

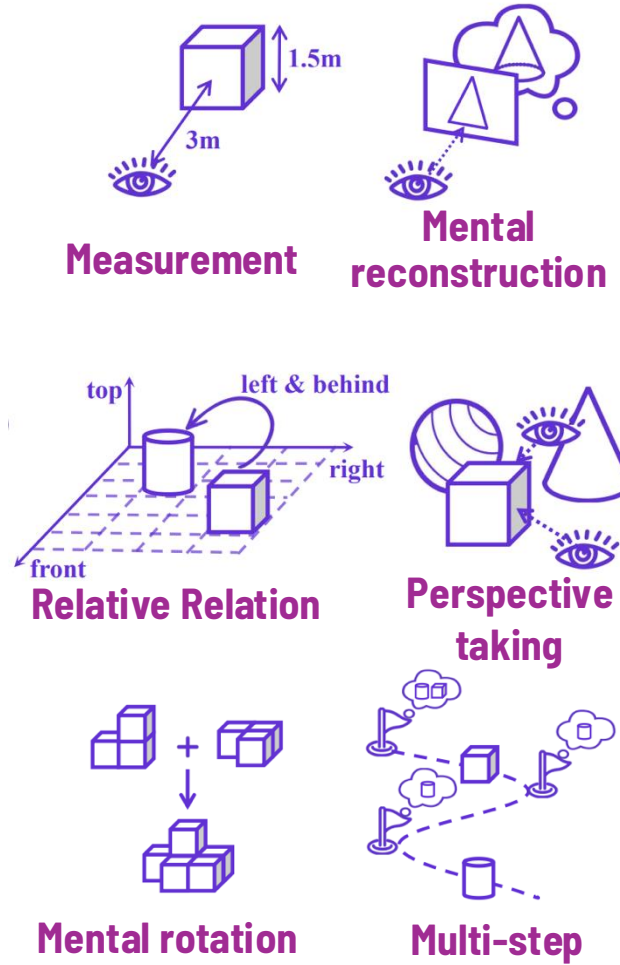
# **Scaling Spatial Reasoning for VLMs at Inference Time**

**Haoming Wang, Qiyao Xue and Wei Gao**  
**University of Pittsburgh**

# Fundamental Capabilities of Spatial Intelligence



**Egocentric  
Visual inputs**



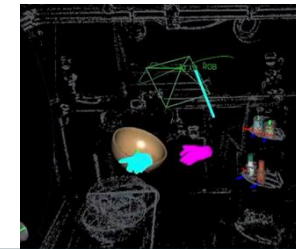
**Basic Spatial  
Reasoning Capabilities**



**Robotics Manipulator**



**Autonomous Drones**



**Wearable AR**

**Spatial-aware  
Applications**

# the performance gap between VLMs and humans



## GPT-5



### Spatial Relations (SR)

**Prompts:** Which object is higher in the 3D world space, the clock or the house in the back?

**GT:** The house in the back.

**Model Responses:** The question is asking which object is higher in physical elevation, the clock tower or the house behind it, not just perceived height. The clock tower seems taller and extends above the house roof, even factoring in the potential slope of the ground. So, the clock tower is likely the higher object in 3D world space.

**Answer:** The clock.

**Reasoning based on 2d pixel space**

### Mental Reconstruction (MR)

**Prompts:** Draw a 3D shape whose top view looks like this, and composed of 3D cubes.

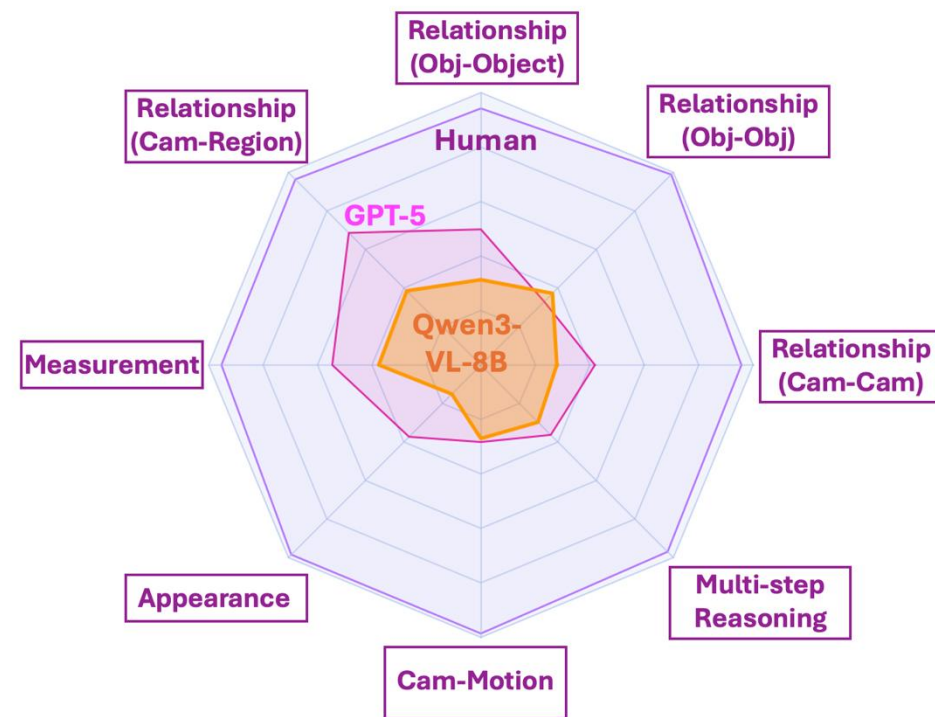
**GT:** -

**Model Responses:** The user requested a 3D object with a top view resembling an L-tetromino, made out of cubes. It's a 4-square footprint, with varied possible heights. The simplest format is an isometric view with a single cube stacked. I'll create the requested 3D shape and generate the image without needing extra edits or references.

