison

- Problems with more spatial information utilization and inference are much harder to MLLMs



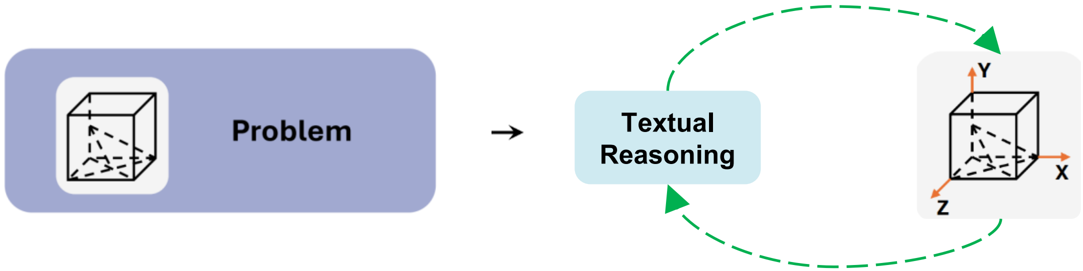
(a) Mathematical Branch

(b) Visual Aid

Accuracies of all LMM on visual-aided mathematical reasoning task across four mathematical branches and six visual aids

Visual-aided Reasoning

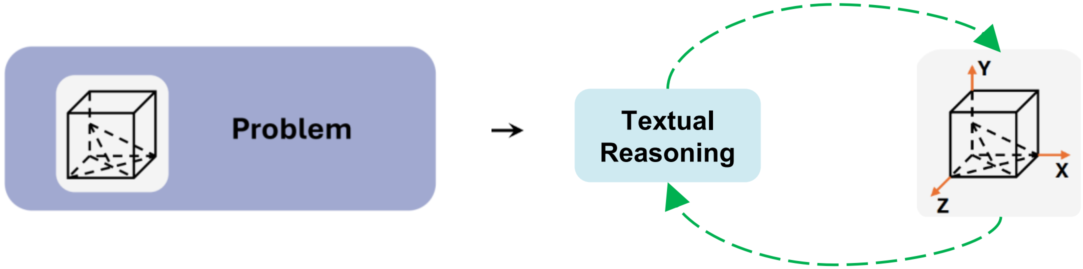
- Testbed: Text-only and multi-modal LLMs, ICL settings.
- Task: Direct Visual-aided Reasoning (D-VAR): Visual Context + Question => Visual-aids + Answer

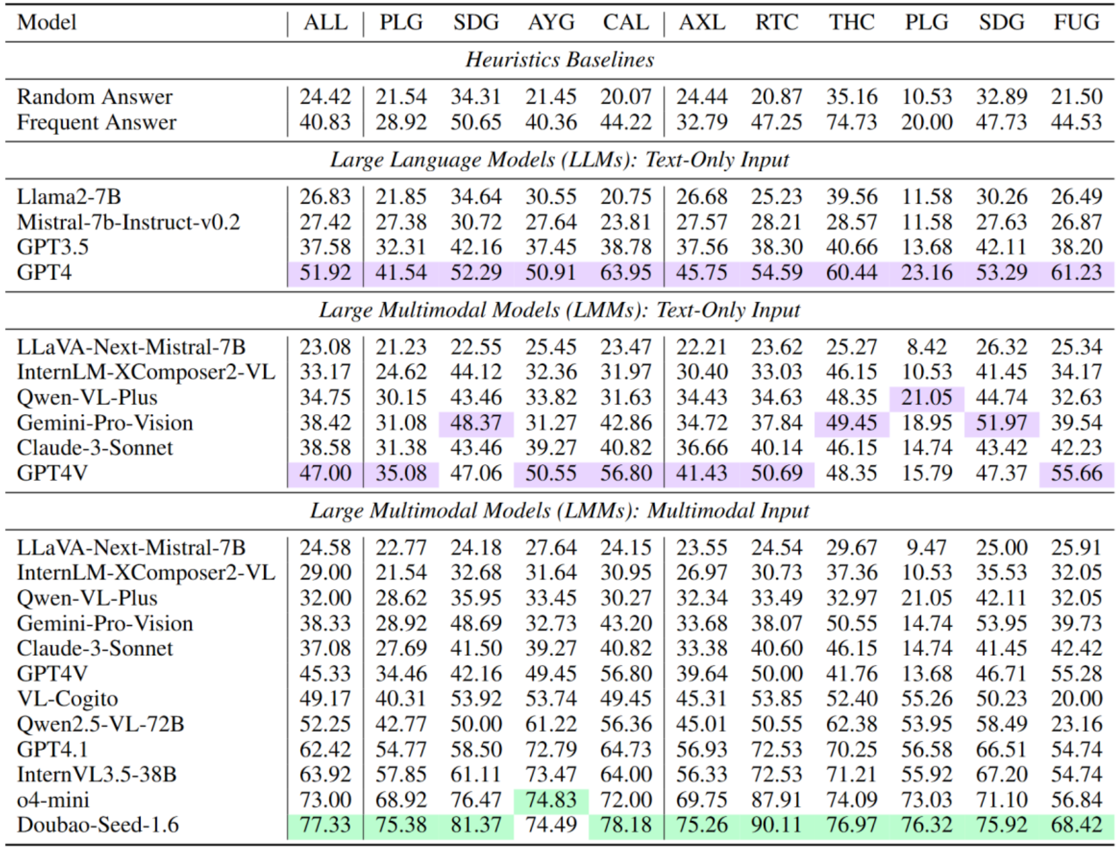


Visual-aided Reasoning

- Testbed: Text-only and multi-modal LLMs, ICL settings.
- Task: Direct Visual-aided Reasoning (D-VAR): Visual Context + Question => Visual-aids + Answer

