VisOnlyQA: Large Vision Language Models Still Struggle with Visual Perception of Geometric Information

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, Rui Zhang

{ryokamoi, rmz5227}@psu.edu Penn State University https://visonlyqa.github.io/

Motivation

The ability of LVLMs to perceive visual information in images has not been sufficiently studied. Specifically, it remains unclear **how accurately LVLMs can perceive geometric information, such as shape, angle, and size**, while geometric perception is fundamental to understanding visual information.

Limitations in existing datasets:

1. Popular datasets, such as MMMU and MathVista, target tasks that require expert-level reasoning and knowledge. They are not suitable for analyzing geometric perception.
2. Datasets for evaluating LVLMs at perceiving visual information include high-level tasks, such as scene understanding, which do not necessarily require accurate geometric perception.

Dataset Creation

We propose *VisOnlyQA*, a new dataset designed to evaluate how accurately LVLMs can perceive basic geometric information in images. Our dataset includes questions that **directly ask about basic and common geometric information (e.g., length, angle, and shape)**.

VisOnlyQA is designed to have favorable properties for analyzing the capability of LVLMs to perceive geometric information:

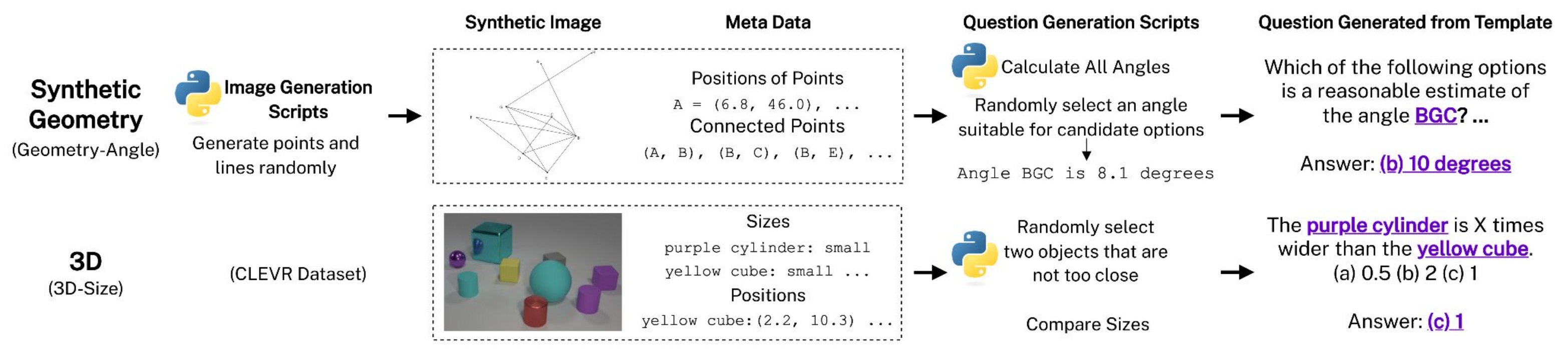
1. The questions in our dataset do not involve challenging reasoning or knowledge, exclusively evaluating geometric perception.
2. We use scientific figures to create unambiguous questions that require accurate geometric perception.

