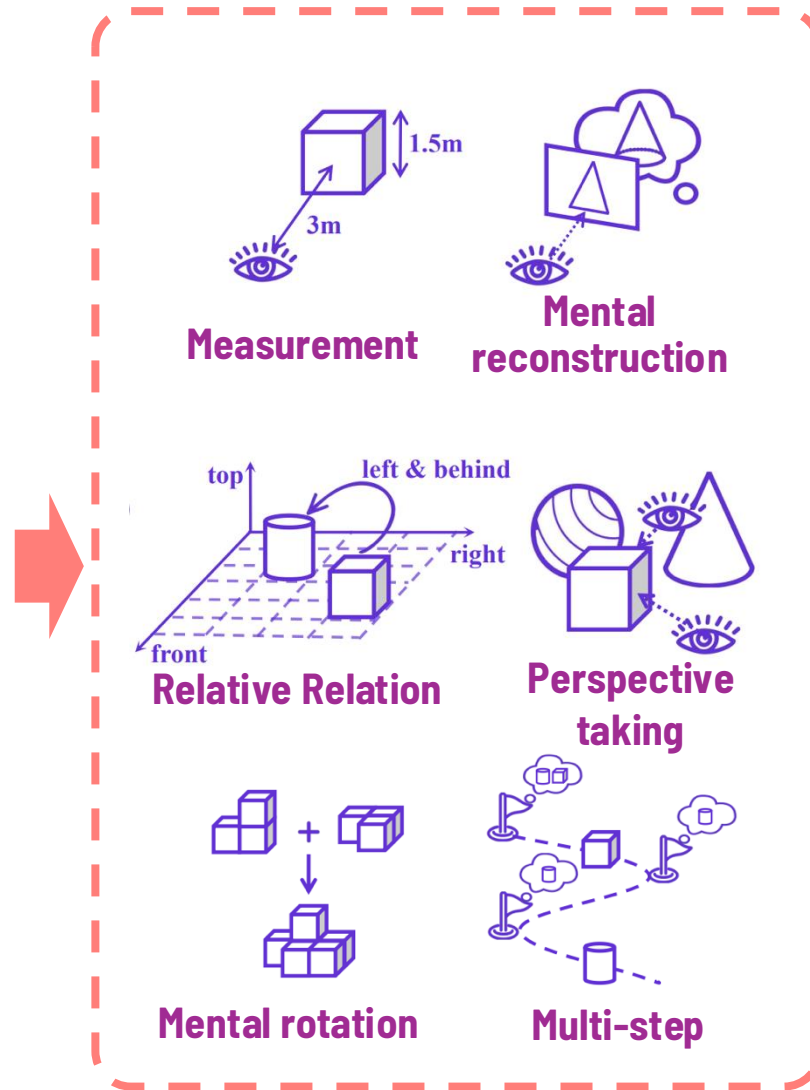


Fundamental Capabilities of Spatial Intelligence



**Egocentric
Visual inputs**



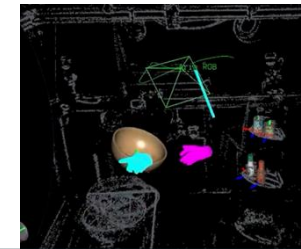
**Basic Spatial
Reasoning Capabilities**



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.