**Answer:**



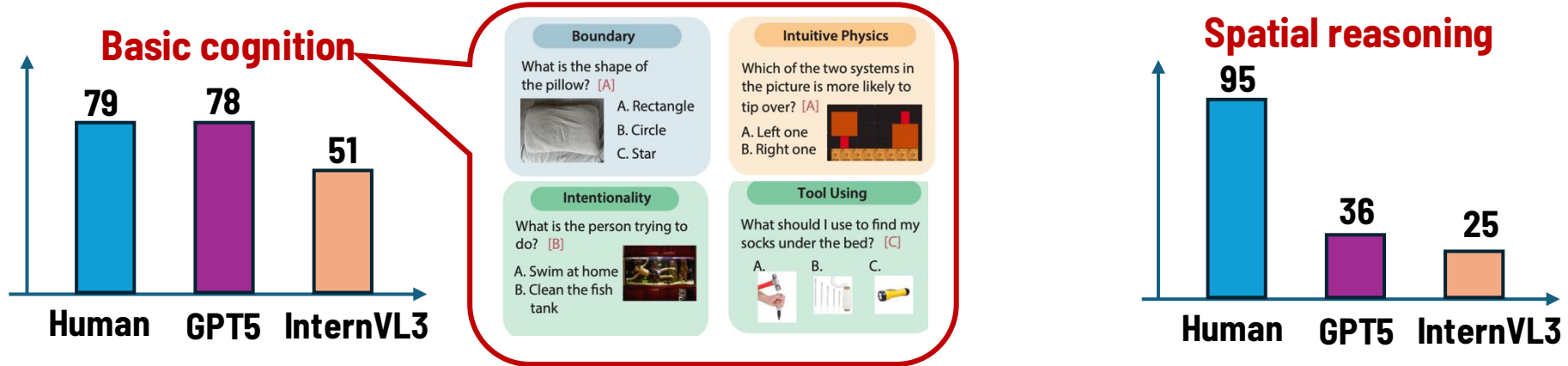
**Struggle in mental rotation**



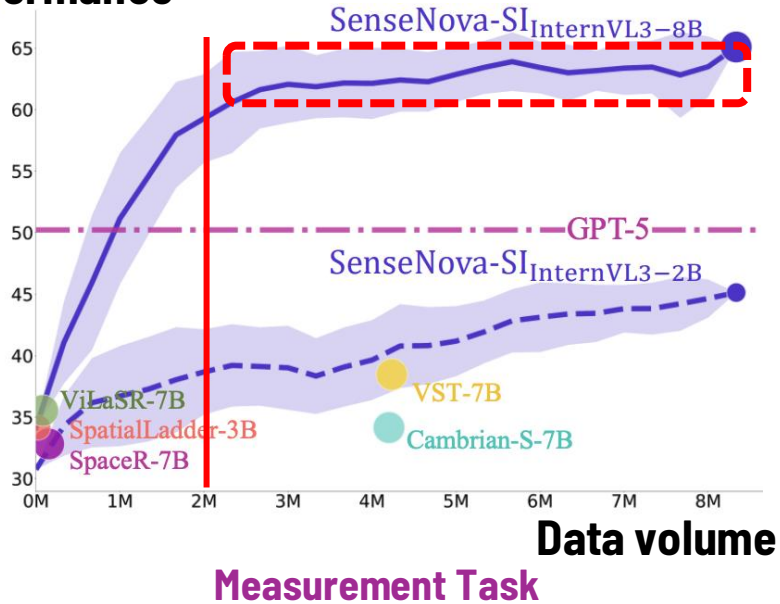
**Human VS Proprietary models  
VS Open-sourced general models**

# Can we solve spatial reasoning by further scaling training data

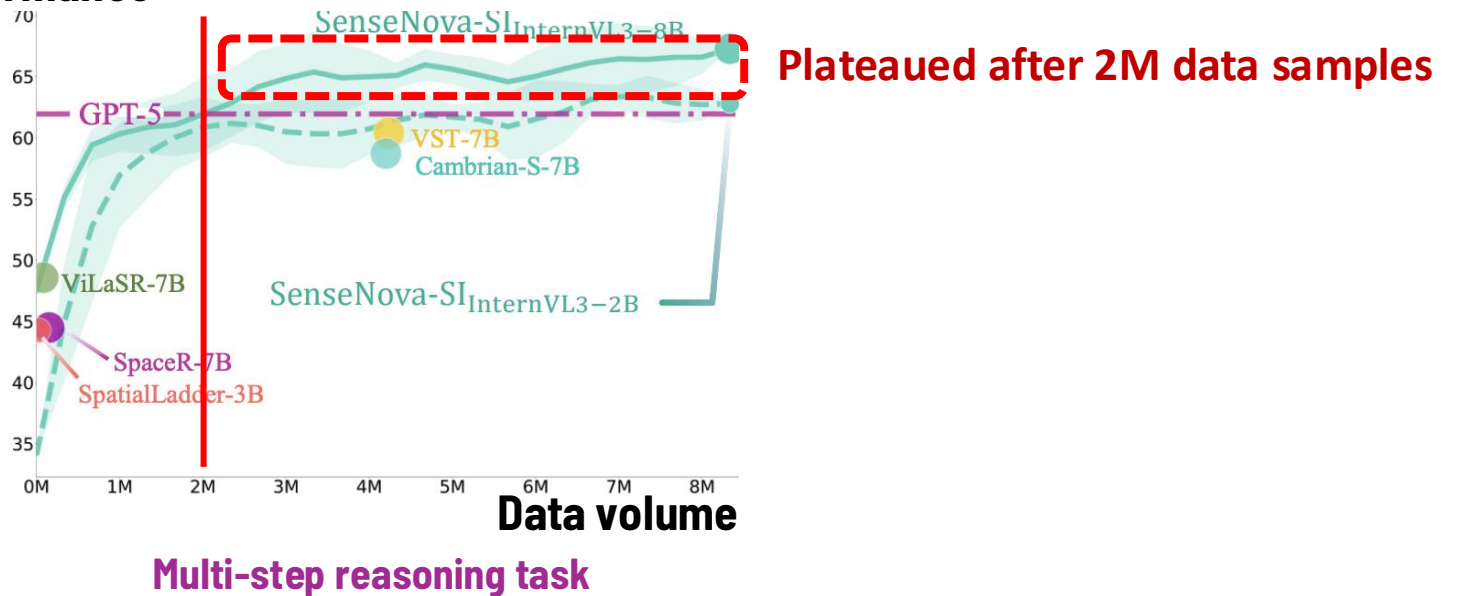
Spatial reasoning is not "emerging" at the same rate as other language or vision capabilities