Real and Synthetic Figures

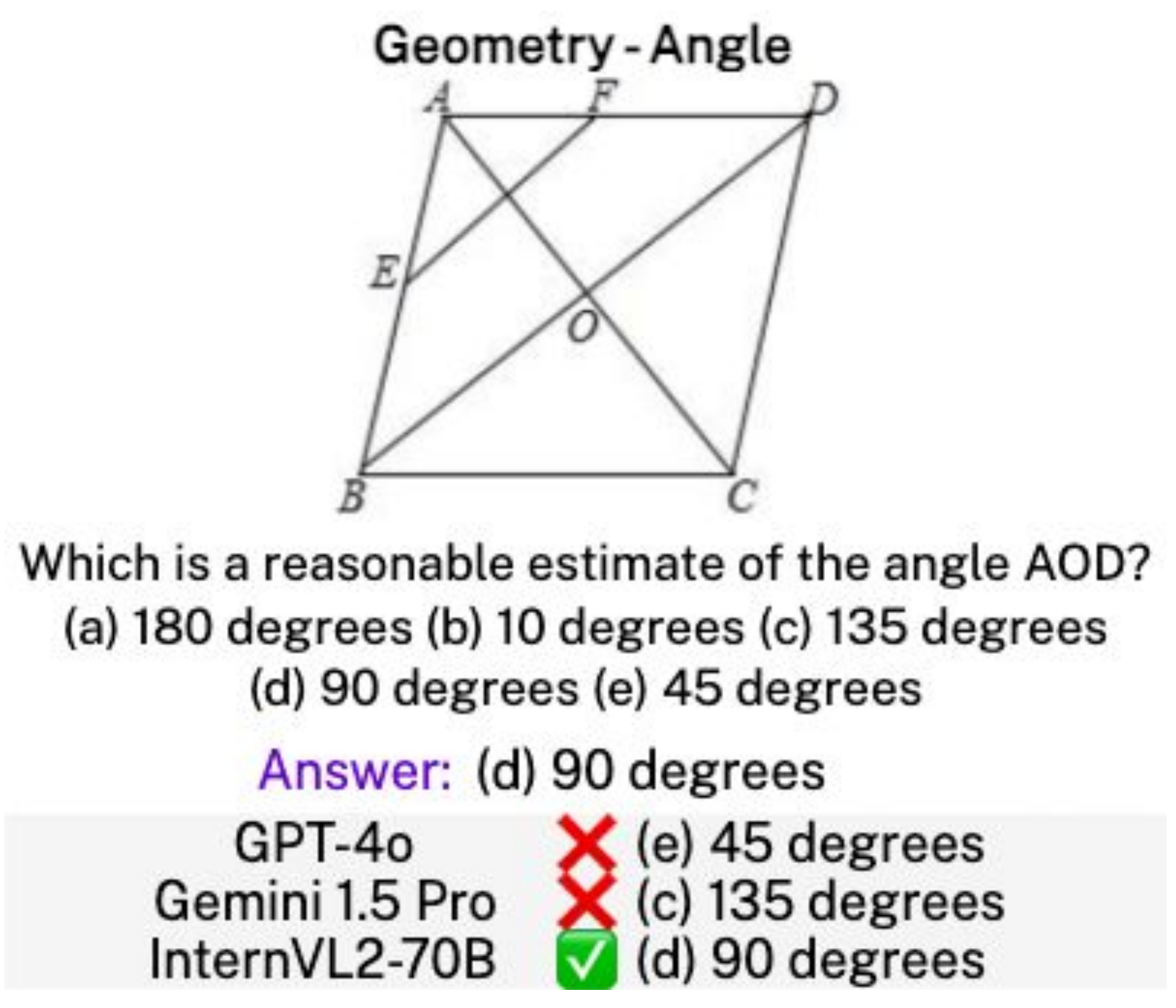
VisOnlyQA consists of two types of figures.

Real Figures: Images in popular datasets, such as MathVista and MMMU. We manually annotate new questions.

Synthetic Figures: We generate synthetic figures where we can get accurate geometric information like positions and size.


Synthetic Image
Meta Data
Question Generation Scripts
Question Generated from Template
Synthetic Geometry
(Geometry-Angle)
Image Generation Scripts
Generate points and lines randomly
Positions of Points
A = (6.8, 46.0), ...
Connected Points
(A, B), (B, C), (B, E), ...
Calculate All Angles
Randomly select an angle suitable for candidate options
Angle BGC is 8.1 degrees
Which of the following options is a reasonable estimate of the angle BGC? ...
Answer: (b) 10 degrees
3D
(3D-Size)
(CLEVR Dataset)
Sizes
purple cylinder: small
yellow cube: small ...
Positions
yellow cube:(2.2, 10.3) ...
Randomly select two objects that are not too close
Compare Sizes
The purple cylinder is X times wider than the yellow cube.
(a) 0.5 (b) 2 (c) 1
Answer: (c) 1

