GePBench: Evaluating Fundamental Geometric Perception for Multimodal Large Language Models

Natural Language Processing Group, Nanjing University

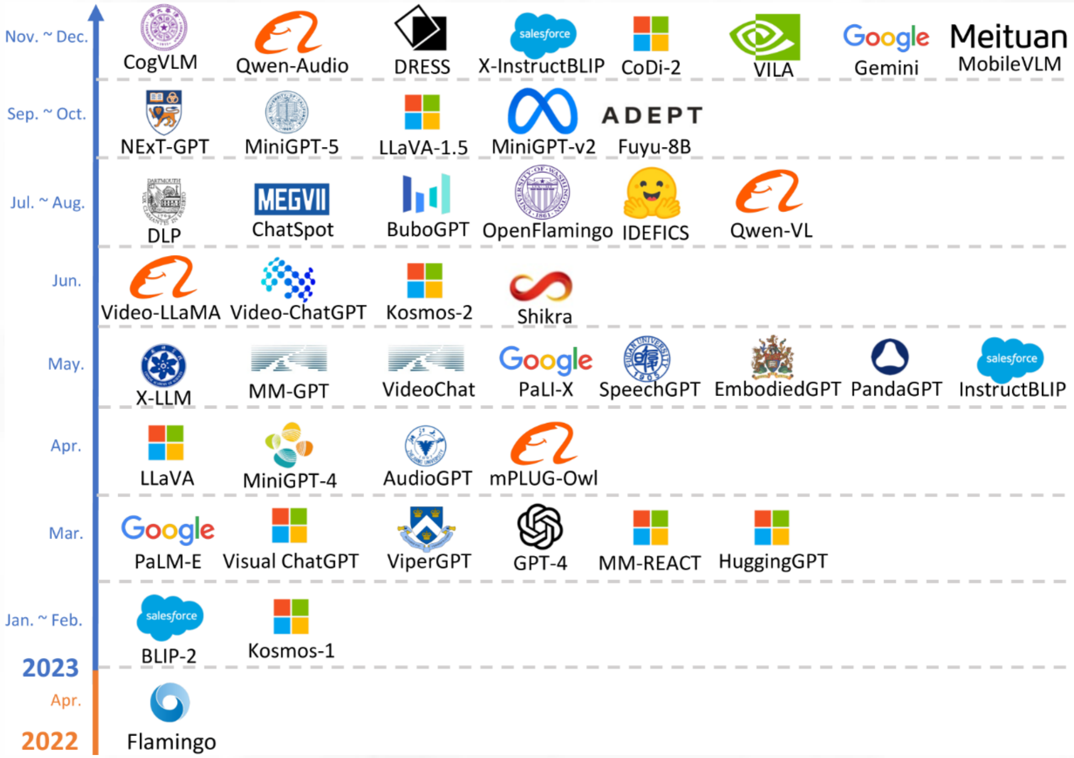
邢尚禹　2025.03

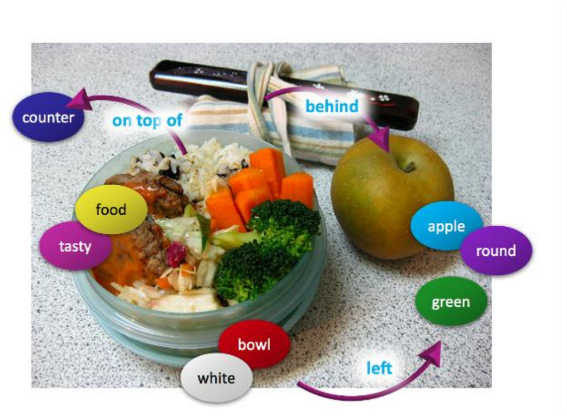
Scan for paper

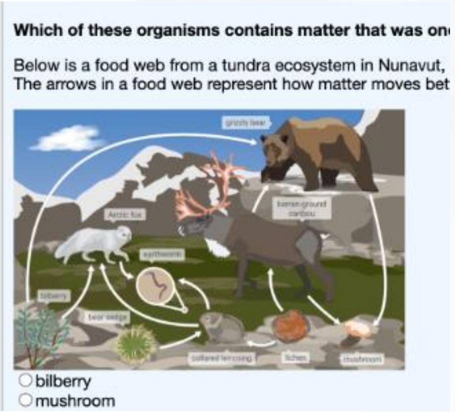
Scan for slides

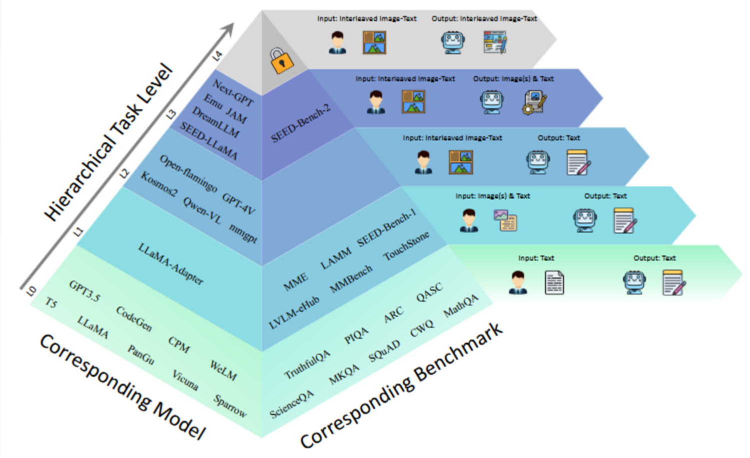
01 Introduction

- Multimodal large language models have achieved significant breakthroughs
- Various benchmarks are proposed to measure their capabilities