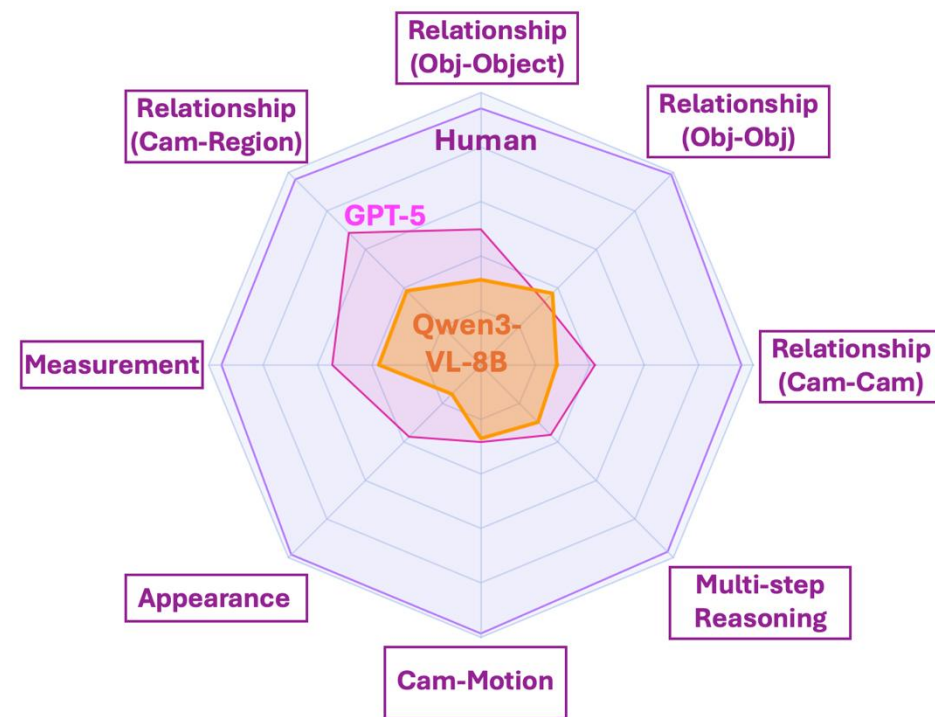
GT: -

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



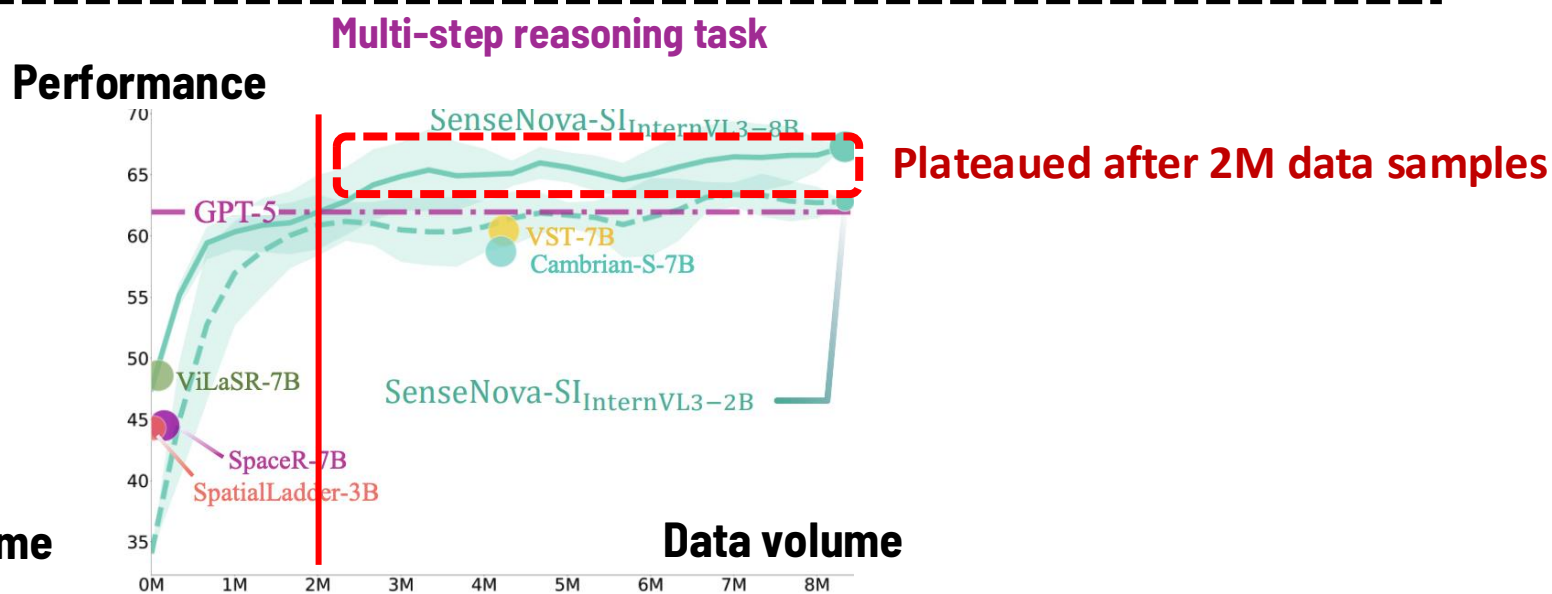