## Performance

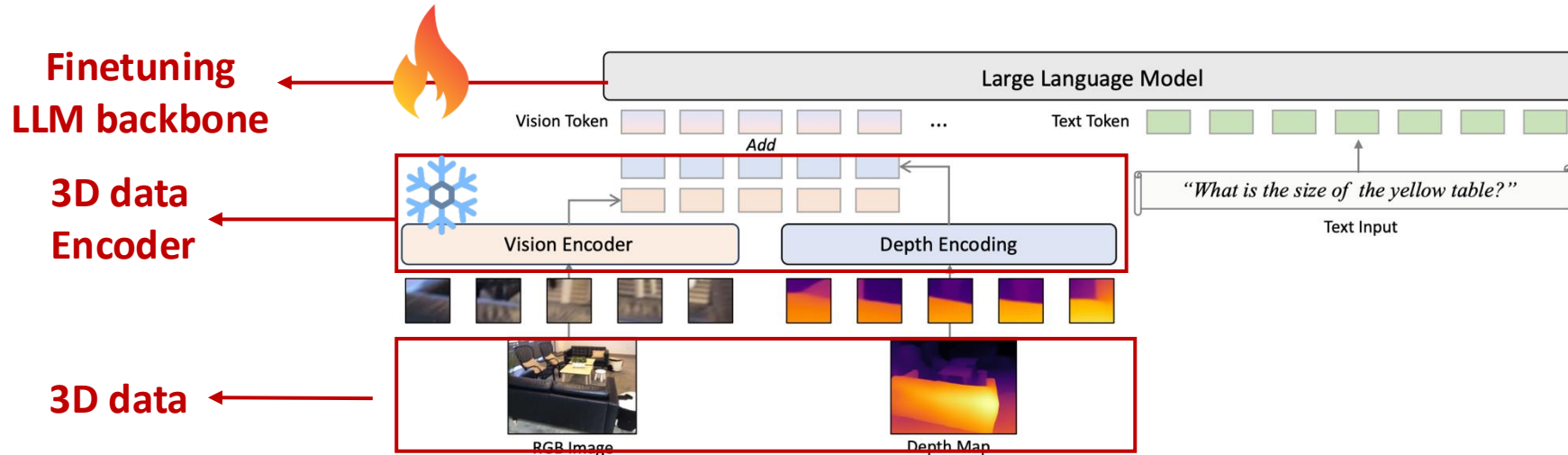


## Performance

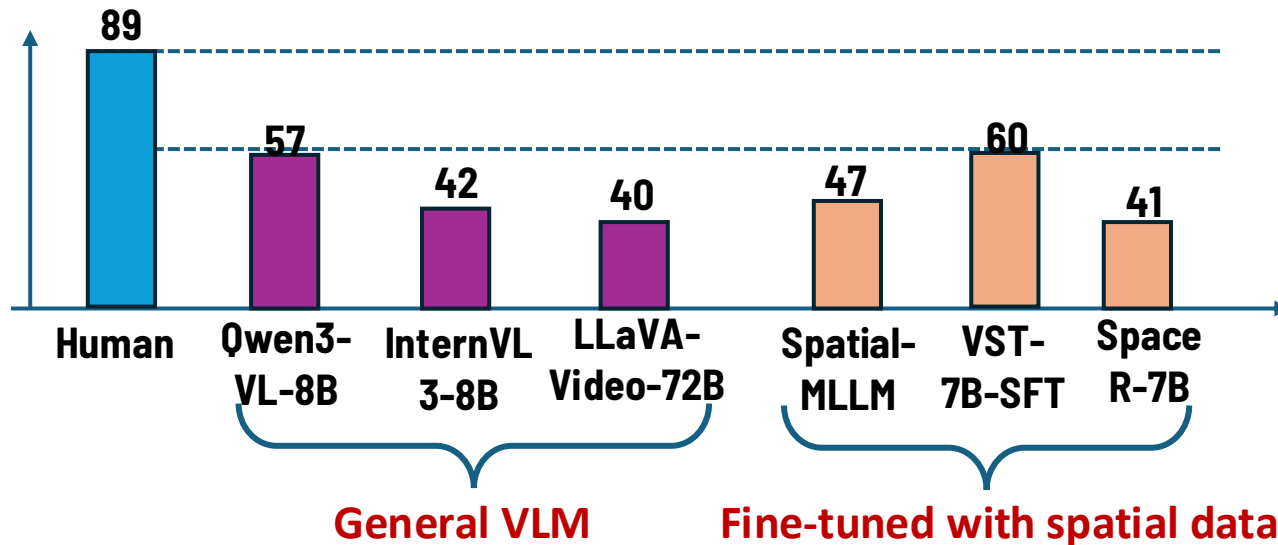


# Injecting 3D data to input layer – limited improvement

Typical pipeline of adding new modality to VLM input layer

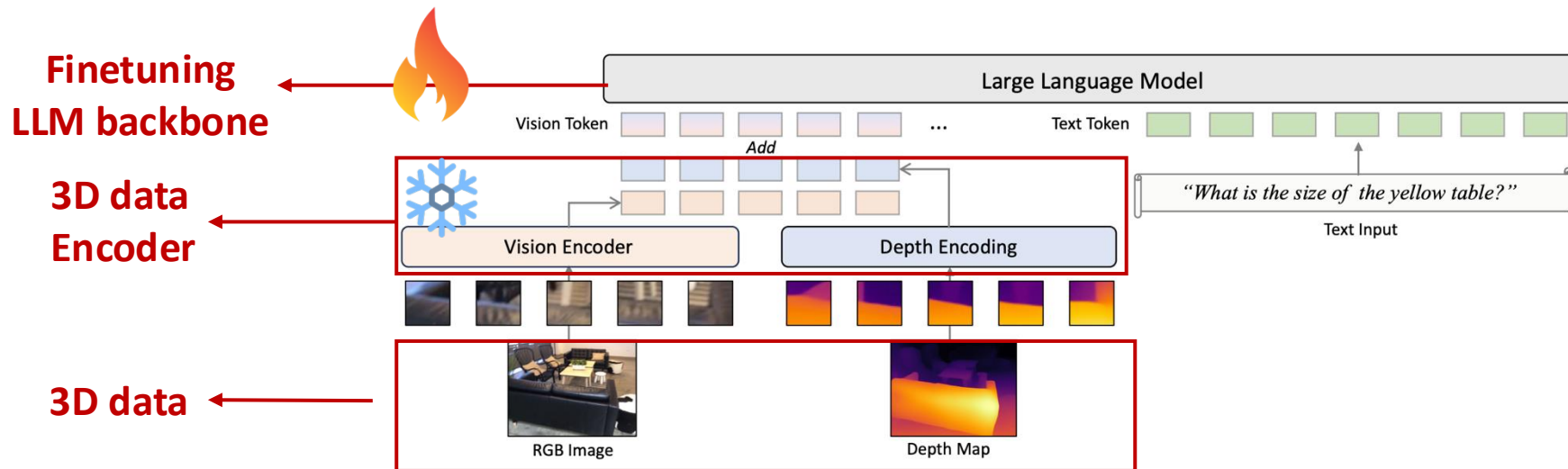


Limited improvement:



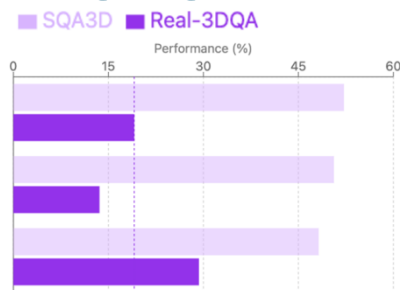
# Injecting 3D data to input layer – limited improvement