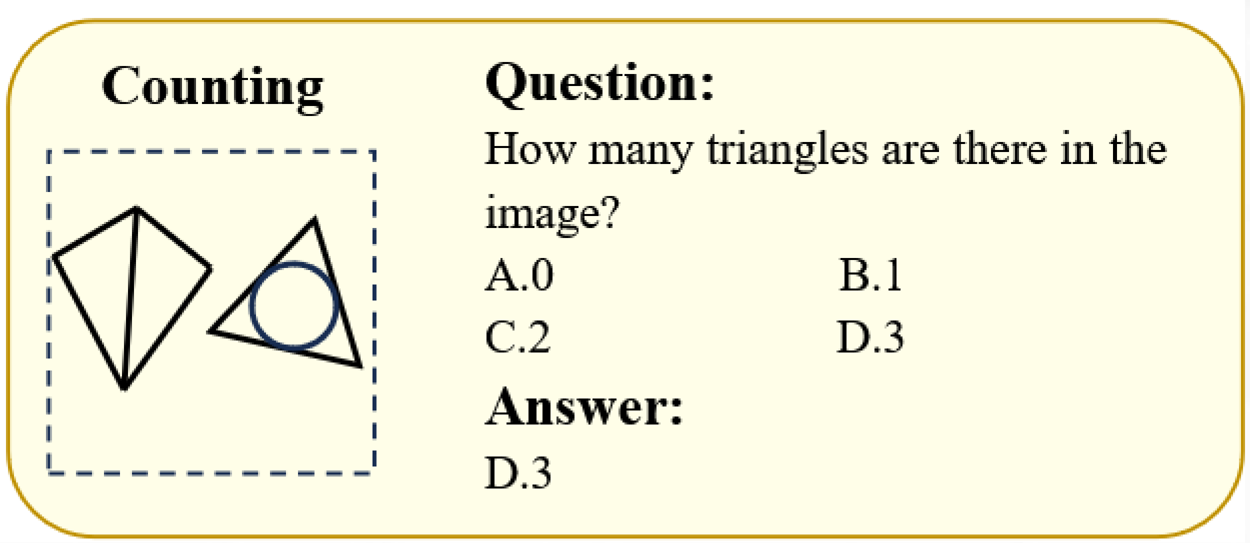






01 Introduction

- A critical gap:
  - focus on real-world scenarios and assess high-level semantic understanding
  - neglect fundamental perceptual challenges such as geometric perception



01 Introduction

Geometric shapes are worth studying because they

1. Provide an ideal testbed for fundamental visual capabilities

    - Geometric shapes require an understanding of spatial relationships and visual details for effective perception

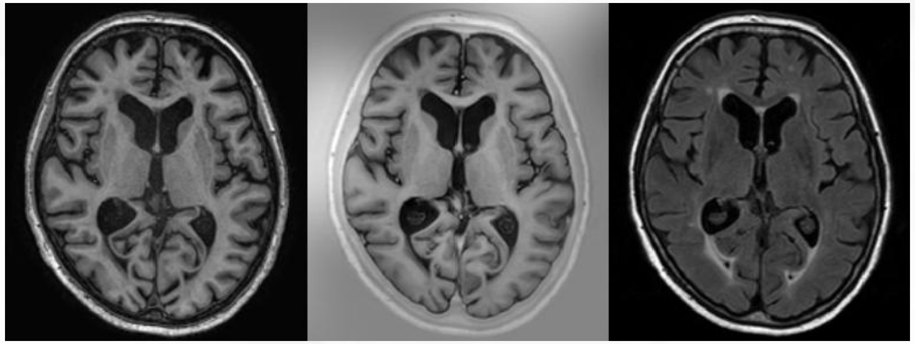
    - More complex tasks like visual reasoning and decision-making build on these basic perceptual abilities

01 **Introduction**

Geometric shapes are worth studying because they

2. Lay the foundation for a wide range of downstream applications

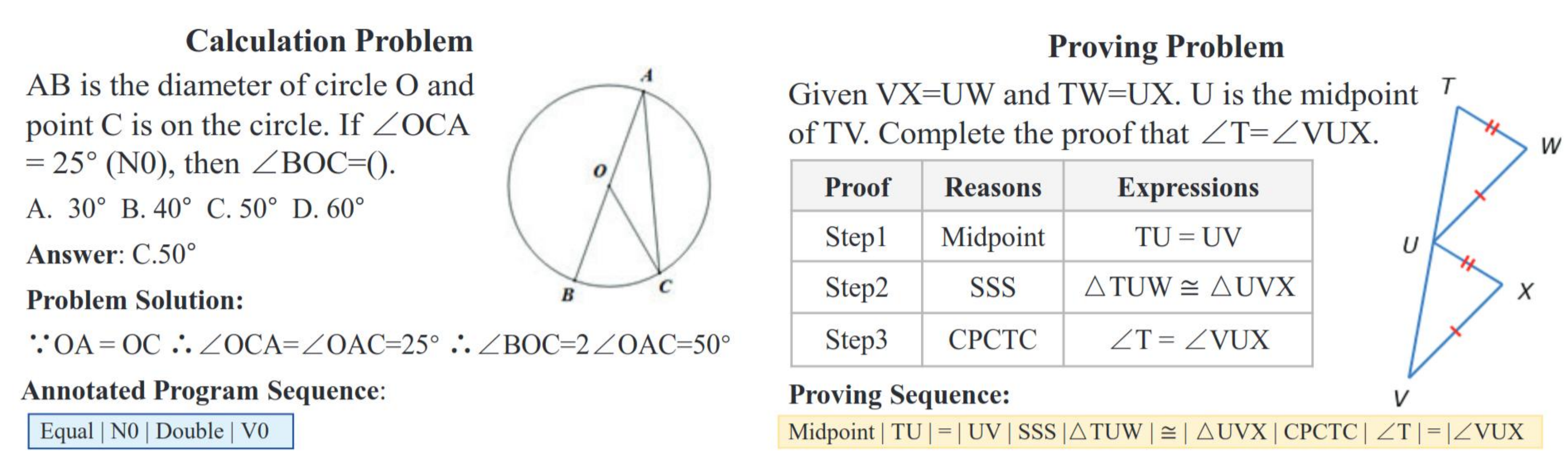
- Tasks like medical image analysis and fossil classification rely heavily on precise spatial perception and the interpretation of abstract visual patterns