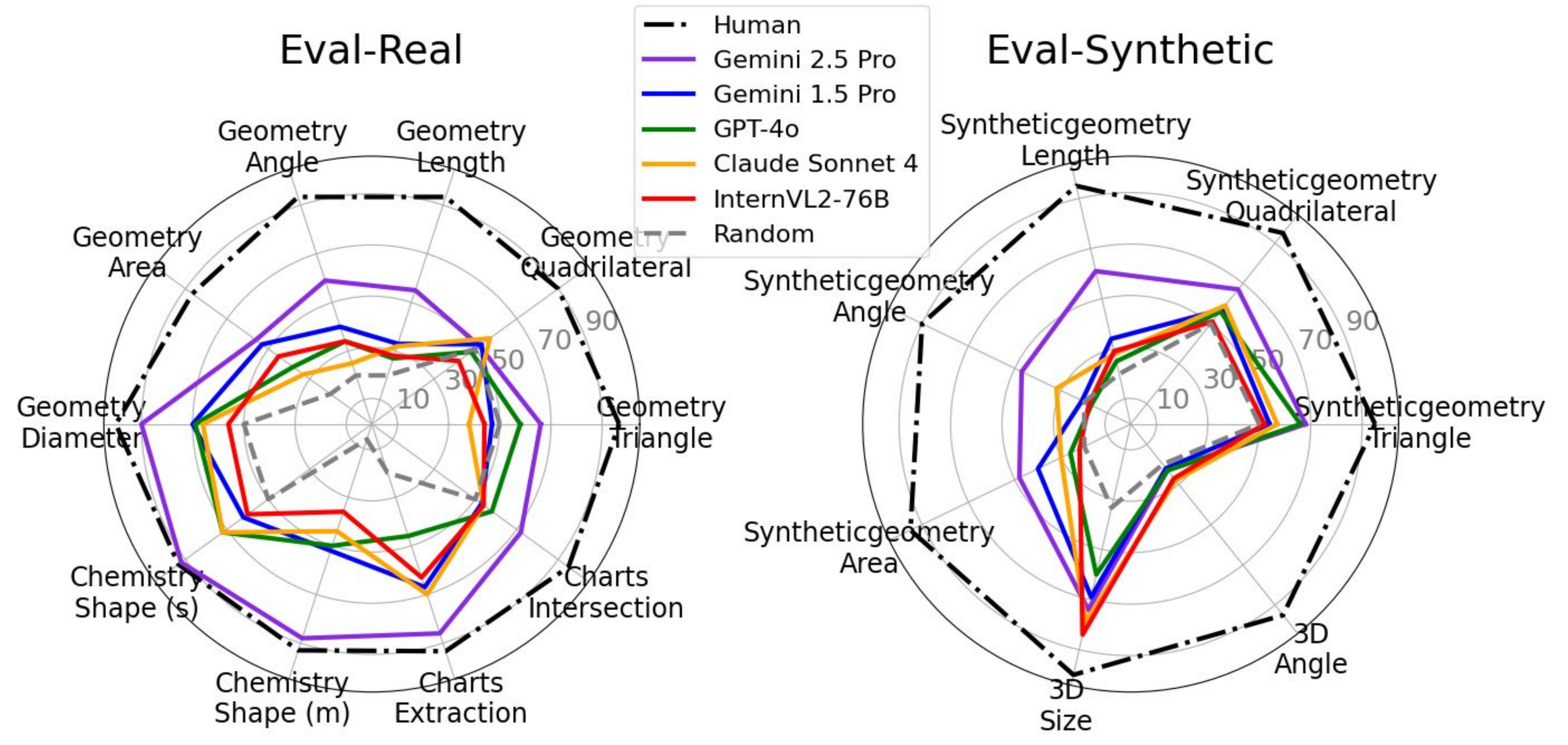
TL;DR - We introduce *VisOnlyQA*, a dataset for evaluating the geometric perception of LVLMs. We observe that even state-of-the-art LVLMs including GPT-4o and Gemini 2.5 Pro struggle with basic geometric perception questions like *"Does this figure include triangle ABC?"*



Which is a reasonable estimate of the angle AOD?
(a) 180 degrees (b) 10 degrees (c) 135 degrees
(d) 90 degrees (e) 45 degrees

Answer: (d) 90 degrees

GPT-4o	✗	(e) 45 degrees
Gemini 1.5 Pro	✗	(c) 135 degrees
InternVL2-70B	✓	(d) 90 degrees



VisOnlyQA consists of simple visual perception tasks, on which human performance is nearly perfect. However, state-of-the-art LVLMs such as GPT-4o, Claude Sonnet 4, and Gemini 2.5 Pro perform poorly.