➢ Doubao-Seed-1.6 **outperforms** most models across **all three modality** settings



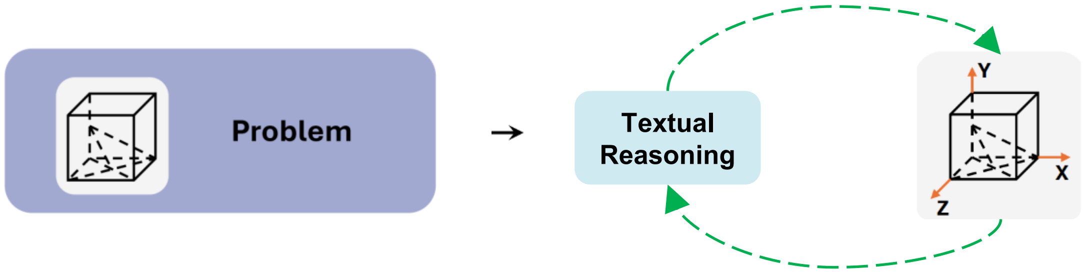


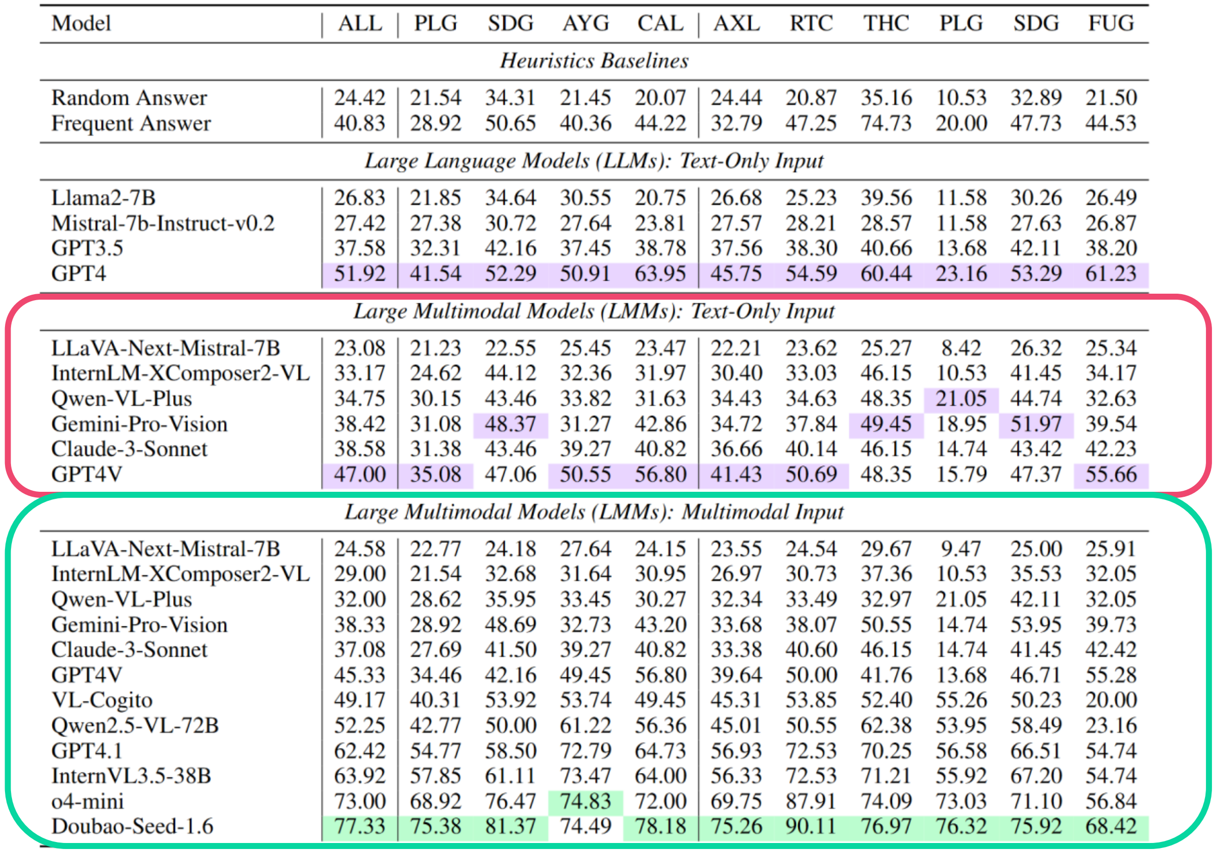
Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
Heuristics Baselines											
Random Answer	24.42	21.54	34.31	21.45	20.07	24.44	20.87	35.16	10.53	32.89	21.50
Frequent Answer	40.83	28.92	50.65	40.36	44.22	32.79	47.25	74.73	20.00	47.73	44.53
Large Language Models (LLMs): Text-Only Input											
Llama2-7B	26.83	21.85	34.64	30.55	20.75	26.68	25.23	39.56	11.58	30.26	26.49
Mistral-7b-Instruct-v0.2	27.42	27.38	30.72	27.64	23.81	27.57	28.21	28.57	11.58	27.63	26.87
GPT3.5	37.58	32.31	42.16	37.45	38.78	37.56	38.30	40.66	13.68	42.11	38.20
GPT4	51.92	41.54	52.29	50.91	63.95	45.75	54.59	60.44	23.16	53.29	61.23
Large Multimodal Models (LMMs): Text-Only Input											
LLaVA-Next-Mistral-7B	23.08	21.23	22.55	25.45	23.47	22.21	23.62	25.27	8.42	26.32	25.34
InternLM-XComposer2-VL	33.17	24.62	44.12	32.36	31.97	30.40	33.03	46.15	10.53	41.45	34.17
Qwen-VL-Plus	34.75	30.15	43.46	33.82	31.63	34.43	34.63	48.35	21.05	44.74	32.63
Gemini-Pro-Vision	38.42	31.08	48.37	31.27	42.86	34.72	37.84	49.45	18.95	51.97	39.54
Claude-3-Sonnet	38.58	31.38	43.46	39.27	40.82	36.66	40.14	46.15	14.74	43.42	42.23
GPT4V	47.00	35.08	47.06	50.55	56.80	41.43	50.69	48.35	15.79	47.37	55.66
Large Multimodal Models (LMMs): Multimodal Input											
LLaVA-Next-Mistral-7B	24.58	22.77	24.18	27.64	24.15	23.55	24.54	29.67	9.47	25.00	25.91
InternLM-XComposer2-VL	29.00	21.54	32.68	31.64	30.95	26.97	30.73	37.36	10.53	35.53	32.05
Qwen-VL-Plus	32.00	28.62	35.95	33.45	30.27	32.34	33.49	32.97	21.05	42.11	32.05
Gemini-Pro-Vision	38.33	28.92	48.69	32.73	43.20	33.68	38.07	50.55	14.74	53.95	39.73
Claude-3-Sonnet	37.08	27.69	41.50	39.27	40.82	33.38	40.60	46.15	14.74	41.45	42.42
GPT4V	45.33	34.46	42.16	49.45	56.80	39.64	50.00	41.76	13.68	46.71	55.28
VL-Cogito	49.17	40.31	53.92	53.74	49.45	45.31	53.85	52.40	55.26	50.23	20.00
Qwen2.5-VL-72B	52.25	42.77	50.00	61.22	56.36	45.01	50.55	62.38	53.95	58.49	23.16
GPT4.1	62.42	54.77	58.50	72.79	64.73	56.93	72.53	70.25	56.58	66.51	54.74
InternVL3.5-38B	63.92	57.85	61.11	73.47	64.00	56.33	72.53	71.21	55.92	67.20	54.74
o4-mini	73.00	68.92	76.47	74.83	72.00	69.75	87.91	74.09	73.03	71.10	56.84
Doubao-Seed-1.6	77.33	75.38	81.37	74.49	78.18	75.26	90.11	76.97	76.32	75.92	68.42