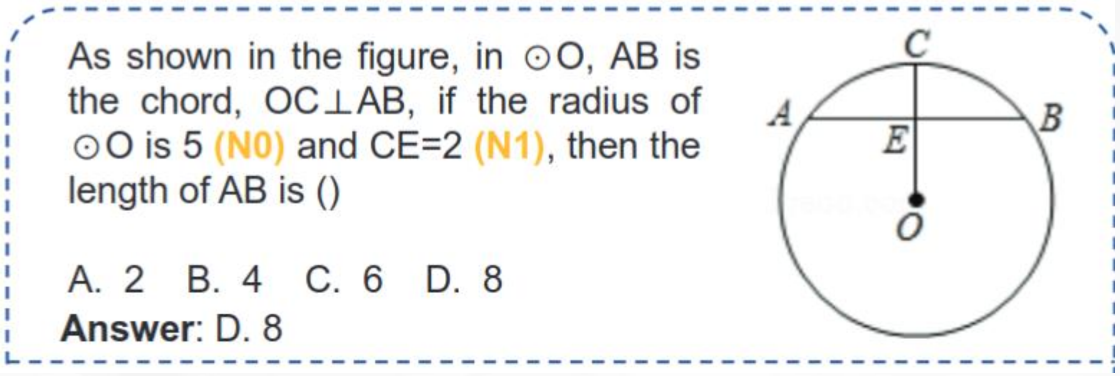


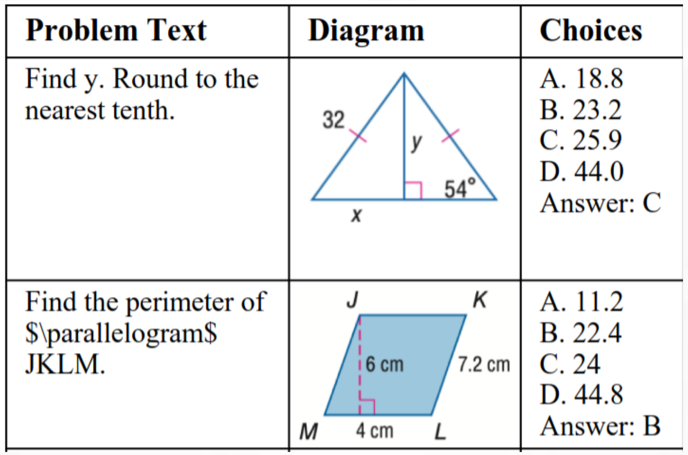
02 Related Work

Other datasets involving geometric figure:

- GeoQA
- Geometry3K
- UniGeo
- GeomVerse
- GeoMM
- MAVIS







Problem Text	Diagram	Choices
Find y. Round to the nearest tenth.	32, y, 54°, x	A. 18.8 B. 23.2 C. 25.9 D. 44.0 Answer: C
Find the perimeter of $\parallelogram$ JKLM.	J, K, 6 cm, 7.2 cm, M, 4 cm, L	A. 11.2 B. 22.4 C. 24 D. 44.8 Answer: B

02 Related Work

Other datasets involving geometric figure …

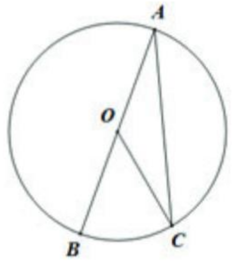
- Their focus is on mathematical reasoning capability
  - Tasks include numeric calculations, proof generation, relationship inference

- They depend on basic perceptual skills like spatial awareness and shape recognition, which we explicitly address

Calculation Problem

AB is the diameter of circle O and point C is on the circle. If $\angle OCA = 25°$ (N0), then $\angle BOC$=().

A. 30° B. 40° C. 50° D. 60°



Answer: C.50°

Problem Solution:

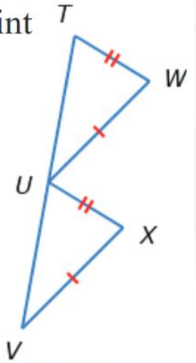
$\because OA = OC \therefore \angle OCA=\angle OAC=25° \therefore \angle BOC=2\angle OAC=50°$

Annotated Program Sequence:

Equal | N0 | Double | V0

Proving Problem

Given VX=UW and TW=UX. U is the midpoint of TV. Complete the proof that $\angle T=\angle VUX$.



Proof	Reasons	Expressions
Step1	Midpoint	TU = UV
Step2	SSS	$\triangle TUW \cong \triangle UVX$
Step3	CPCTC	$\angle T = \angle VUX$

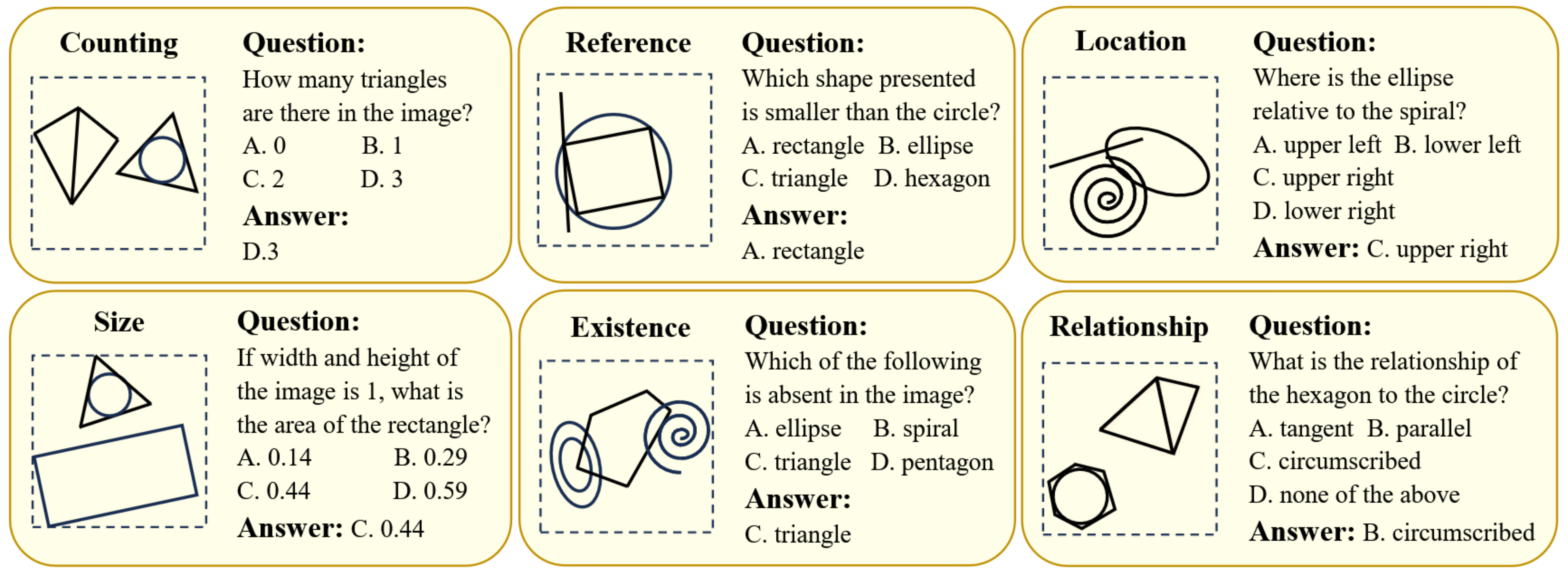
Proving Sequence:

Midpoint | TU | = | UV | SSS | $\triangle$TUW | $\cong$ | $\triangle$UVX | CPCTC | $\angle$T | = | $\angle$VUX

03 Our Dataset

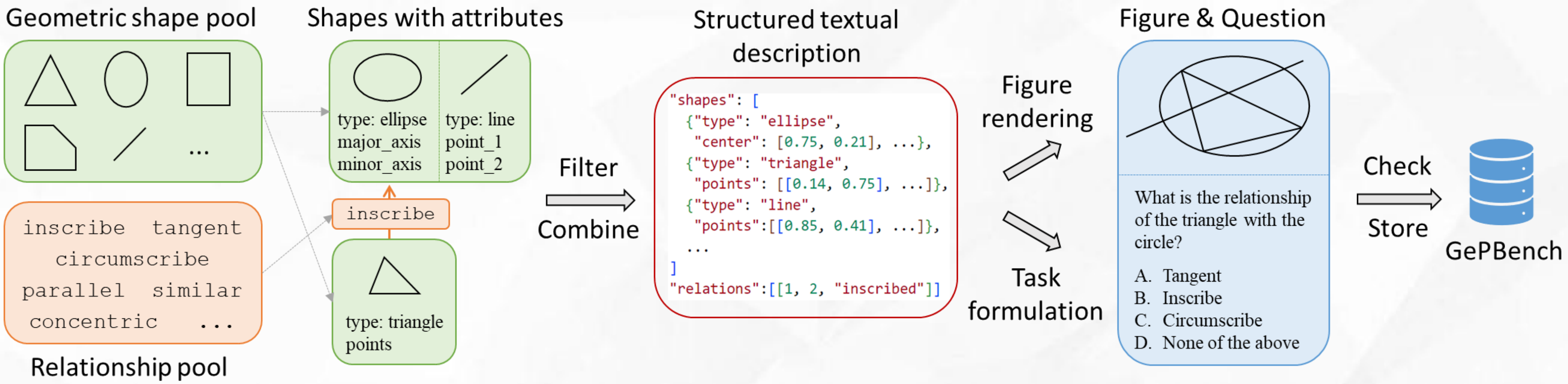
GePBench: a Novel Geometric Perception Benchmark

- 80K images and 285K standard multiple-choice questions
- 6 key dimensions of spatial and shape perception



03 Our Dataset

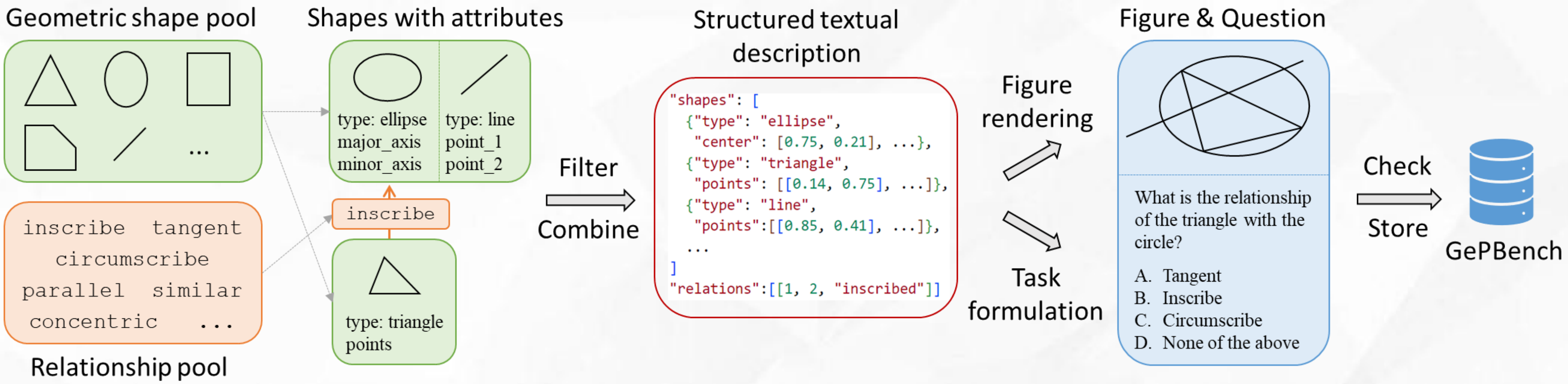
Data construction pipeline:



1. Structured Description Generation (→ descriptions)
   i. Sample shapes from pool and randomly assign attributes
   ii. Sample relationships and shapes from pool and add to figure

03 Our Dataset

Data construction pipeline:

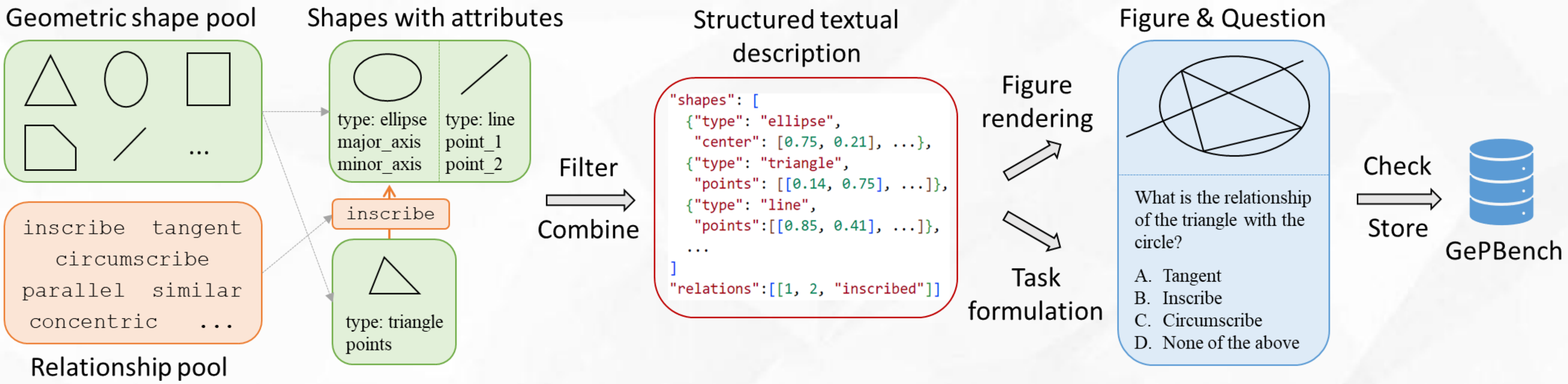


2. Figure rendering (descriptions → image)

i. Use Matplotlib to draw figure and add noise to part of the shapes

03 Our Dataset

Data construction pipeline:



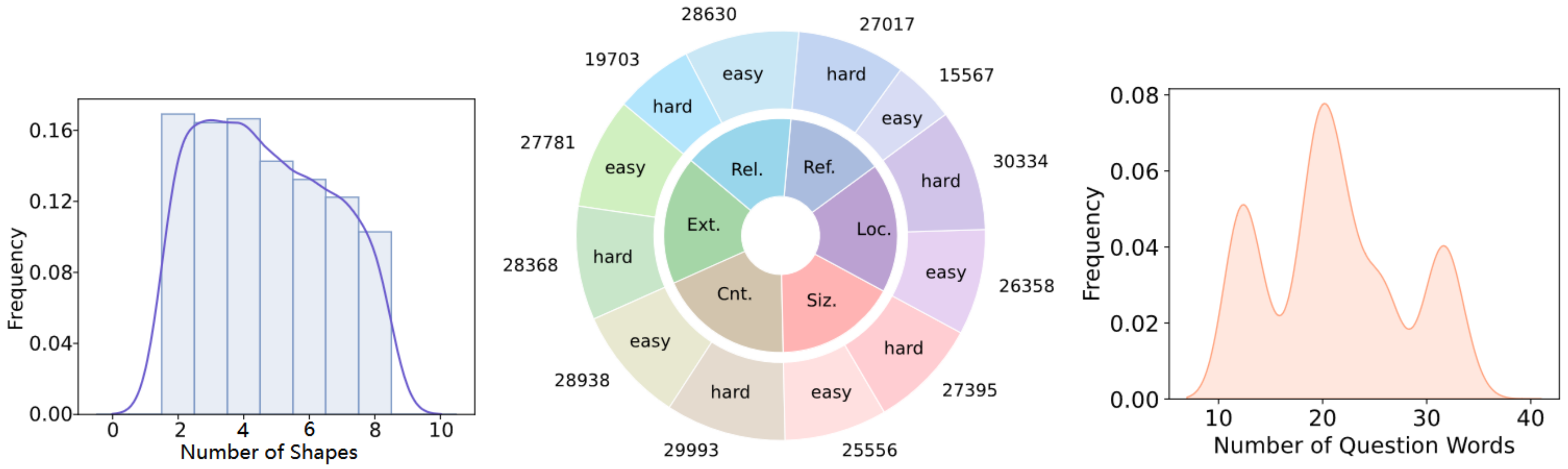
3. Task Formulation (descriptions → questions)

i. Create questions using pre-defined templates for each of the 6 dimensions

ii. Category into easy/hard split according to shape number and noise level

03 Our Dataset

Key statistics:



(a) Number of shapes per figure.

(b) Sizes of different aspects.

(c) Number of words per question.

Figure 3: Key data distributions of GePBench.

04 Experiments

Model Class	Size	Avg.	Easy Ext.	Easy Cnt.	Easy Siz.	Easy Loc.	Easy Ref.	Easy Rel.	Hard Ext.	Hard Cnt.	Hard Siz.	Hard Loc.	Hard Ref.	Hard Rel.
Random guessing	-	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Human	-	99.3	99.8	99.6	98.5	98.7	99.3	99.9	99.1	99.5	98.9	98.7	99.4	99.6
GPT-4o	-	59.3	**78.9**	62.6	17.5	68.4	78.6	70.7	71.7	55.7	20.0	67.6	68.0	52.2
Claude-3.5-Sonnet	-	50.3	76.2	69.6	17.9	55.3	62.5	64.5	**72.4**	57.8	16.9	24.9	30.4	54.7
Gemini-1.5-pro	-	68.4	72.7	**74.1**	**64.0**	**74.0**	**80.3**	**75.0**	66.9	**64.0**	**44.4**	**71.5**	**73.3**	**61.1**
BLIP2	3B	34.1	40.1	14.3	30.7	25.2	54.2	38.2	39.7	28.9	**35.1**	26.0	46.0	31.3
InstructBLIP	3B	34.1	40.9	26.4	20.4	26.7	59.4	41.8	42.1	23.7	15.0	28.8	52.3	31.5
MiniGPTv2	7B	31.4	29.8	42.6	28.2	24.6	33.7	32.7	27.5	40.4	28.1	26.2	31.9	30.6
LLaVA-1.5	7B	39.2	40.0	48.9	28.4	45.5	42.1	51.2	32.1	31.0	28.0	40.2	38.1	45.1
LLaVA-1.5	13B	40.8	46.2	57.4	13.0	51.9	52.0	47.4	37.3	39.4	11.4	49.5	45.1	39.2
LLaVA-OneVision	7B	56.7	61.7	72.8	29.8	57.8	77.4	64.5	57.8	55.7	31.3	57.1	71.3	43.6
LLaVA-OneVision	72B	61.7	75.8	**74.3**	25.0	62.2	81.2	76.7	67.9	58.7	24.0	65.1	76.6	52.9
mPLUG-Owl3	7B	46.8	56.7	66.3	26.1	33.0	62.0	53.4	56.3	54.5	21.4	33.7	60.7	37.8
InternVL-2.5	8B	55.7	68.7	64.8	15.4	64.2	72.8	63.8	64.7	52.0	20.8	67.2	62.0	51.6
InternVL-2.5	26B	61.1	70.6	64.1	25.5	63.2	81.3	74.1	67.2	56.8	27.4	**74.8**	73.6	55.1
InternVL-2.5	78B	**65.2**	75.7	72.1	**37.3**	72.6	80.9	**77.4**	**72.1**	**62.1**	25.3	73.3	**78.2**	**55.5**
MiniCPM-V-2.6	8B	57.4	68.9	69.8	33.5	58.6	78.6	53.4	61.9	54.8	29.8	58.7	74.6	46.2
GLM-4V	9B	54.2	64.2	73.9	20.2	52.7	80.9	62.7	50.7	54.3	19.1	54.4	72.6	44.9
Mantis-Idefics2	8B	47.5	60.0	65.9	15.6	43.0	68.4	49.5	56.3	50.0	13.9	48.7	63.9	35.2
Qwen2-VL	7B	56.5	65.8	72.3	22.6	62.2	82.7	62.4	59.7	55.2	17.6	60.6	74.5	42.3
Qwen2-VL	72B	63.0	**76.7**	71.9	25.1	**77.3**	**85.9**	70.1	67.3	57.2	25.1	73.4	76.5	49.1
LLaMA-3.2-Vision	90B	55.3	61.1	64.4	19.6	67.3	71.6	68.9	58.3	54.9	21.2	62.2	64.3	50.2