## Typical pipeline of adding new modality to VLM input layer



## The model didn't actually reason with 3d data

### - Language shortcut



performance plummeted after linguistic cues were removed

### - Raw 3d data is too information-dense

#### Heuristic Shortcuts

Q. Is the chair facing the window?

#### Non-minimal Set:

yaw(chair)=+90°  
window position  
office desk is in front of chair  
notebook and mug on desk

#### Reasoning:

Desk is in front of chair, as chair typically faces desk...No. ❌

#### Attention Dilution

Q. Is the chair between the desk and the window?

#### Non-minimal Set:

position (chair, desk, window)  
chair near to the door  
desk on the right  
window on the left wall

#### Reasoning:

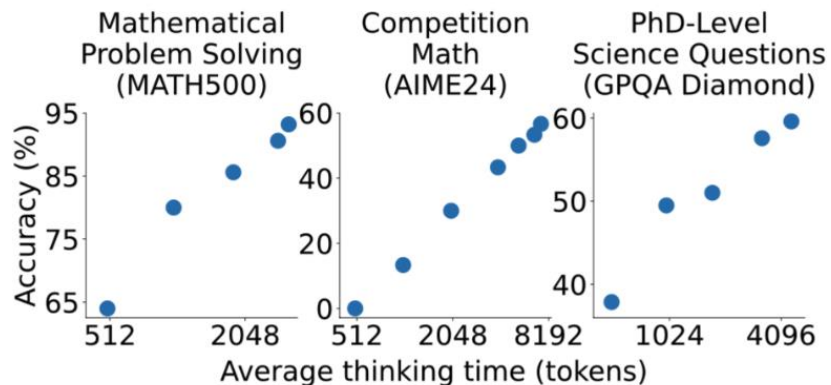
Desk is on the right, chair near to the door. There is no direct line from this...No. ❌



flooding the context with distracting geometric details and inducing hallucinations.

# How to scale at inference time

## Inference time scaling in text domain



Performance scales with thinking time

```

To find the real numbers  $a \neq 0$  and  $b$  such that
 $f(x) - f(a) = (x - a)^2 (x - b)$ ,
we first need to compute  $f(a)$ :
 $f(a) = a^3 + 3a^2 + 1$ .

.....

Let's check if  $b = 0$  is a valid solution:
Substituting  $b = 0$  into  $b + 2a = -3$ , we get:
 $0 + 2a = -3 \implies 2a = -3 \implies a = -\frac{3}{2}$ .
Now, we can verify that this value of  $a$  also satisfies
the equation  $2ab + a^2 = 0$ :
 $2 \cdot (-\frac{3}{2}) \cdot 0 + (-\frac{3}{2})^2 = 0 + \frac{9}{4} = \frac{9}{4} \neq 0$ .

However, this does not equal  $0$ . So,  $b = 0$  is not a valid solution.
Let's try another possibility:

.....

Therefore, the ordered pair  $(a, b)$  is  $(-2, 1)$ .
    
```

Verification

Backtracking

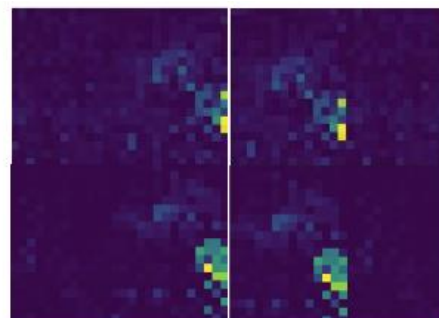
Ahah moment in reasoning

## Thinking in text domain impair visual spatial reasoning

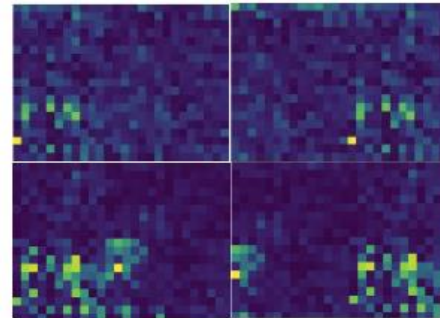
Model	Inference Method	Occlusion Restoration	Distance Comparison	Azimuth Transfer
Qwen2.5-VL	Vanilla	34.52	<b>43.41</b>	27.25
	Explicit Stepwise CoT	33.0 (-1.51)	34.4 (-9.00)	23.0 (-4.25)
	Explicit Multi-view CoT	35.04 (+0.52)	34.16 (-9.25)	24.29 (-2.96)



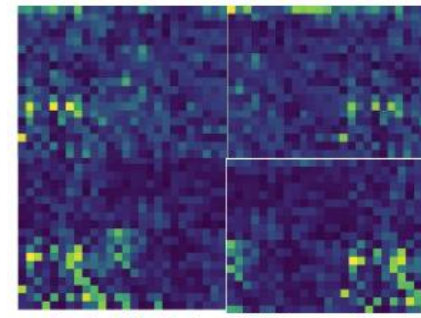
Q: Which object is closest to the bike in both images?  
 A. Sofa  
 B. blue mat  
 C. indoor\_plant  
 D. Other



vanilla (Answer: A ✓)



Explicit Stepwise CoT (Answer: C ✗)



Explicit Multi-view CoT (Answer: B ✗)