Visual-aided Reasoning

- Testbed: Text-only and multi-modal LLMs, ICL settings.
- Task: Direct Visual-aided Reasoning (D-VAR): Visual Context + Question => Visual-aids + Answer

- Doubao-Seed-1.6 **outperforms** most models across **all three modality** settings
- Cross-modality reasoning is challenging for current MLLMs
  - **Accuracy reduce** when performing spatial reasoning upon mathematical **image** instead of **caption**



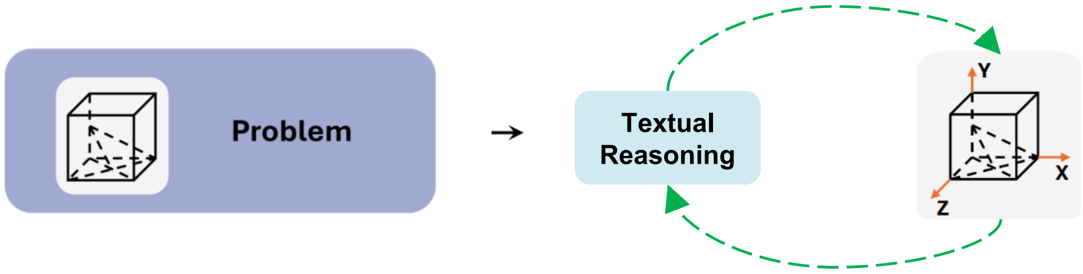


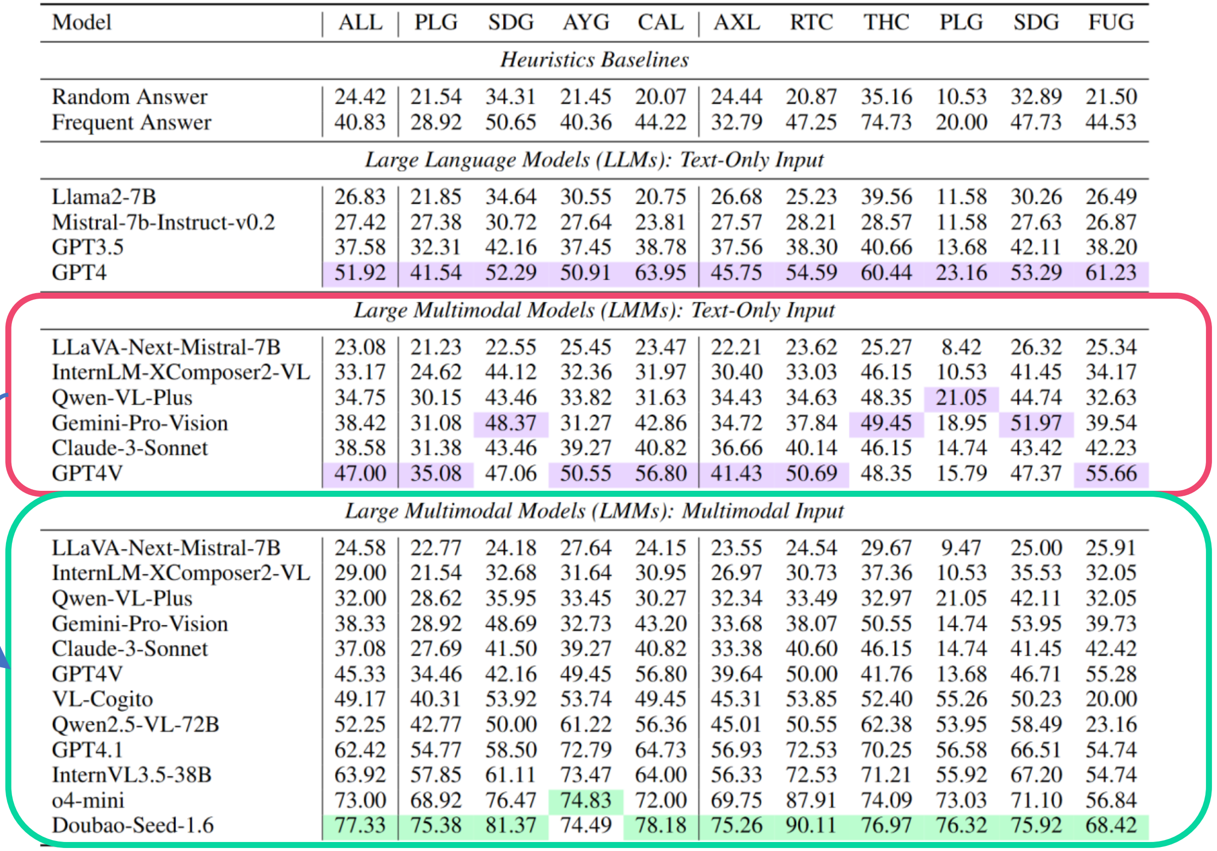
Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
Heuristics Baselines											
Random Answer	24.42	21.54	34.31	21.45	20.07	24.44	20.87	35.16	10.53	32.89	21.50
Frequent Answer	40.83	28.92	50.65	40.36	44.22	32.79	47.25	74.73	20.00	47.73	44.53
Large Language Models (LLMs): Text-Only Input											
Llama2-7B	26.83	21.85	34.64	30.55	20.75	26.68	25.23	39.56	11.58	30.26	26.49
Mistral-7b-Instruct-v0.2	27.42	27.38	30.72	27.64	23.81	27.57	28.21	28.57	11.58	27.63	26.87
GPT3.5	37.58	32.31	42.16	37.45	38.78	37.56	38.30	40.66	13.68	42.11	38.20
GPT4	51.92	41.54	52.29	50.91	63.95	45.75	54.59	60.44	23.16	53.29	61.23
Large Multimodal Models (LMMs): Text-Only Input											
LLaVA-Next-Mistral-7B	23.08	21.23	22.55	25.45	23.47	22.21	23.62	25.27	8.42	26.32	25.34
InternLM-XComposer2-VL	33.17	24.62	44.12	32.36	31.97	30.40	33.03	46.15	10.53	41.45	34.17