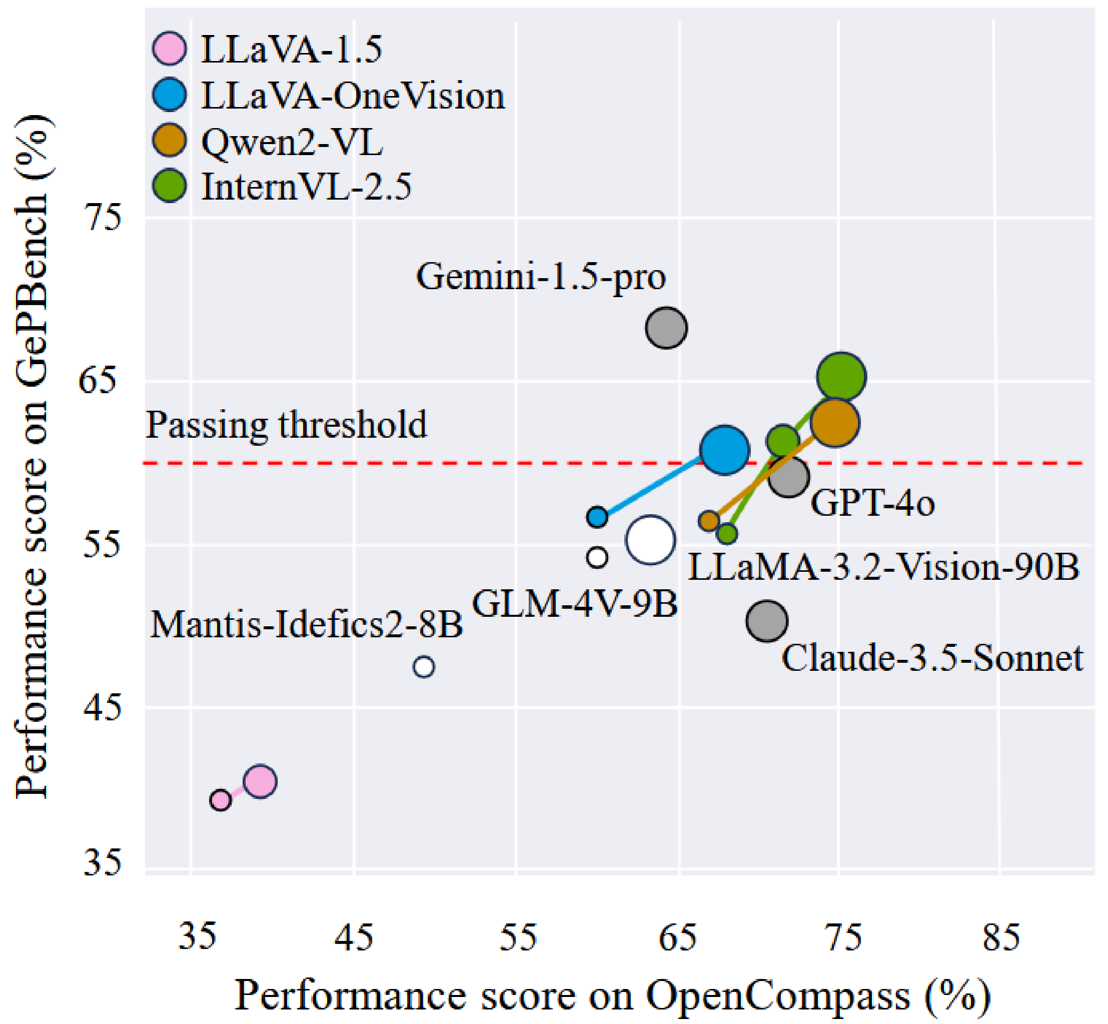
04 Experiments

Main observations:

1. Both closed-source and open-source models face significant challenges
   - Few reach the passing threshold

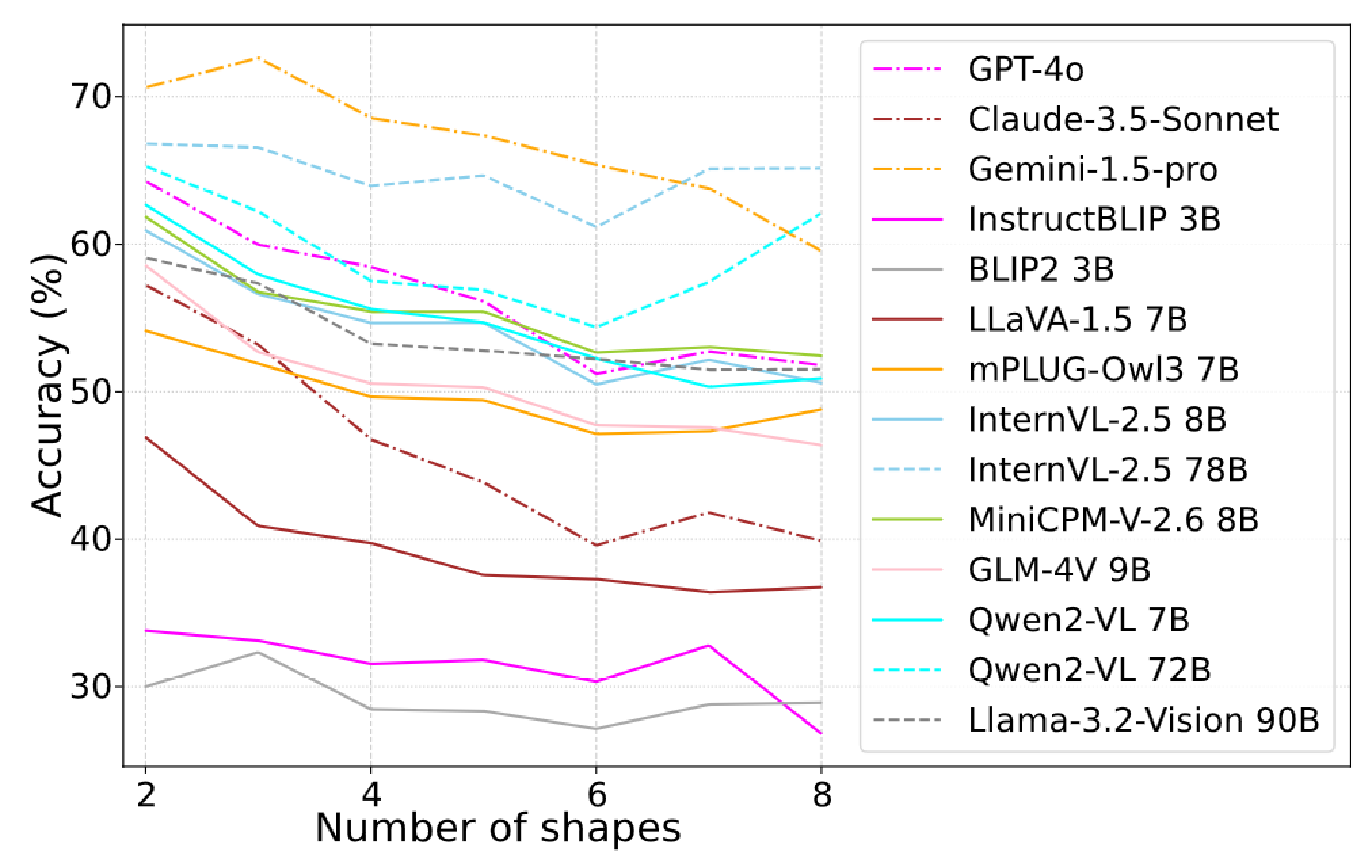
2. Scaling model size yields limited improvements
   - Compared with OpenCompass, improvements are lower

3. Size and Location are generally more challenging than other aspect
   - Most models perform worse than random guessing on Size aspect



04 Experiments

Ablation study on number of shapes per image



04 Experiments

Ablation study on visual encoders

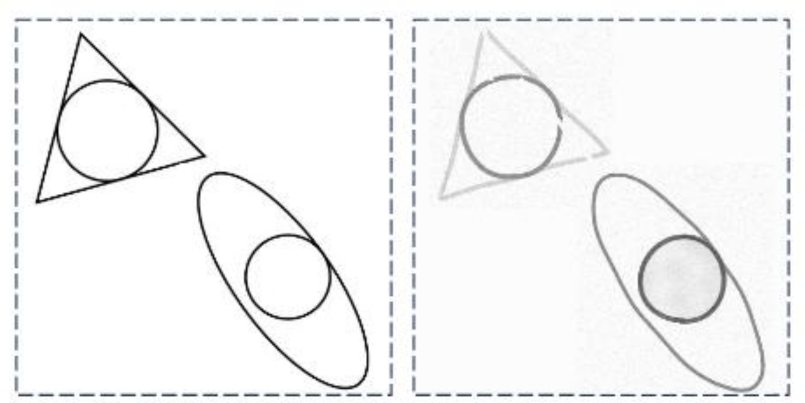
- Higher resolution improves detail recognition but impacts spatial accuracy
- Different encoders specialize in different dimensions
- Mixed encoders underperform in geometric tasks