# How to scale at inference time

Alternative method other than thinking in text

Think with image

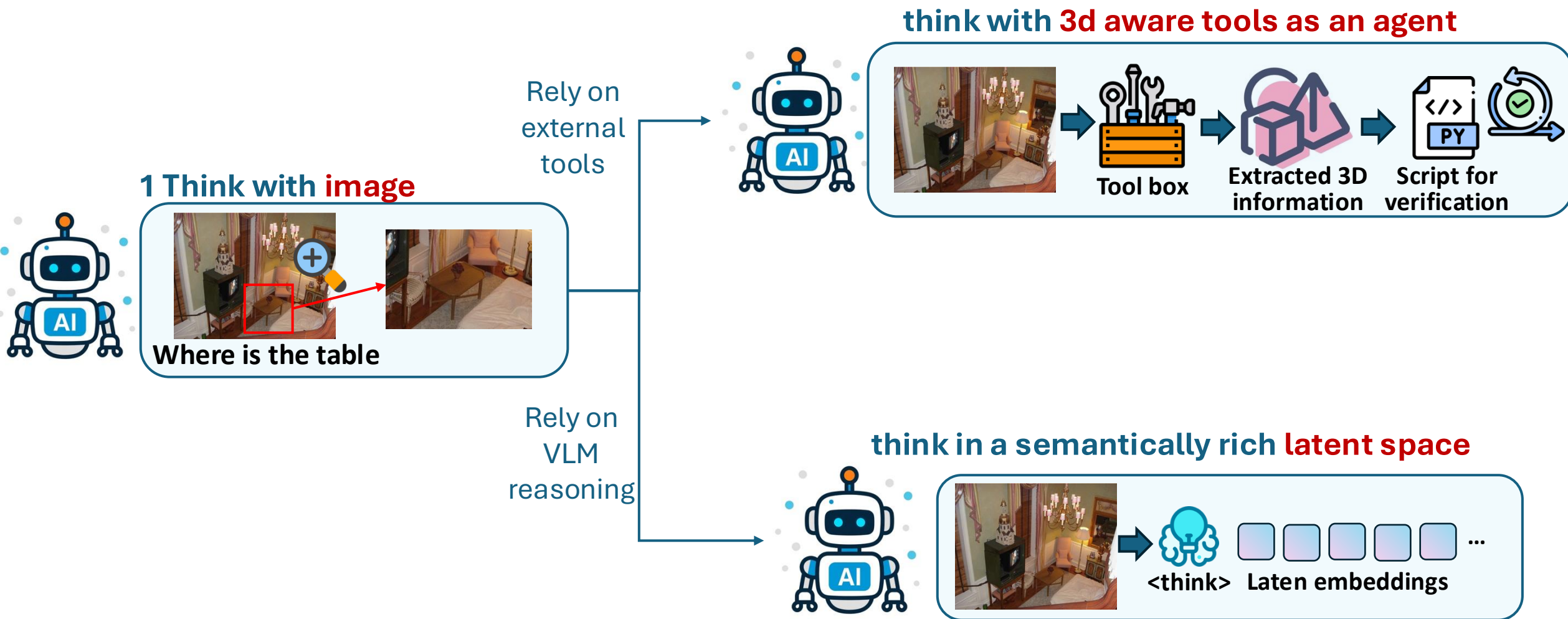


**Limitation:**

- only guide the model to focus on task-relevant regions
- but does not enable it to extract spatial information

# How to scale at inference time

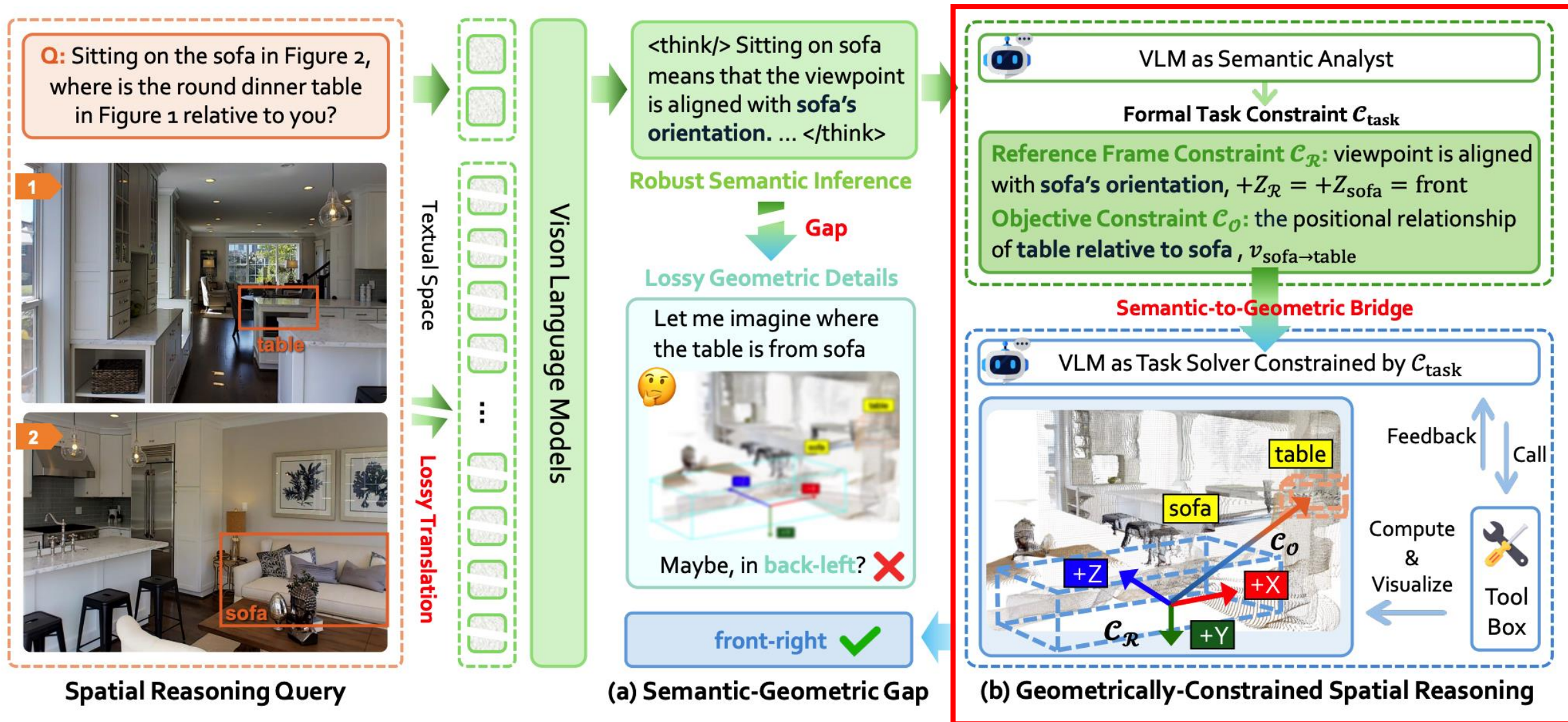
## Alternative method other than thinking in text



