

26 FEB 2024

Spatial Reasoning in large language models

MOTIVATION

Humans extract and use
knowledge of the physical world
- referred to as *world models*.

Are language models, who've seen a vast majority of all text ever produced by humans, capable of inferring spatial relations from their sequential, text-based inputs?

EXAMPLE 1

“

I have a stereo-vision system with a focal length of 16mm and a baseline of 30cm.

One of the cameras is red.

I rotate the left camera 87 degrees towards left.

Can you calculate the new baseline?

ANSWER FROM GPT4

To calculate the new baseline after rotating the left camera by 87° to the left, let's analyze the situation.

Given:

- **Original baseline** (B) = 30 cm
- **Rotation angle** (θ) = 87°
- **Focal length** (f) = 16 mm (not needed for baseline calculation)
- **One camera is rotated left** (assumed counterclockwise in the horizontal plane)

Step-by-Step Calculation:

When the left camera is rotated by θ , the new effective baseline (B') is the perpendicular distance between the optical centers of both cameras. The effective baseline is given by:

$$B' = B \cos(\theta)$$

Substituting the values:








$$B' = 30 \times \cos(87^\circ)$$

Using $\cos(87^\circ) \approx 0.0506$,

$$B' = 30 \times 0.0506 = 1.518 \text{ cm}$$

Conclusion:

After rotating the left camera by 87° to the left, the new effective baseline reduces to **approximately 1.52 cm**. This means the stereo system will now have significantly less disparity and a much weaker depth perception compared to the original configuration.



✗ Doesn't get that the question is part of a “special case”

ANSWER FROM GPT4

did you take into account that one camera is red?

The color of the camera (red) does not affect the geometric calculations of the baseline unless there's some specific optical property related to the color that alters the system (e.g., a wavelength-dependent focal shift). However, in standard stereo vision baseline calculations, the color is purely cosmetic and does not influence the math.

So, yes, I accounted for all the relevant factors—rotation, baseline, and trigonometry—but if you meant something specific about the red camera affecting the baseline, let me know! 😊



Gets that the camera being **red** has nothing to do with the baseline

ANSWER FROM GPT4

what happens if the rotation is decreased to 45 degrees. What will be the resulting baseline?

If the rotation angle of the left camera is decreased to **45°**, we recalculate the new baseline using the same formula:

$$B' = B \cos(\theta)$$

where:

- $B = 30$ cm (original baseline)
- $\theta = 45^\circ$

Using $\cos(45^\circ) \approx 0.7071$:

$$B' = 30 \times 0.7071$$

$$B' \approx 21.21 \text{ cm}$$

Conclusion:

With a **45° rotation**, the new baseline is approximately **21.21 cm**. This is a significant increase compared to the **1.52 cm** at **87°**, meaning the stereo vision system will have better depth perception than before but still less than the original 30 cm baseline.



Doesn't question the formula used

EXAMPLE 2

“

I am collecting images to form a panoramic image.

My camera captures approximately 60 degrees.

After taking the first image, I rotate the camera 180 degrees and take a second image.

How many degrees does my panorama span after stitching?

ANSWER FROM GPT4 - PART 1

Okay, let's break this down. The first image covers 60 degrees, and after rotating the camera by 180 degrees, a second image is taken. That's pretty straightforward — the two images will be stitched together to form a panoramic image.

So, the panorama span should be the sum of these two angles, right? If the first image covers 60 degrees and the camera rotates 180 degrees, the total panoramic span should be $180 + 60$ degrees, making it 240 degrees.

So, the camera rotates 180 degrees between images, and each image has a 60-degree field of view. The question is a bit tricky — it's not just about adding the two FOVs. The images are actually 180 degrees apart in orientation, and each image covers a range of angles.

For instance, if the first image is centered at 0 degrees, it spans from -30 degrees to +30 degrees. The second image is centered at 180 degrees, spanning from 150 degrees to 210 degrees.

So, the total panorama span might be 120 degrees, but with a gap in between.

Okay, there are two 60-degree FOV images, each 180 degrees apart. The first covers from -30 to +30 (or equivalently 330 to 30 degrees), and the second image covers from 150 to 210 degrees. These segments are not adjacent, so stitching them forms two disjoint sections on a circular panorama.

✗ Starts off with the wrong reasoning

✓ Gets that the questions is “tricky”

✓ Correct spatial model of the problem!

✓ Still on the correct reasoning path

ANSWER FROM GPT4 - PART II

If the question is asking for the total angular span, it seems to be the angular distance between the minimum and maximum angles. This would be 180 degrees plus 60 degrees, so the total span is 240 degrees, despite the gap in between them.

Each image covers 60 degrees, with the camera rotation placing them 180 degrees apart. The entire panorama isn't continuous because of the 120-degree gap between the two images. The question of "span" becomes tricky since the images aren't contiguous.

If we consider the span as the total angular distance from the leftmost to the rightmost part of the panorama (i.e., the arc covering both images), then the span will be 240 degrees (after filling the gap).

Okay, let's clarify: the first image's left edge is at -30 degrees (or 330 degrees after wrapping), and the second image's right edge is at 210 degrees. Measuring between those extreme points, one way gives a span of 120 degrees, but another gives 240 degrees.

The key issue is the 180-degree separation between the images. The actual "panorama span" is ambiguous, but if we assume we're joining them side by side on a circle, the total span could be 240 degrees, meaning the angular extent from the leftmost to rightmost edge after stitching.

- ✓ Still on the correct reasoning path
- ✗ Gets confused with what's asked
- ✗ Leads to the wrong conclusion & answer