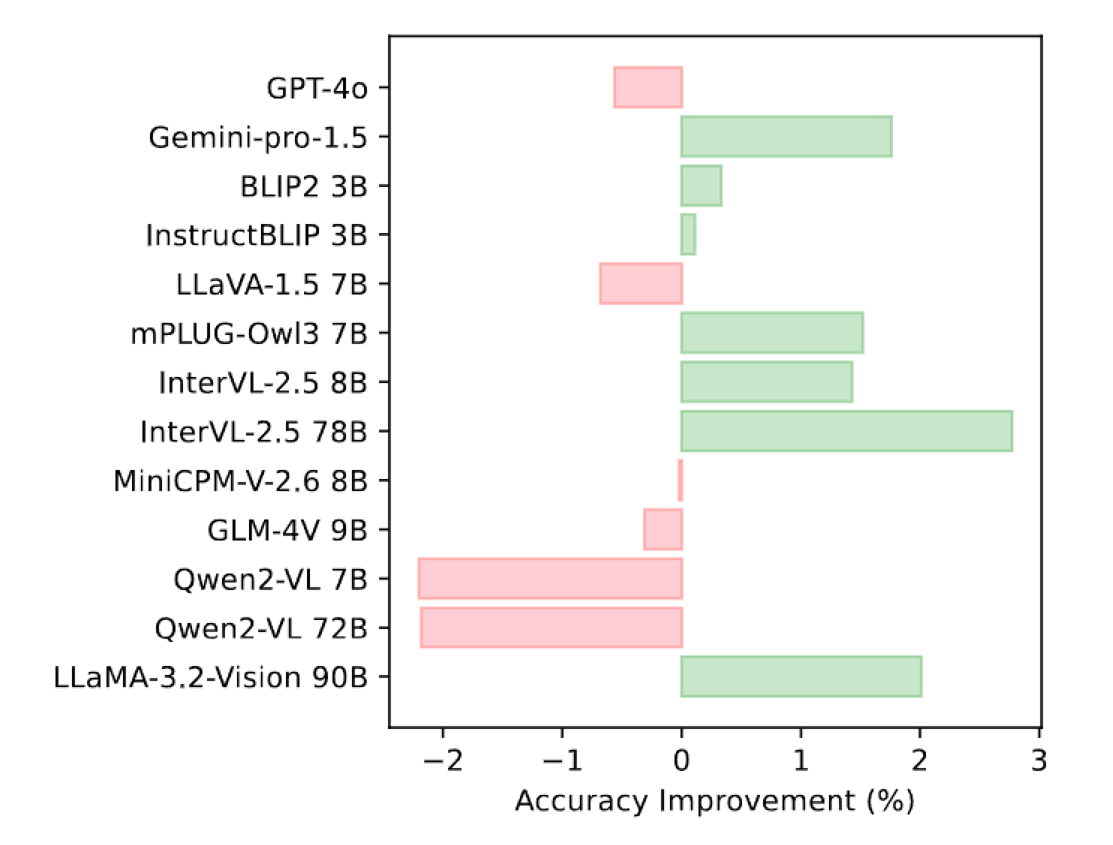
Encoder class	Resolution	Avg.	Easy						Hard					
			Ext.	Cnt.	Siz.	Loc.	Ref.	Rel.	Ext.	Cnt.	Siz.	Loc.	Ref.	Rel.
CLIP	224^2	40.8	35.4	44.2	40.9	49.2	**42.5**	50.5	31.8	26.8	42.3	46.5	36.1	43.4
CLIP	336^2	**43.1**	**40.0**	48.9	**50.5**	45.5	42.1	51.2	32.1	**31.0**	**52.1**	40.2	38.1	**45.1**
OpenCLIP	224^2	**43.1**	35.8	42.1	50.4	**53.6**	42.4	52.4	33.7	24.7	51.6	**49.4**	**39.3**	42.0
DINOv2	224^2	37.3	38.6	33.7	49.5	31.8	32.7	49.7	33.7	23.8	50.2	30.1	29.8	43.9
SigLIP	224^2	42.6	37.5	**49.4**	**50.5**	43.9	39.7	**53.7**	**35.6**	24.9	52.0	45.1	36.1	42.7
CLIP + DINOv2	$224^2 + 224^2$	38.2	35.9	43.6	40.1	38.3	39.2	47.9	34.1	24.5	38.1	37.3	35.3	44.0
CLIP + DINOv2	$336^2 + 224^2$	37.3	34.4	30.0	50.2	28.5	38.2	52.8	33.1	16.8	51.6	29.8	37.7	44.6

04 Experiments

Ablation study on noise

- Not all models experience a performance decline
- May be attributed to training data:
    i. Training data include scenarios with visual degradation
    ii. Noisy figures might align more closely with the distribution of the real-world images

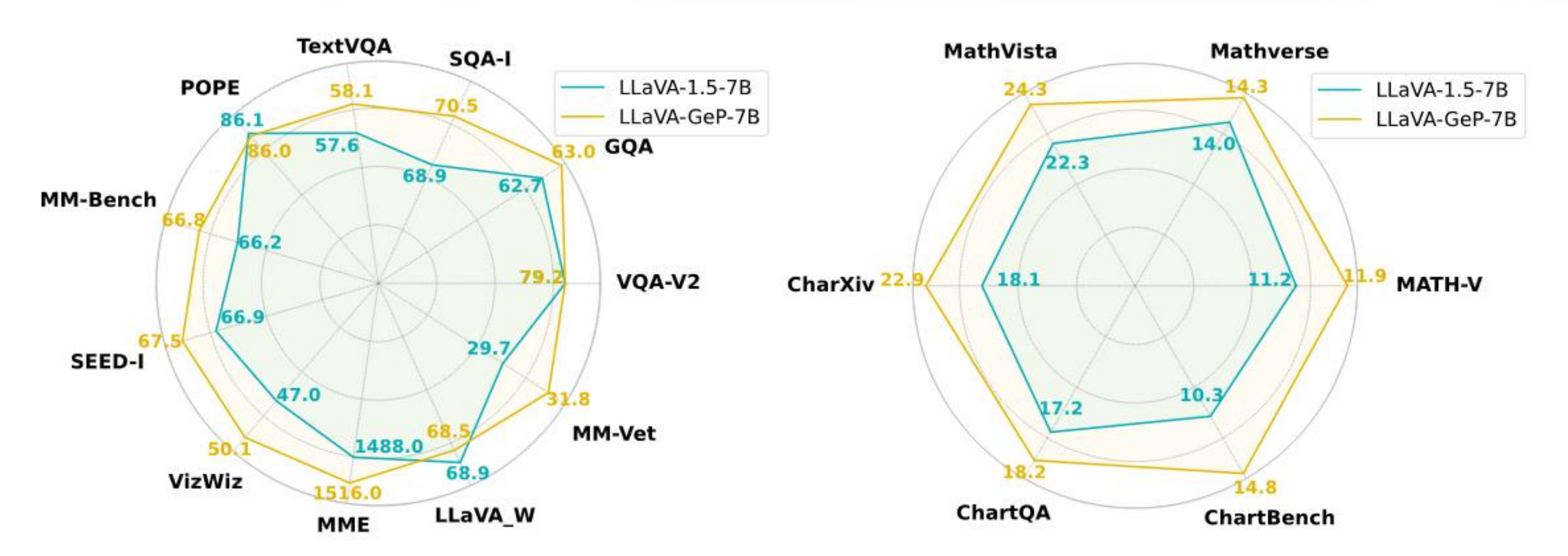




04 Experiments

Geometric perception is beneficial to downstream tasks

- Source 300K samples from the same distribution
- Mix with LLaVA-1.5 two stage training data, train LLaVA-1.5-GeP from scratch



05 Conclusion

- We introduce GePBench, a large-scale benchmark dataset designed to evaluate geometric perception in MLLMs
- Extensive experiments highlight substantial room for improvements
- Enhancing geometric perception contributes to improved performance in downstream tasks

Thanks