# Existing work: Geometrically-Constrained Agent

## ❖ Existing work: Geometrically-Constrained Agent

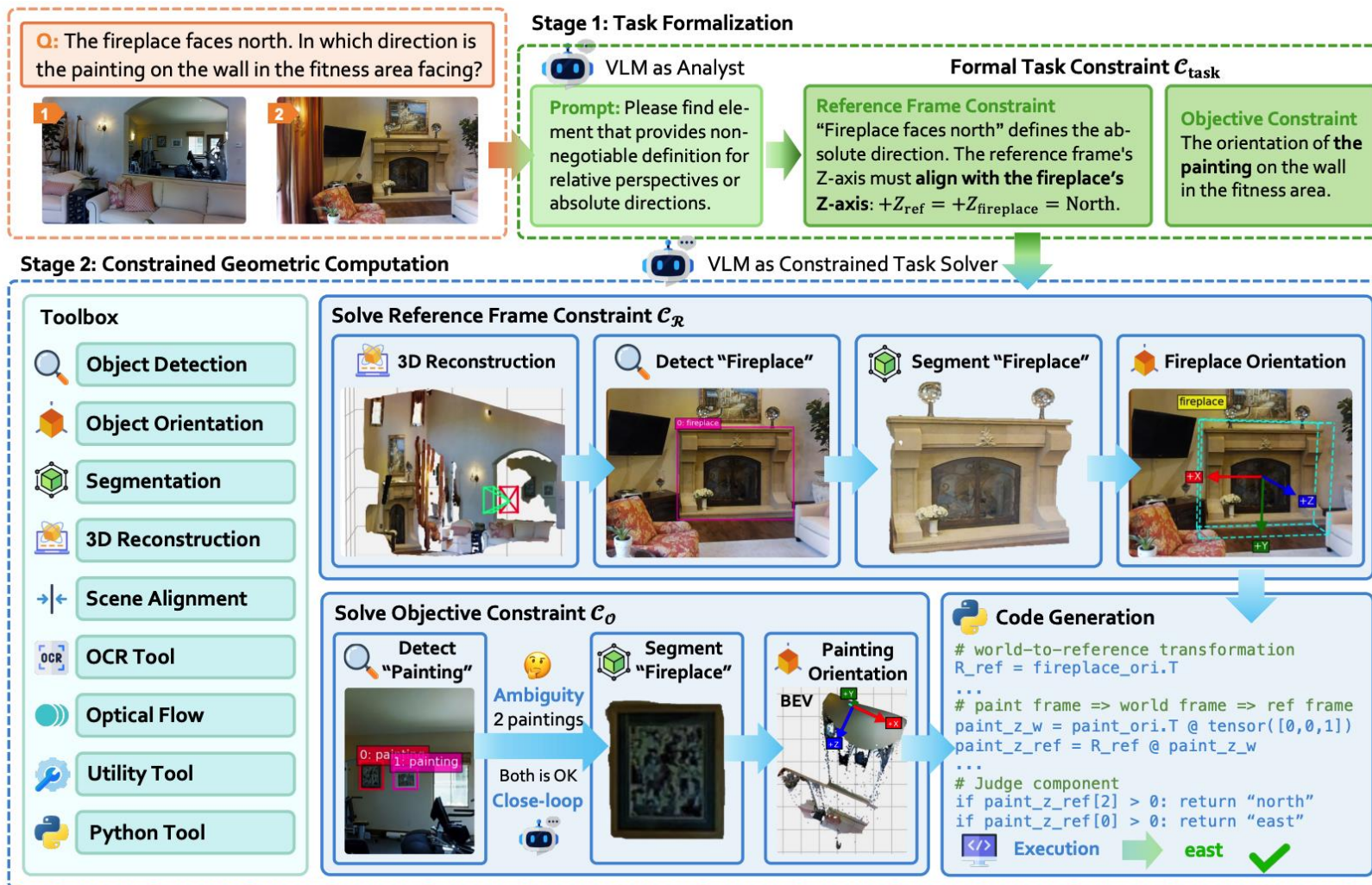
Provide a path of separating semantic reasoning from geometric computation



# Existing work: Geometrically-Constrained Agent

## ❖ Geometrically-Constrained Agent limitation

- Decompose reasoning into coarse task formulation and constrained geometric computation handled by a VLM agent.
- Lack of scalability to more complex downstream applications require spatial reasoning

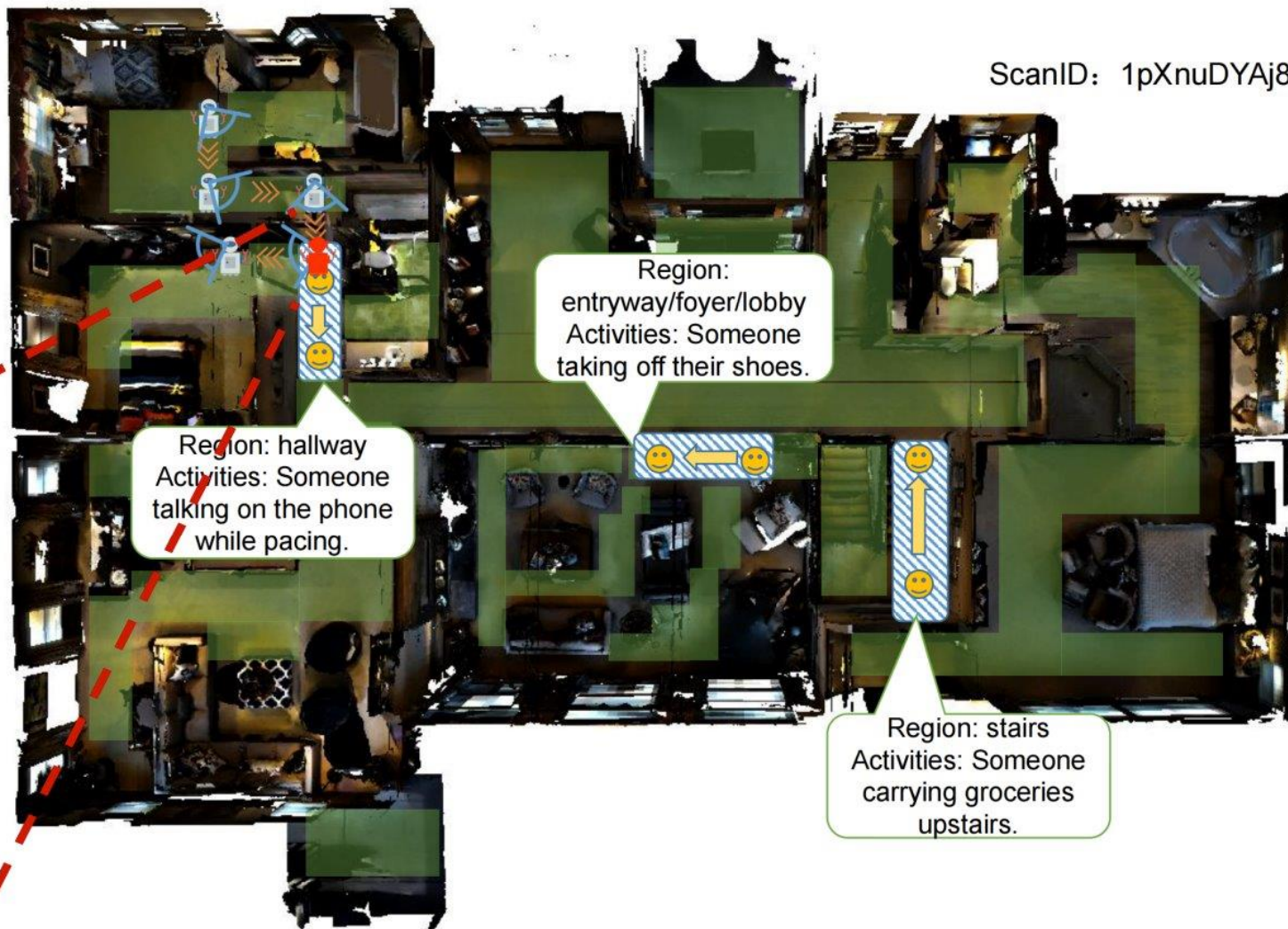
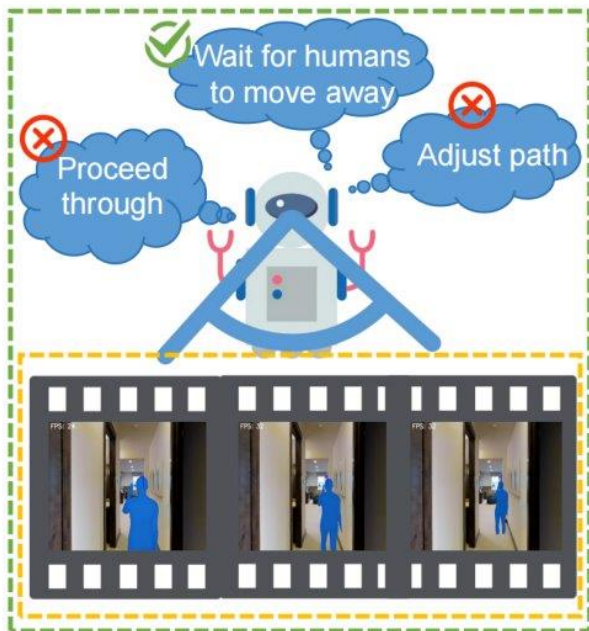