Qwen-VL-Plus	34.75	30.15	43.46	33.82	31.63	34.43	34.63	48.35	21.05	44.74	32.63
Gemini-Pro-Vision	38.42	31.08	48.37	31.27	42.86	34.72	37.84	49.45	18.95	51.97	39.54
Claude-3-Sonnet	38.58	31.38	43.46	39.27	40.82	36.66	40.14	46.15	14.74	43.42	42.23
GPT4V	47.00	35.08	47.06	50.55	56.80	41.43	50.69	48.35	15.79	47.37	55.66
Large Multimodal Models (LMMs): Multimodal Input											
LLaVA-Next-Mistral-7B	24.58	22.77	24.18	27.64	24.15	23.55	24.54	29.67	9.47	25.00	25.91
InternLM-XComposer2-VL	29.00	21.54	32.68	31.64	30.95	26.97	30.73	37.36	10.53	35.53	32.05
Qwen-VL-Plus	32.00	28.62	35.95	33.45	30.27	32.34	33.49	32.97	21.05	42.11	32.05
Gemini-Pro-Vision	38.33	28.92	48.69	32.73	43.20	33.68	38.07	50.55	14.74	53.95	39.73
Claude-3-Sonnet	37.08	27.69	41.50	39.27	40.82	33.38	40.60	46.15	14.74	41.45	42.42
GPT4V	45.33	34.46	42.16	49.45	56.80	39.64	50.00	41.76	13.68	46.71	55.28
VL-Cogito	49.17	40.31	53.92	53.74	49.45	45.31	53.85	52.40	55.26	50.23	20.00
Qwen2.5-VL-72B	52.25	42.77	50.00	61.22	56.36	45.01	50.55	62.38	53.95	58.49	23.16
GPT4.1	62.42	54.77	58.50	72.79	64.73	56.93	72.53	70.25	56.58	66.51	54.74
InternVL3.5-38B	63.92	57.85	61.11	73.47	64.00	56.33	72.53	71.21	55.92	67.20	54.74
o4-mini	73.00	68.92	76.47	74.83	72.00	69.75	87.91	74.09	73.03	71.10	56.84
Doubao-Seed-1.6	77.33	75.38	81.37	74.49	78.18	75.26	90.11	76.97	76.32	75.92	68.42

