# Tri-Bench: Stress-Testing VLM Reliability on Spatial Reasoning under Camera Tilt and Object Interference

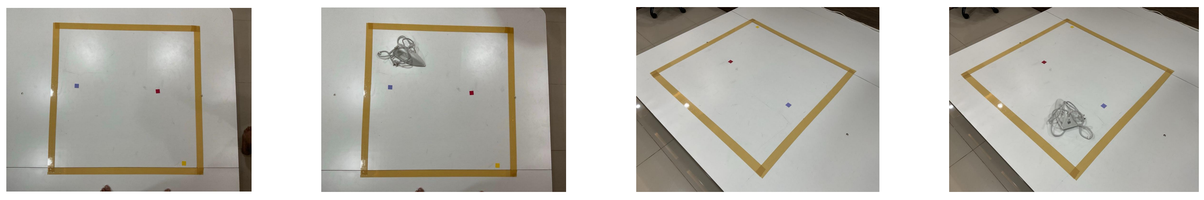
Amit Bendkhale | Independent Researcher, India

## Motivation

- VLMs are widely used, but 3D spatial reasoning is unclear.
- Do VLMs reason in 3D or rely on 2D image cues?
- What natural factors can affect VLM spatial reasoning?
- We propose a compact benchmark for fast stress-testing.

## New Dataset

- 100 triangles, with 4 views → 400 images



- Diversity in triangle shapes along with their 2D projections:

| 3D → 2D ↓ | S | I | E |
|---|---|---|---|
| S | 237 | 65 | 20 |
| I | 17 | 39 | 5 |
| E | 2 | 0 | 15 |

scalene (S) = 64%
isosceles (I) = 26%
equilateral (E) = 10%

| 3D → 2D ↓ | A | O | R |
|---|---|---|---|
| A | 113 | 8 | 35 |
| O | 34 | 116 | 49 |
| R | 5 | 4 | 36 |

acute (A) = 38%
obtuse (O) = 32%
right (R) = 30%

- 10 everyday objects for interference
- Square border provided as a guardrail for homography

## VLM Tasks

- Single prompt for 6 relative measurement prediction tasks:

Q1. Triangle Side Type
Q2. Triangle Angle Type
Q3. Ratio of AB and AC
Q4. Difference: $|\angle ABC - \angle ACB|$
Q5. Ratio of max/min sides
Q6. Maximum angle difference

## Result 1: accuracy w.r.t. 3D vs 2D

| Model | w.r.t. 3D Ground Truth | w.r.t. 2D Image Projection |
|---|---|---|
| Gemini-2.5-Pro | 75.30 | 80.89 |
| Gemini-2.5-Flash | 71.58 | 77.14 |
| GPT-5 | 64.32 | 65.04 |
| Qwen2.5-VL-32B | 64.70 | 66.22 |
| **AVERAGE** | **68.98** | **72.32** |

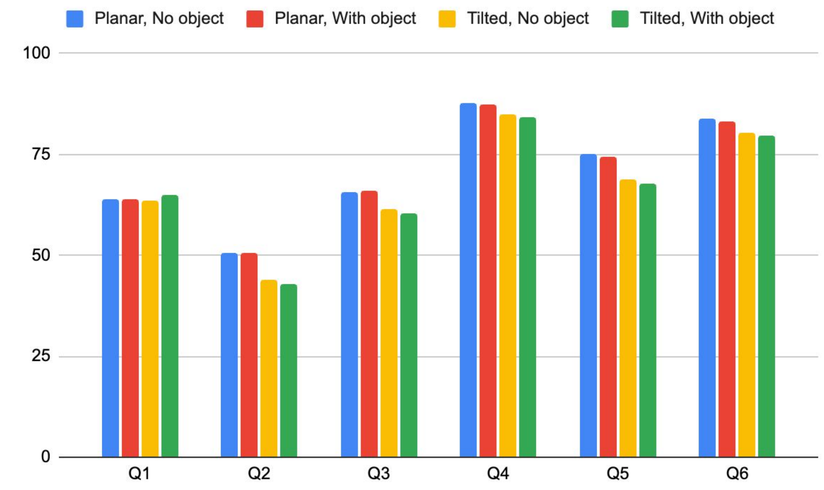
*Average accuracy (%) of VLMs across all tasks*

## Result 2: majority class bias

| Model | S (Q1) | I (Q1) | E (Q1) | A (Q2) | O (Q2) | R (Q2) |
|---|---|---|---|---|---|---|
| Gemini-2.5-Pro | 99.61 | 2.88 | **0.00** | 78.29 | 88.28 | **0.00** |
| Gemini-2.5-Flash | 98.83 | 1.92 | **0.00** | 72.37 | 80.47 | 5.83 |
| GPT-5 | 99.61 | 0.96 | **0.00** | 92.11 | 3.91 | 1.67 |
| Qwen2.5-VL-32B | **100.00** | **0.00** | **0.00** | **100.00** | **0.00** | **0.00** |
| **AVERAGE** | **99.51** | **1.44** | **0.00** | 85.69 | 43.16 | **1.88** |

*Average accuracy (%) of VLMs across Q1 and Q2, as per 3D triangle shapes*

## Result 3: question-wise analysis




*Average accuracy (%) across all VLMs for each task Q1-Q6, under four capture conditions*

## Key Takeaways

1. VLMs are accurate ~70% of the time
2. VLM answers are closer to image-plane relations instead of actual 3D
3. VLMs show very high majority class bias under precision tasks Q1-Q2
4. Non-planar camera tilt shows consistent accuracy decline of ~4%
5. Object interference impact is negligible

## Future Work

- Multi-view geometry
- Lighting variations
- Granular Camera Tilt
- Sequential Prompting
- Reasons for inaccuracy
- Complex shapes
- Relative vs Absolute
- Better error metrics
- Heavy object occlusion
- Analyzing Training Bias

connect

arXiv

github