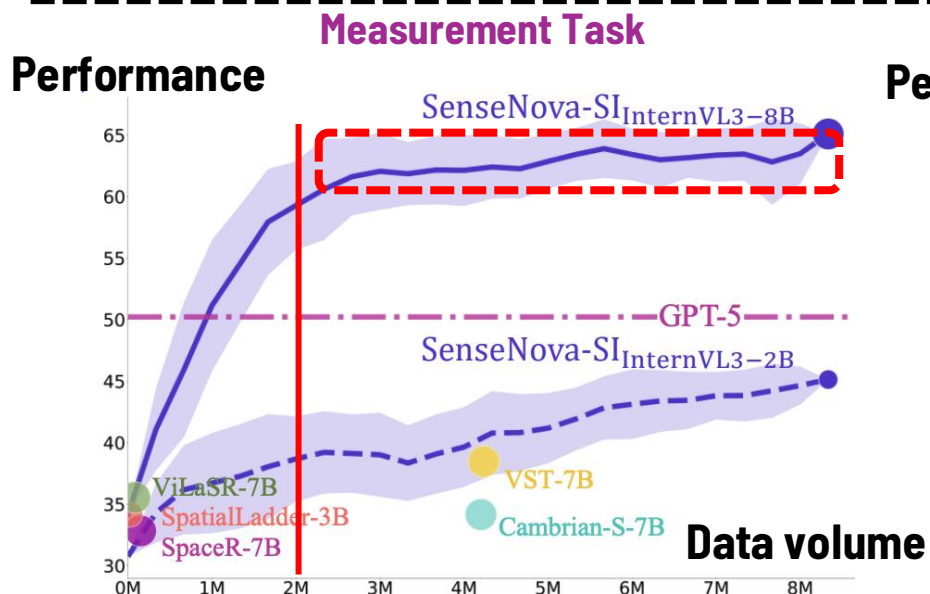
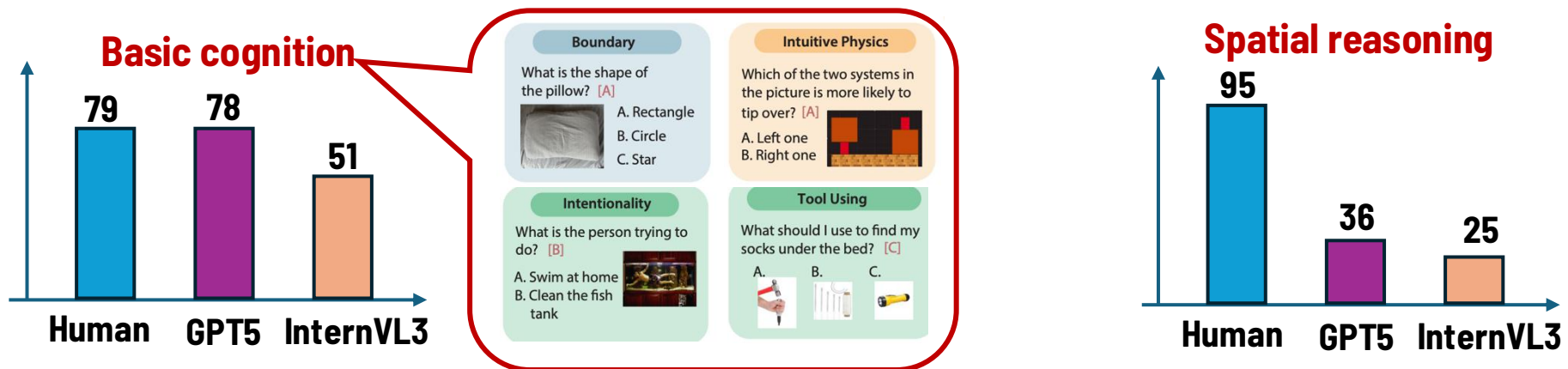
**Human VS Proprietary models
VS Open-sourced general models**