# Potential idea: Agentic workflow on spatial reasoning task (navigation)

Design an agentic workflow with 3D tool use and navigation skills to improve VLM-based robot navigation under dynamic blocking conditions.

## Instruction

Make a **left turn** and **exit** through the door on your right. Proceed to **turn right** and enter the room **next door**, pausing at the entrance. Be aware of someone in the hallway, pacing and potentially talking on the phone, which could serve as an obstacle to your path. Ensure you do not disturb them during your movement.

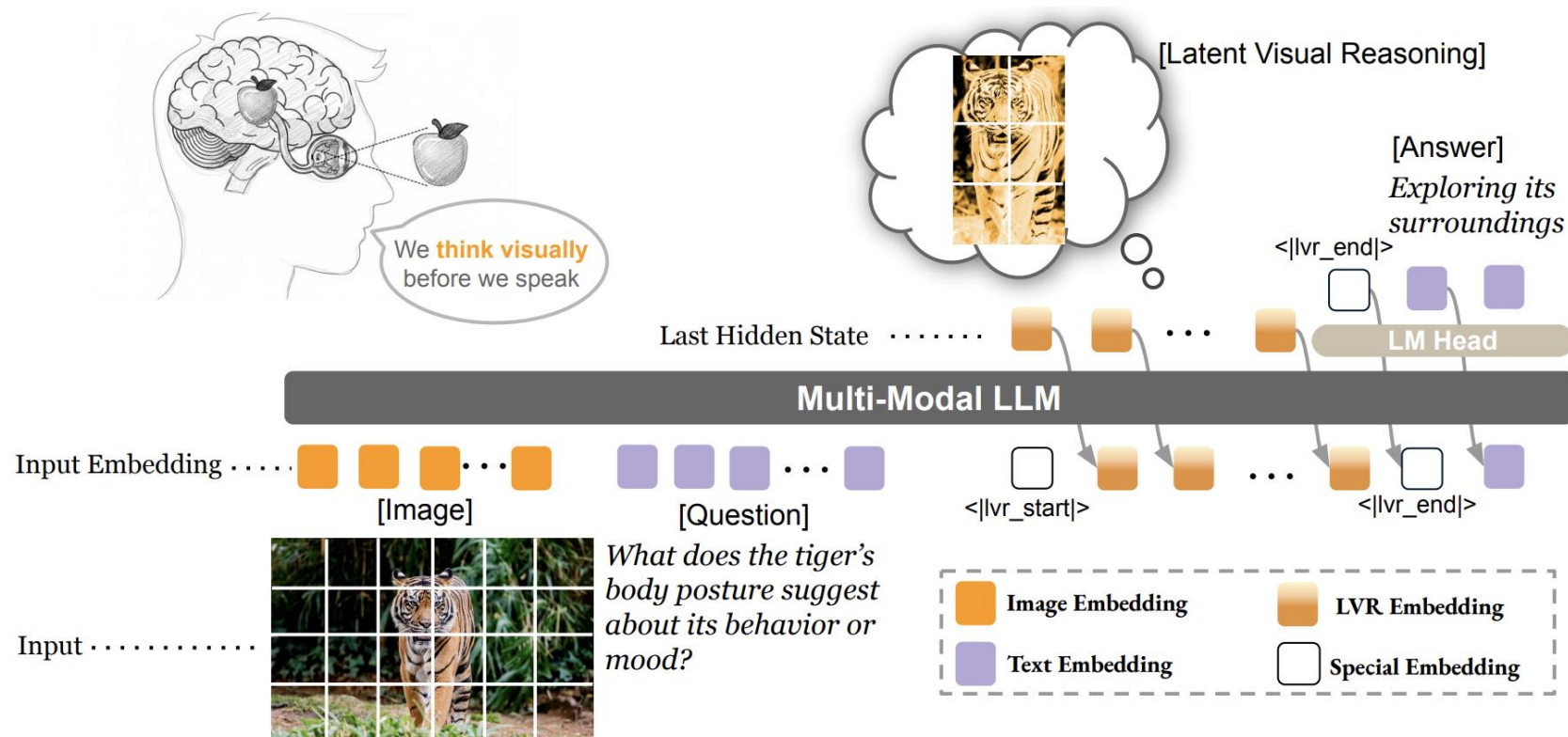


- Agent Path
- Human Path
- Human
- Agent
- Interaction with human
- Human Movable Area
- Agent Navigable Area