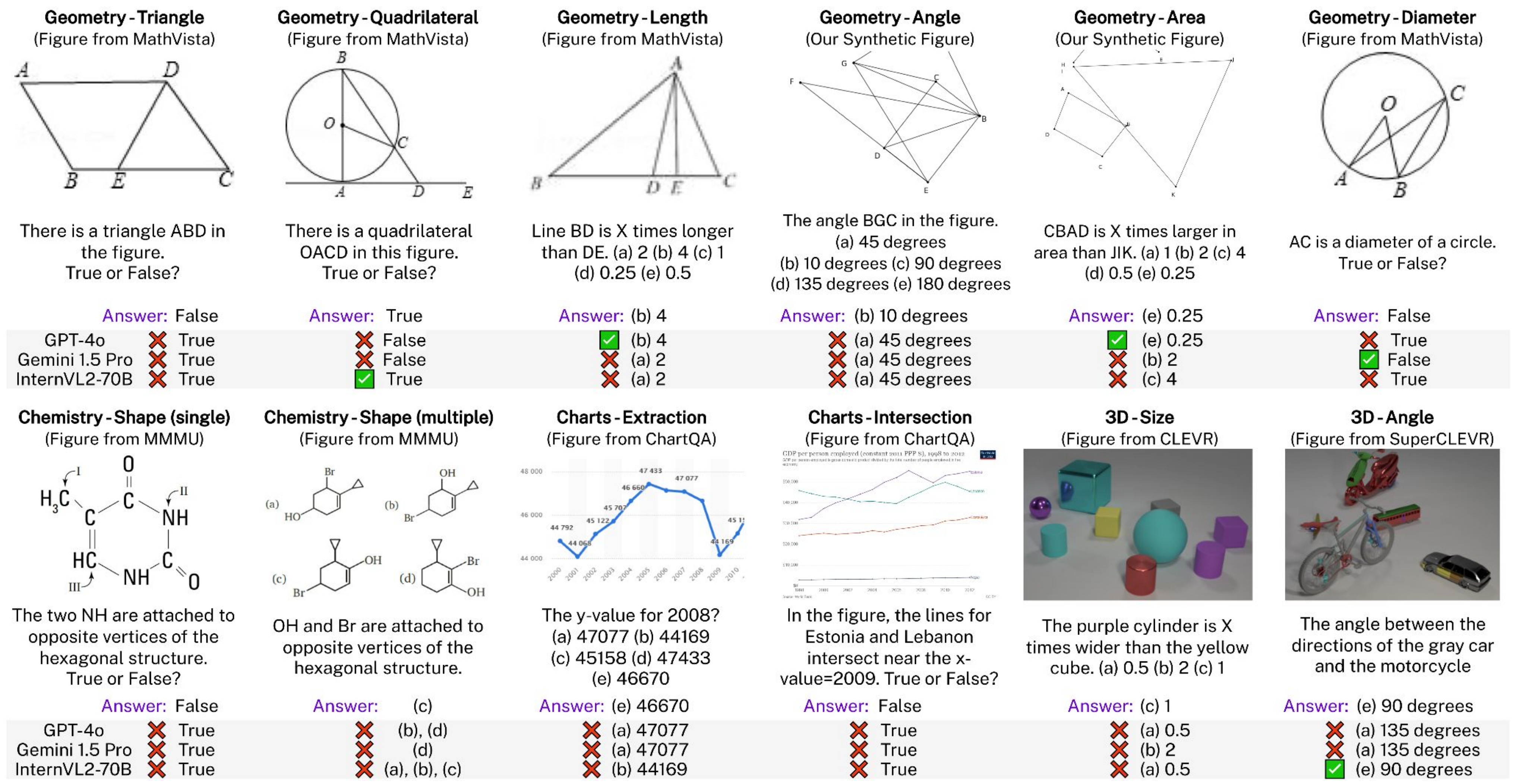
VisOnlyQA Dataset

VisOnlyQA includes **12 tasks** in total

Geometric Shapes	5 Tasks
Chemical Structure	2 Tasks
Charts	2 Tasks
3D shapes	2 Tasks

VisOnlyQA includes the evaluation and training set, consisting of the Real and Synthetic figures

	# Tasks	# Questions
Eval-Real	10 Tasks	900
Eval-Synthetic	7 Tasks	700
Train	7 tasks	70k



Geometry - Triangle (Figure from MathVista)
There is a triangle ABD in the figure. True or False?
Answer: False
GPT-4o ✗ True
Gemini 1.5 Pro ✗ True
InternVL2-70B ✗ True