[1] Has GPT-5 Achieved Spatial Intelligence? An Empirical Study

[2] How Far are VLMs from Visual Spatial Intelligence? A Benchmark-Driven Perspective

Can we solve spatial reasoning by further scaling training data and model size

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

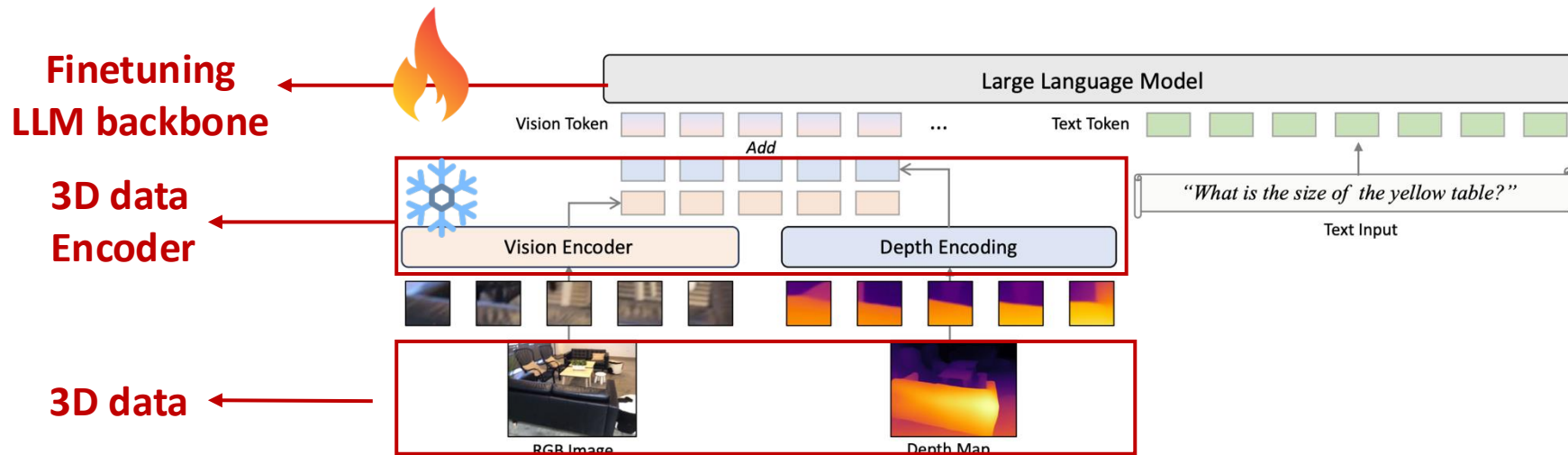


[3] Why Do MLLMs Struggle with Spatial Understanding? A Systematic Analysis from Data to Architecture