Visual-aided Reasoning

- Testbed: Text-only and multi-modal LLMs, ICL settings.
- Task: Direct Visual-aided Reasoning (D-VAR): Visual Context + Question => Visual-aids + Answer

- ➢ Doubao-Seed-1.6 **outperforms** most models across **all three modality** settings
- ➢ Cross-modality reasoning is challenging for current MLLMs
  - **Accuracy reduce** when performing spatial reasoning upon mathematical **image** instead of **caption**



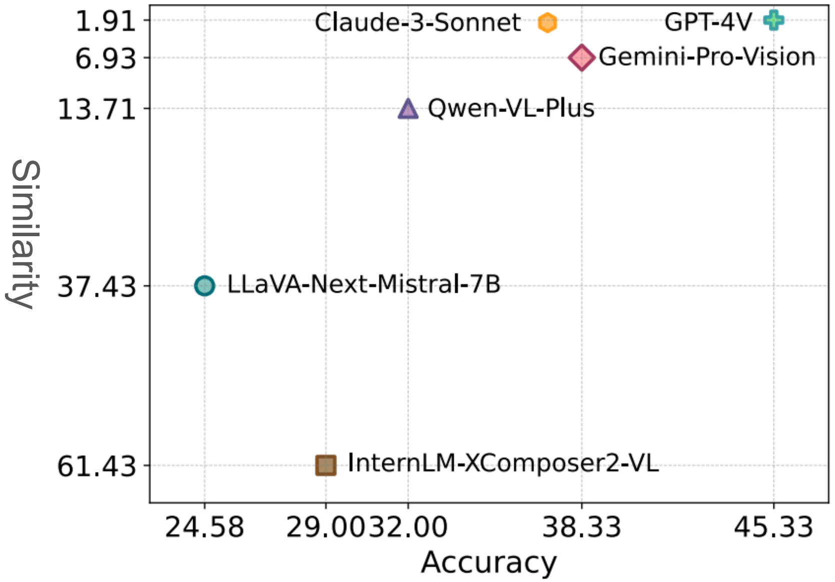


Model	ALL	PLG	SDG	AYG	CAL	AXL	RTC	THC	PLG	SDG	FUG
Heuristics Baselines											
Random Answer	24.42	21.54	34.31	21.45	20.07	24.44	20.87	35.16	10.53	32.89	21.50
Frequent Answer	40.83	28.92	50.65	40.36	44.22	32.79	47.25	74.73	20.00	47.73	44.53
Large Language Models (LLMs): Text-Only Input											
Llama2-7B	26.83	21.85	34.64	30.55	20.75	26.68	25.23	39.56	11.58	30.26	26.49
Mistral-7b-Instruct-v0.2	27.42	27.38	30.72	27.64	23.81	27.57	28.21	28.57	11.58	27.63	26.87
GPT3.5	37.58	32.31	42.16	37.45	38.78	37.56	38.30	40.66	13.68	42.11	38.20
GPT4	51.92	41.54	52.29	50.91	63.95	45.75	54.59	60.44	23.16	53.29	61.23
Large Multimodal Models (LMMs): Text-Only Input											
LLaVA-Next-Mistral-7B	23.08	21.23	22.55	25.45	23.47	22.21	23.62	25.27	8.42	26.32	25.34
InternLM-XComposer2-VL	33.17	24.62	44.12	32.36	31.97	30.40	33.03	46.15	10.53	41.45	34.17
Qwen-VL-Plus	34.75	30.15	43.46	33.82	31.63	34.43	34.63	48.35	21.05	44.74	32.63
Gemini-Pro-Vision	38.42	31.08	48.37	31.27	42.86	34.72	37.84	49.45	18.95	51.97	39.54
Claude-3-Sonnet	38.58	31.38	43.46	39.27	40.82	36.66	40.14	46.15	14.74	43.42	42.23
GPT4V	47.00	35.08	47.06	50.55	56.80	41.43	50.69	48.35	15.79	47.37	55.66
Large Multimodal Models (LMMs): Multimodal Input											
LLaVA-Next-Mistral-7B	24.58	22.77	24.18	27.64	24.15	23.55	24.54	29.67	9.47	25.00	25.91