Geometry - Quadrilateral (Figure from MathVista)
There is a quadrilateral OACD in this figure. True or False?
Answer: True
✗ False
✗ False
✓ True

Geometry - Length (Figure from MathVista)
Line BD is X times longer than DE. (a) 2 (b) 4 (c) 1 (d) 0.25 (e) 0.5
Answer: (b) 4
✓ (b) 4
✗ (a) 2
✗ (a) 2

Geometry - Angle (Our Synthetic Figure)
The angle BGC in the figure.
(a) 45 degrees
(b) 10 degrees (c) 90 degrees
(d) 135 degrees (e) 180 degrees
Answer: (b) 10 degrees
✗ (a) 45 degrees
✗ (a) 45 degrees
✗ (a) 45 degrees

Geometry - Area (Our Synthetic Figure)
CBAD is X times larger in area than JIK. (a) 1 (b) 2 (c) 4 (d) 0.5 (e) 0.25
Answer: (e) 0.25
✓ (e) 0.25
✗ (b) 2
✗ (c) 4

Geometry - Diameter (Figure from MathVista)
AC is a diameter of a circle. True or False?
Answer: False
✗ True
✓ False
✗ True

Chemistry - Shape (single) (Figure from MMMU)
The two NH are attached to opposite vertices of the hexagonal structure. True or False?
Answer: False
GPT-4o ✗ True
Gemini 1.5 Pro ✗ True
InternVL2-70B ✗ True

Chemistry - Shape (multiple) (Figure from MMMU)
OH and Br are attached to opposite vertices of the hexagonal structure.
Answer: (c)
✗ (b), (d)
✗ (d)
✗ (a), (b), (c)

Charts - Extraction (Figure from ChartQA)
The y-value for 2008?
(a) 47077 (b) 44169
(c) 45158 (d) 47433
(e) 46670
Answer: (e) 46670
✗ (a) 47077
✗ (a) 47077
✗ (b) 44169

Charts - Intersection (Figure from ChartQA)
In the figure, the lines for Estonia and Lebanon intersect near the x-value=2009. True or False?
Answer: False
✗ True
✗ True
✗ True

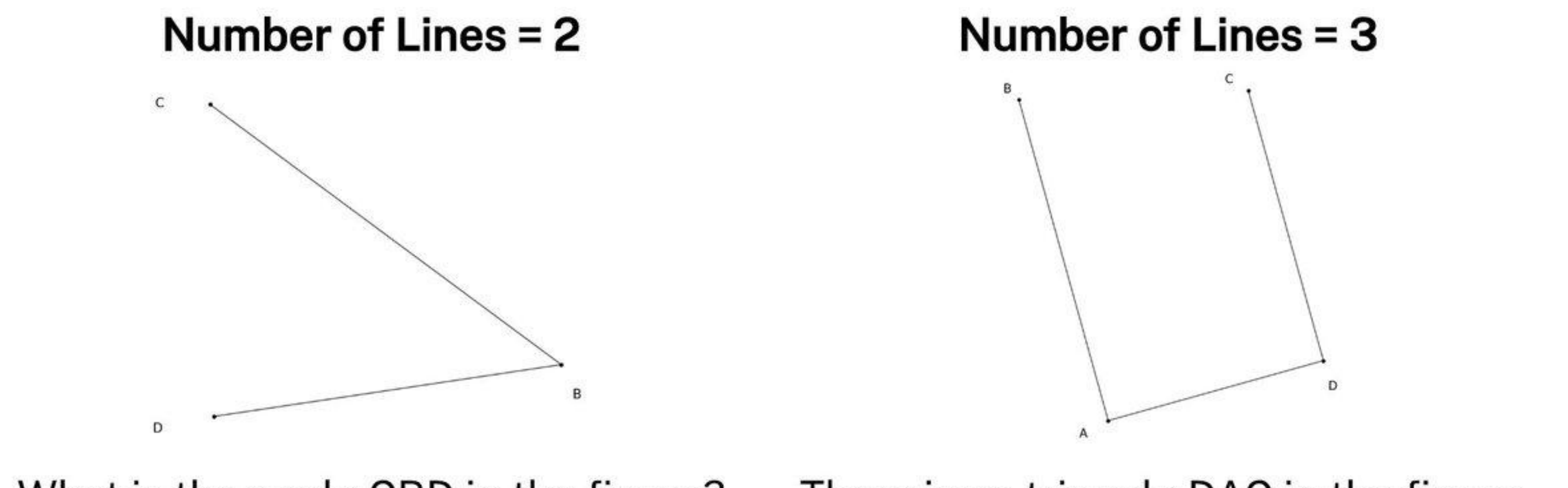
3D - Size (Figure from CLEVR)
The purple cylinder is X times wider than the yellow cube. (a) 0.5 (b) 2 (c) 1
Answer: (c) 1
✗ (a) 0.5
✗ (b) 2
✗ (a) 0.5