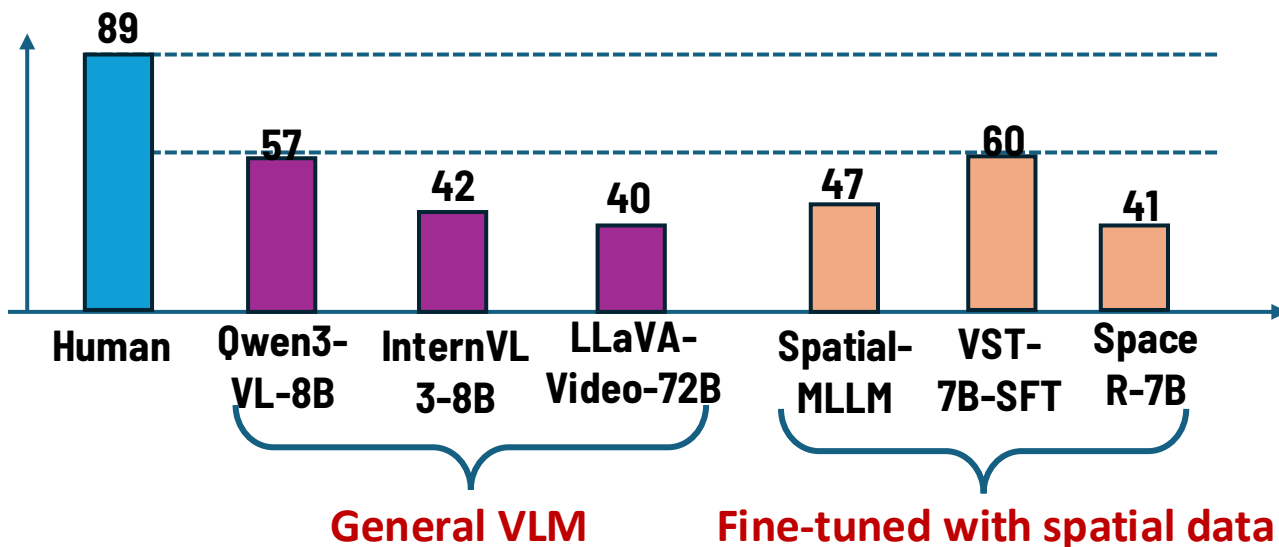
[4] Scaling and Beyond: Advancing Spatial Reasoning in MLLMs Requires New Recipes

Scaling with diverse input modalities

Typical pipeline of adding new modality to VLM input layer



Limited improvement:

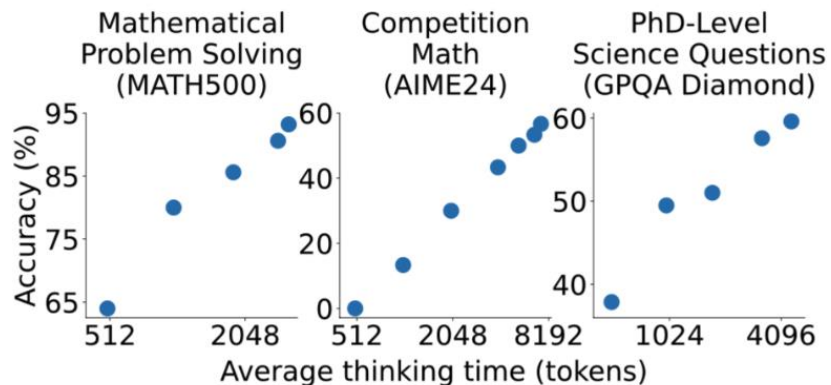


Scaling with diverse input modalities

- [5] **Learning from Videos for 3D World: Enhancing MLLMs with 3D Vision Geometry Priors**
- [6] **S2-MLLM: Boosting Spatial Reasoning Capability of MLLMs for 3D Visual Grounding with Structural Guidance**
- [7] **SD-VLM: Spatial Measuring and Understanding with Depth-Encoded Vision-Language Models**
- [8] **SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models**

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  $2ab + a^2 = 0$ :
 $2 \left(-\frac{3}{2}\right) \cdot 0 + \left(-\frac{3}{2}\right)^2 = 0 + \frac{9}{4} = \frac{9}{4}$ .