# Existing work: Latent spatial reasoning

## ❖ Motivation

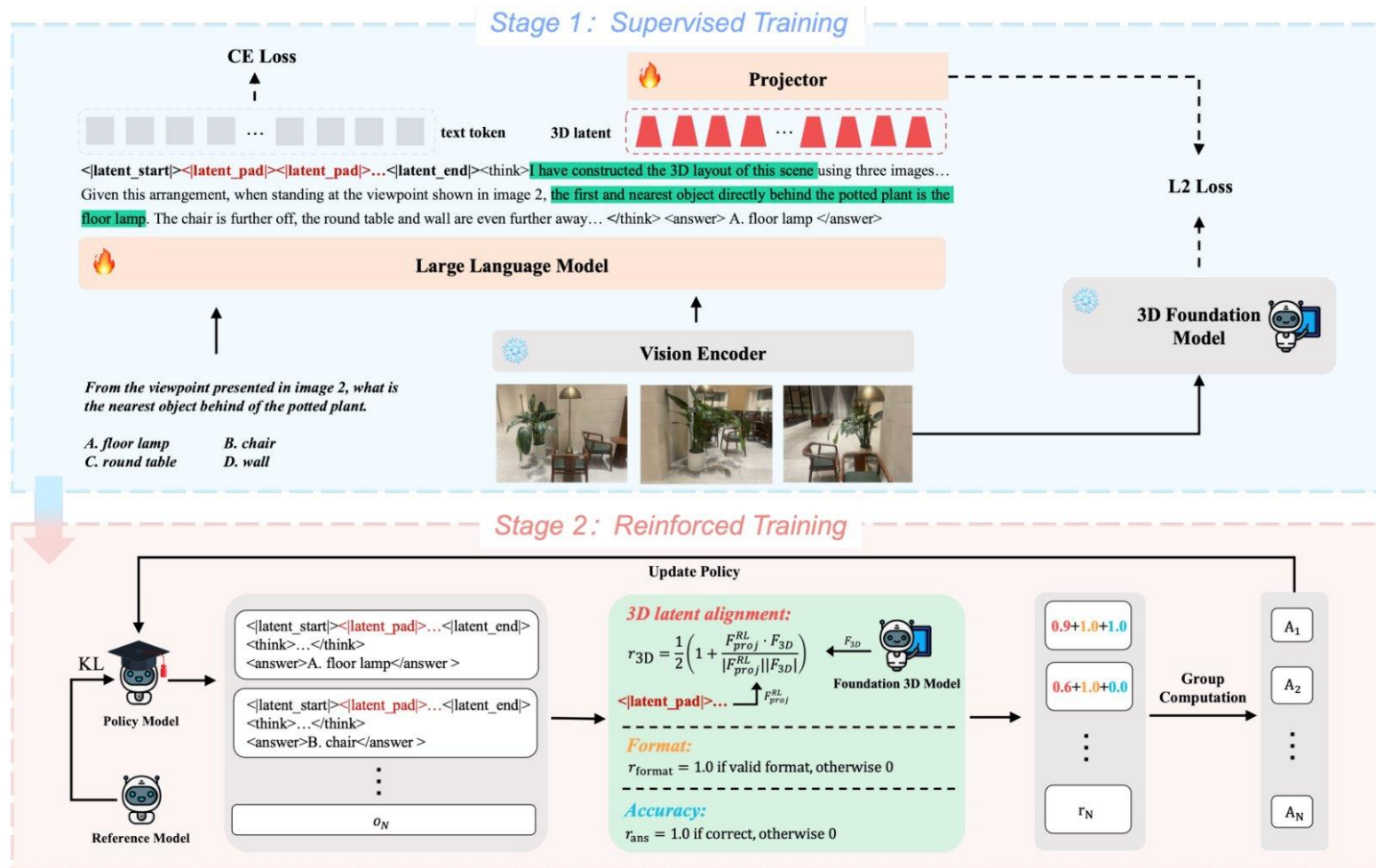
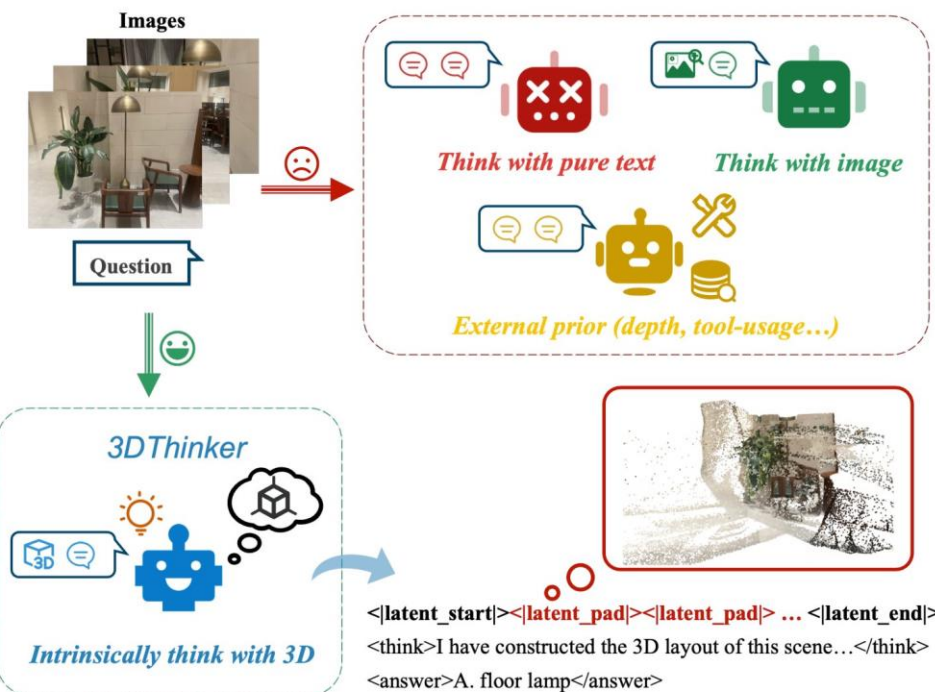
- A unified spatial representation should be embedded in continuous latent space, enabling 3D mental modeling for human-like spatial reasoning.
- Most of current method focusing on 2D image perception rather than spatial reasoning with necessary 3D latent feature guidance.



# Existing work: Latent spatial reasoning

## ❖ Latent spatial reasoning existing work & limitation

- Latent spatial reasoning enables VLMs to reason in continuous vision-language latent space, guided by 3D models like VGGT and trained with SFT plus GRPO-based RLVR.
- Only dense 3D features are used in latent reasoning, while fine-grained 2D spatial supervision at intermediate reasoning steps is missing.



# Potential idea: Latent spatial reasoning

## ❖ Latent spatial reasoning potential idea

Extract step-level spatial cues to supervise reasoning during view shifts.