InternLM-XComposer2-VL	29.00	21.54	32.68	31.64	30.95	26.97	30.73	37.36	10.53	35.53	32.05
Qwen-VL-Plus	32.00	28.62	35.95	33.45	30.27	32.34	33.49	32.97	21.05	42.11	32.05
Gemini-Pro-Vision	38.33	28.92	48.69	32.73	43.20	33.68	38.07	50.55	14.74	53.95	39.73
Claude-3-Sonnet	37.08	27.69	41.50	39.27	40.82	33.38	40.60	46.15	14.74	41.45	42.42
GPT4V	45.33	34.46	42.16	49.45	56.80	39.64	50.00	41.76	13.68	46.71	55.28
VL-Cogito	49.17	40.31	53.92	53.74	49.45	45.31	53.85	52.40	55.26	50.23	20.00
Qwen2.5-VL-72B	52.25	42.77	50.00	61.22	56.36	45.01	50.55	62.38	53.95	58.49	23.16
GPT4.1	62.42	54.77	58.50	72.79	64.73	56.93	72.53	70.25	56.58	66.51	54.74
InternVL3.5-38B	63.92	57.85	61.11	73.47	64.00	56.33	72.53	71.21	55.92	67.20	54.74
o4-mini	73.00	68.92	76.47	74.83	72.00	69.75	87.91	74.09	73.03	71.10	56.84
Doubao-Seed-1.6	77.33	75.38	81.37	74.49	78.18	75.26	90.11	76.97	76.32	75.92	68.42

FIND MORE EXPERIMENTS ON OTHER TASKS IN APPENDIX

Reasoning Comparison

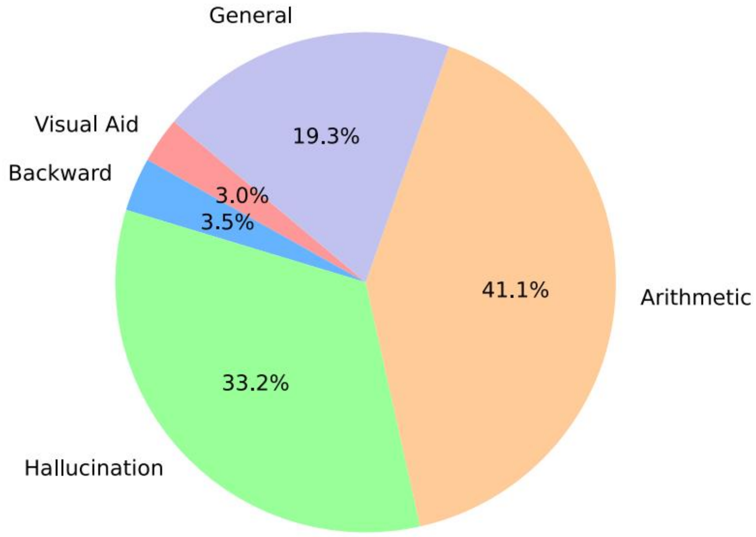
- **Observation 1: Low similarity** between **general reasoning** and **visual-aided reasoning** answers
  - ➢ Visual-aided reasoning task **differs significantly** from general reasoning tasks



(a) N-gram similarity of Answer between general reasoning (CQ2A) and visual-aided reasoning (CQ2VA).

Reasoning Tendency

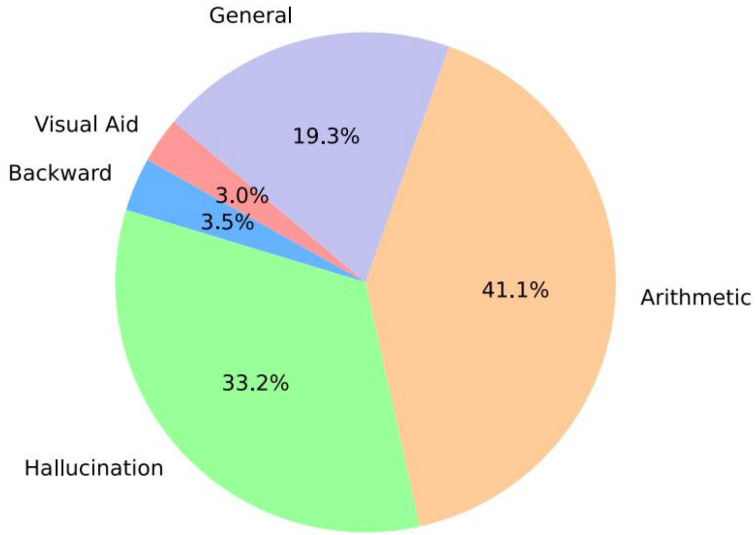
- Reasoning Trajectories:
  - General: Correct reasoning without relying on visual aids.
  - Arithmetic: Correct reasoning using pure arithmetic methods.
  - Visual-Aided: Correct reasoning incorporating the use of visual aids.
  - Backward: Correct reasoning derived from provided choices or the final conclusion.
  - Hallucination



Model reasoning patterns in direct mathematical problem solving with visual context (CQ2VA).