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

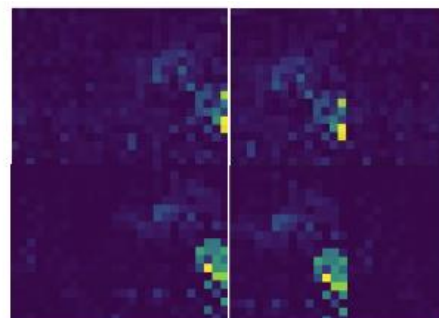
Ahah moment in reasoning

Thinking in text domain impair visual spatial reasoning

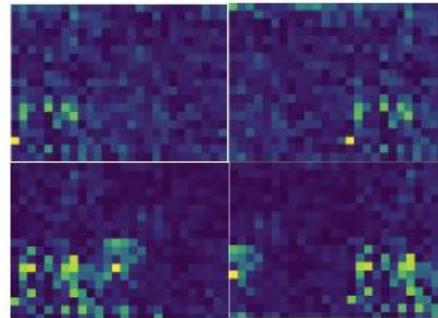
Model	Inference Method	Occlusion Restoration	Distance Comparison	Azimuth Transfer
Qwen2.5-VL	Vanilla	34.52	43.41	27.25
	Explicit Stepwise CoT	33.0 (-1.51)	34.4 (-9.00)	23.0 (-4.25)
	Explicit Multi-view CoT	35.04 (+0.52)	34.16 (-9.25)	24.29 (-2.96)



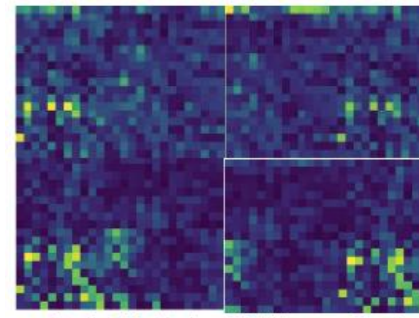
Q: Which object is closest to the bike in both images?
 A. Sofa
 B. blue mat
 C. indoor_plant
 D. Other



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

[9] Mind journey: Test-Time Scaling with World Models for Spatial Reasoning

[10] When and How Much to Imagine: Adaptive Test-Time Scaling with World Models for Visual Spatial Reasoning

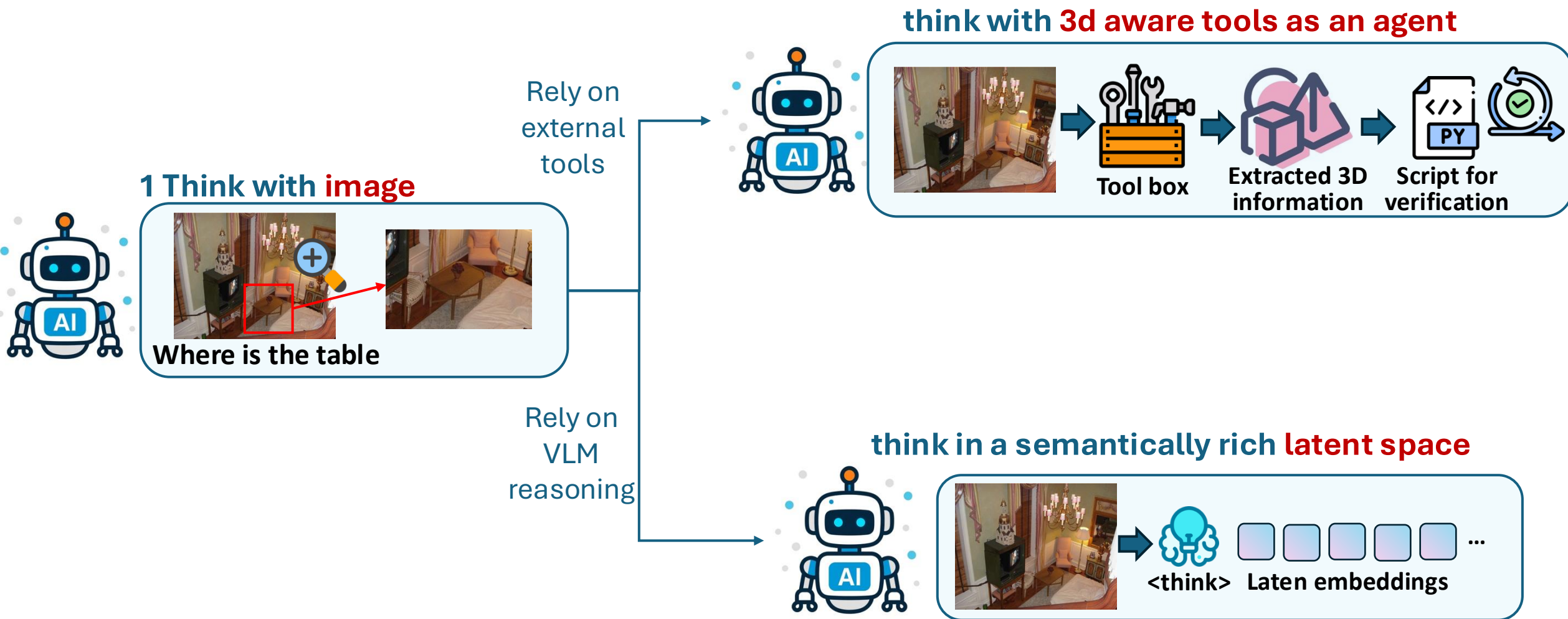
[11] SpatialDreamer: Incentivizing Spatial Reasoning via Active Mental Imagery