3D - Angle (Figure from SuperCLEVR)
The angle between the directions of the gray car and the motorcycle
Answer: (e) 90 degrees
✗ (a) 135 degrees
✗ (a) 135 degrees
✓ (e) 90 degrees

Analysis

Geometric Shapes with Only Two Lines are Already Difficult

Surprisingly, **LVLMS perform poorly even on very simple geometric shapes that only include two or three lines**.



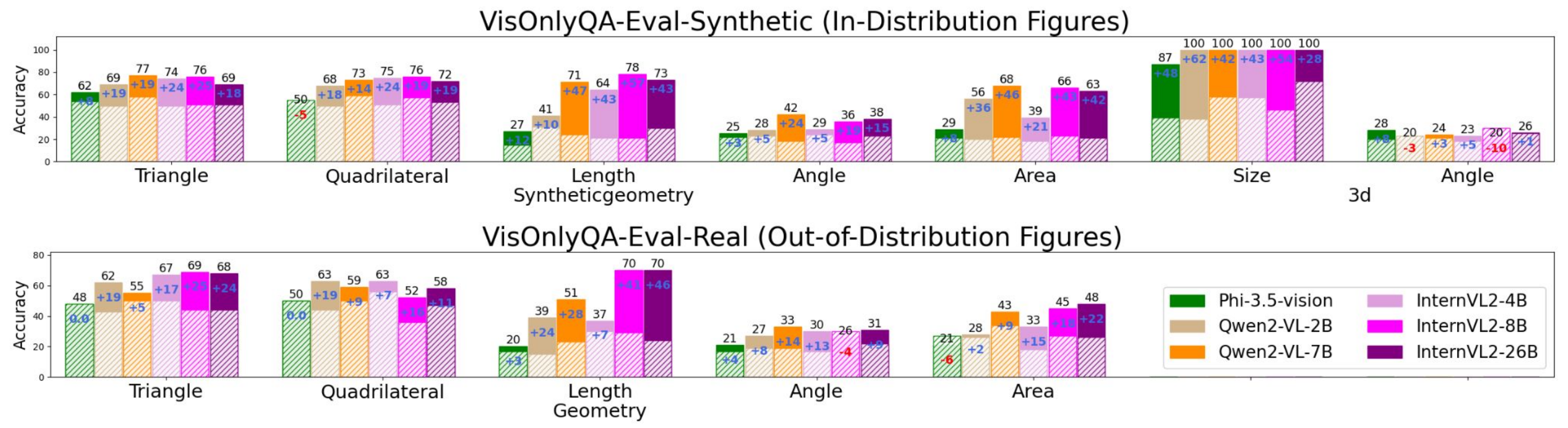
What is the angle CBD in the figure?
Answer: (b) 45 degrees

There is no triangle DAC in the figure.
Answer: True

Gemini 1.5 Pro	✗	(d) 135 degrees	✗	False
InternVL2-70B	✗	(d) 135 degrees	✗	False

Fine-tuning Does Not Fully Solve This Issue

Fine-tuning on VisOnlyQA-Train does not always improve geometric perception even on in-distribution images. This result suggests that simply scaling training data does not solve this issue



Larger LMs Improve the Geometric Perception of LVLMs

LVLMs with **larger language models exhibit better visual perception**. This result suggests that language models are crucial in processing visual information encoded by ViT.

	ViT Size	LLM Size	Real	Synthetic
InternVL2-4B	304M	3.8B	38.4	34.1
InternVL2-8B	304M	7.7B	**40.7**	**35.0**
Qwen2-VL-2B	675M	1.5B	32.3	33.6
Qwen2-VL-7B	675M	7.6B	**38.9**	**37.1**