Reasoning Tendency

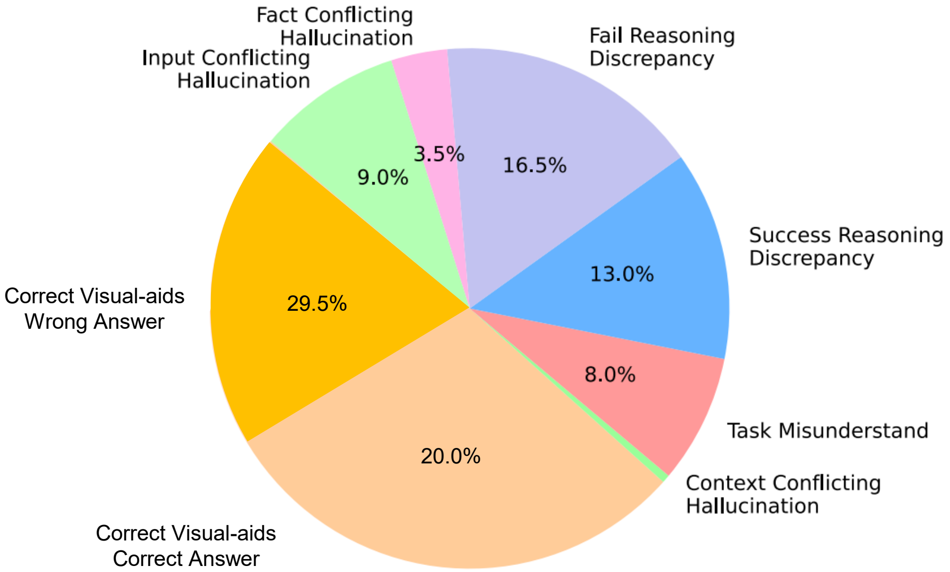
- Reasoning Trajectories:
  - General: Correct reasoning without relying on visual aids.
  - Arithmetic: Correct reasoning using pure arithmetic methods.
  - Visual-Aided: Correct reasoning incorporating the use of visual aids.
  - Backward: Correct reasoning derived from provided choices or the final conclusion.
  - Hallucination
- MLLMs tend to proceed reasoning along a **text-only trajectory**, **disregarding** the potential benefits of visual aids



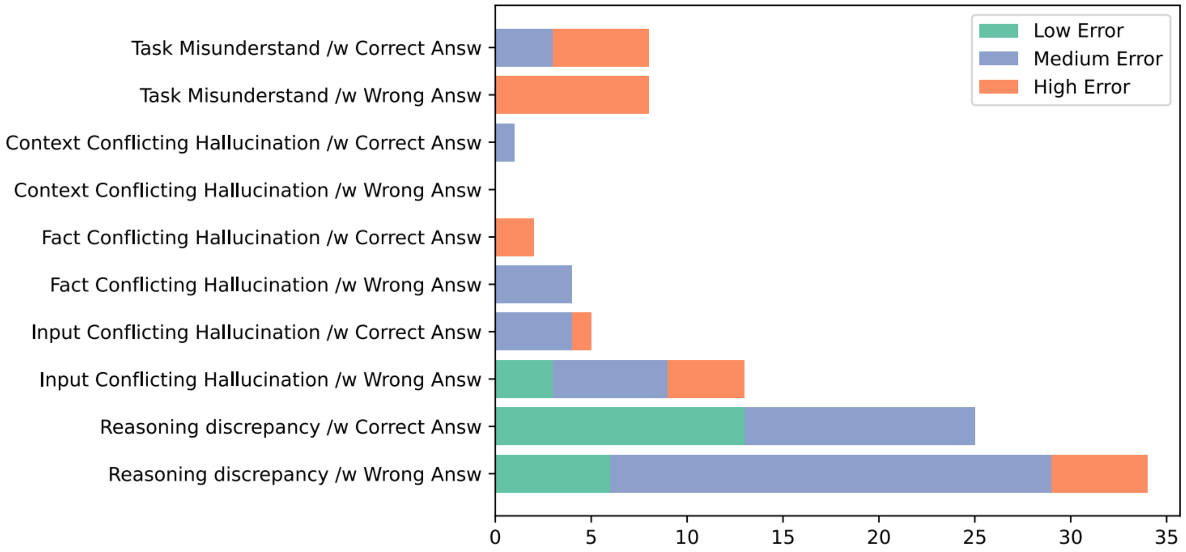
Model reasoning patterns in direct mathematical problem solving with visual context (CQ2VA).