[12] WMNav: Integrating Vision-Language Models into World Models for Object Goal Navigation

[13] Sparse Imagination for Efficient Visual World Model Planning

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

Q: Sitting on the sofa in Figure 2, where is the round dinner table in Figure 1 relative to you?



1

2

Spatial Reasoning Query



Visual Language Models

<think/> Sitting on sofa means that the viewpoint is aligned with **sofa's orientation**. ... </think>

Robust Semantic Inference

Gap

Lossy Geometric Details

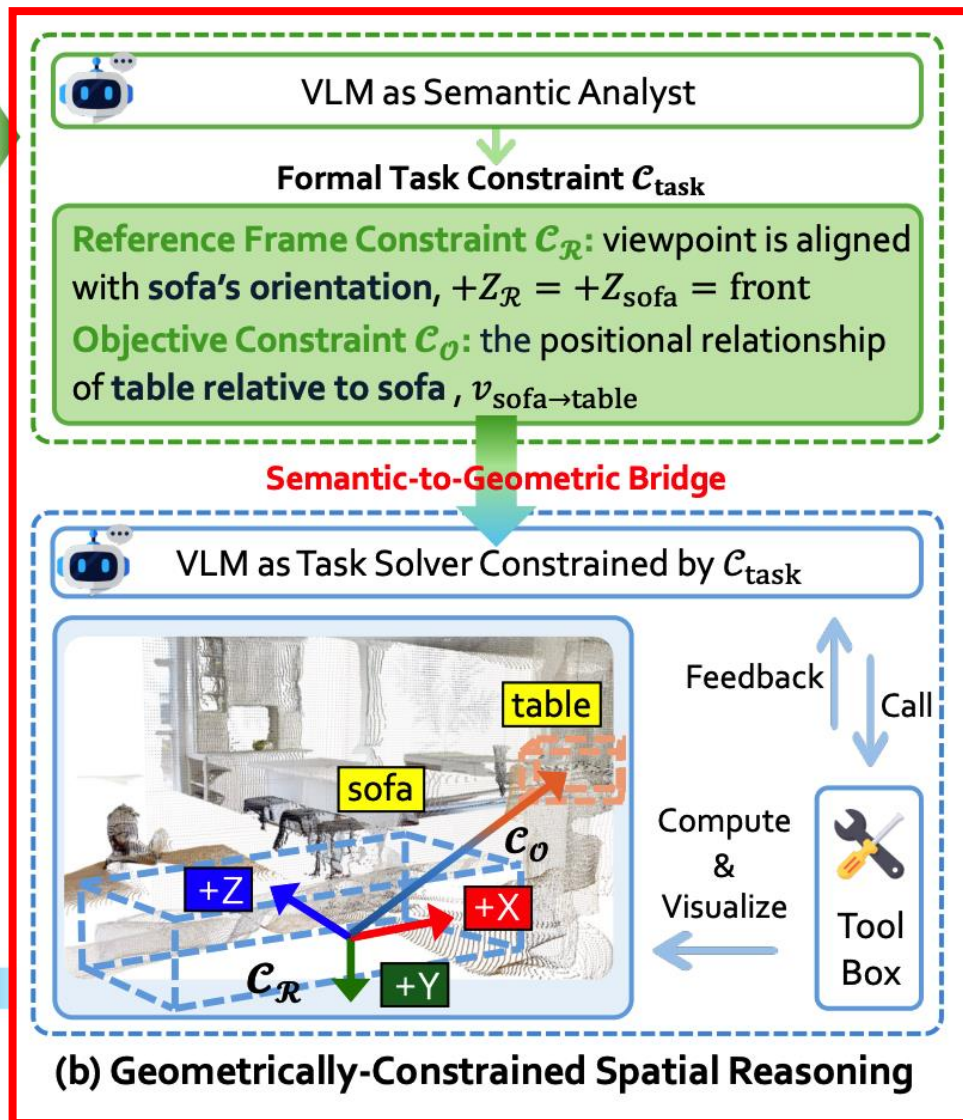
Let me imagine where the table is from sofa