Visual-aids Inference

- Critical Factors
  - Hallucination
  - Poor Task Understanding
  - Low performance on reasoning based on correct visual-aids



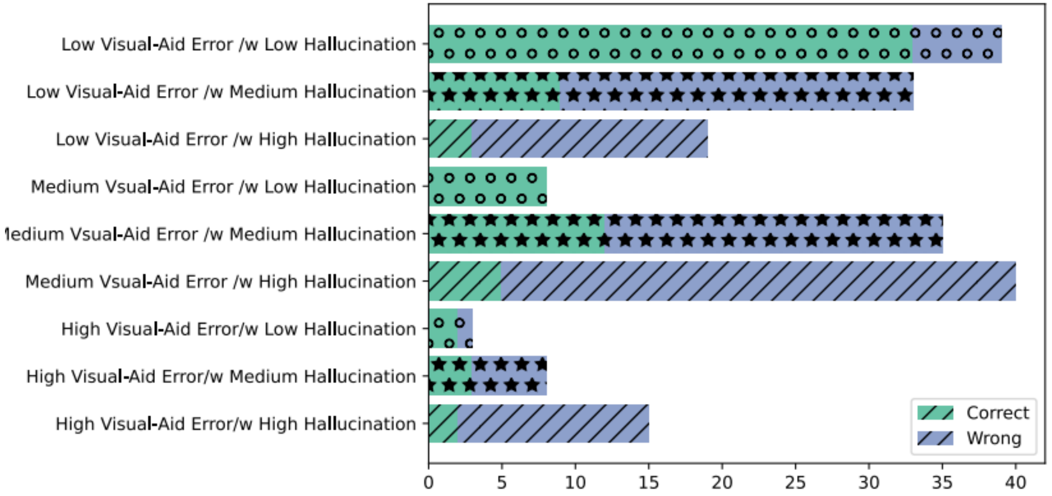
Model distributions of generated visual-aids during visual-aided reasoning process (CQ2VA).



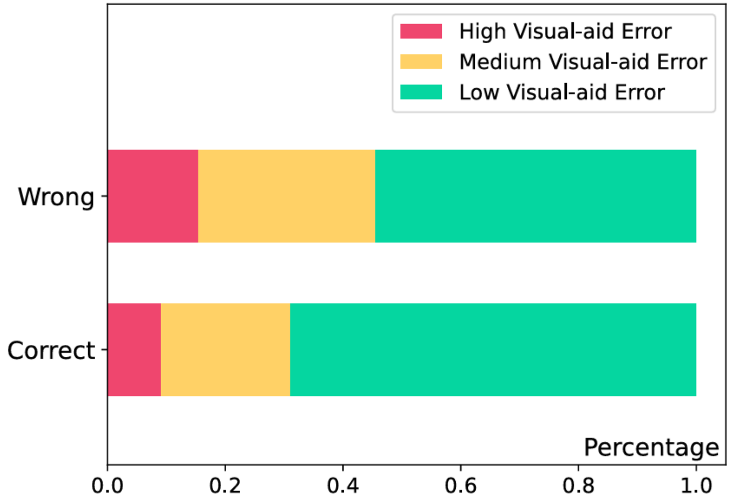
Correlation between error causes of visual aid and answer correctness

Visual-aids and Reasoning Hallucination

- Factor: Low performance on reasoning based on correct visual-aids
- Correct visual aids can
  - effectively **alleviate hallucinations** during **reasoning**
  - significantly **increase** the **success rate** of the **reasoning** process
- Hallucination in reasoning offsets the positive effect of correct visual-aids



Correlation between visual-aids and reasoning hallucination.



Correlation between errors of visual-aids and answer correctness.

Conclusions

Conclusion

- Cross-modality evaluation aspects are rarely taken into account in the evaluation

Conclusions

Conclusion

- Cross-modality evaluation aspects are rarely taken into account in the evaluation
- **Deficiencies** of mainstream LLMs in **deducing visual aids** and the **corresponding textual reasoning** steps

Conclusions

Conclusion

- Cross-modality evaluation aspects are rarely taken into account in the evaluation
- **Deficiencies** of mainstream LLMs in **deducing visual aids** and the **corresponding textual reasoning** steps
- We propose a benchmark focus on evaluating more **comprehensive cross-modality evaluation**

Conclusions

Conclusion

- Cross-modality evaluation aspects are rarely taken into account in the evaluation
- **Deficiencies** of mainstream LLMs in **deducing visual aids** and the **corresponding textual reasoning** steps
- We propose a benchmark focus on evaluating more **comprehensive cross-modality evaluation**
- **Significant impact of hallucination** in both visual-aid inference and visual-aided reasoning demonstrate **models' lack of confidence** in this novel cross-modality task.

Conclusions

Conclusion

- Cross-modality evaluation aspects are rarely taken into account in the evaluation
- **Deficiencies** of mainstream LLMs in **deducing visual aids** and the **corresponding textual reasoning** steps
- We propose a benchmark focus on evaluating more **comprehensive cross-modality evaluation**
- **Significant impact of hallucination** in both visual-aid inference and visual-aided reasoning demonstrate **models' lack of confidence** in this novel cross-modality task.

Future Work

- ➢ Further explore cause of weak visual-aids inference and visual-aided reasoning
- ➢ Propose fine-grained metrics to evaluate visual-aids inference capability

Thank You!