Maybe, in **back-left?** ❌

front-right ✅

(a) Semantic-Geometric Gap



(b) Geometrically-Constrained Spatial Reasoning

More existing works on thinking with tools

[14] World2Mind: Cognition Toolkit for Allocentric Spatial Reasoning in Foundation Models

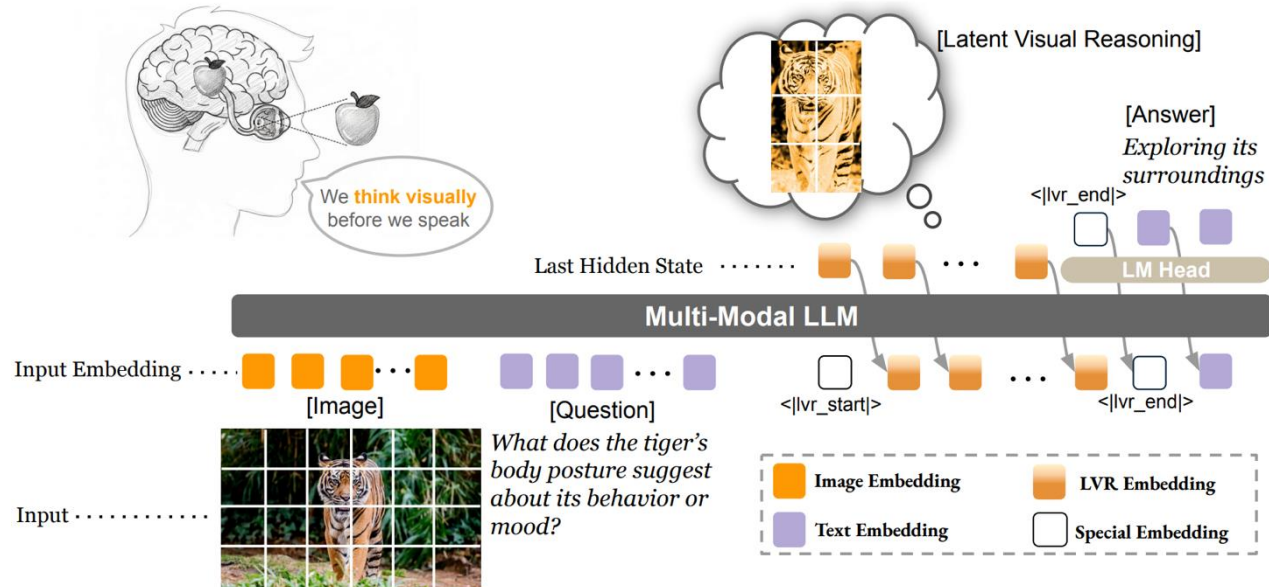
[15] Geometrically-Constrained Agent for Spatial Reasoning

[16] RieMind: Geometry-Grounded Spatial Agent for Scene Understanding

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.



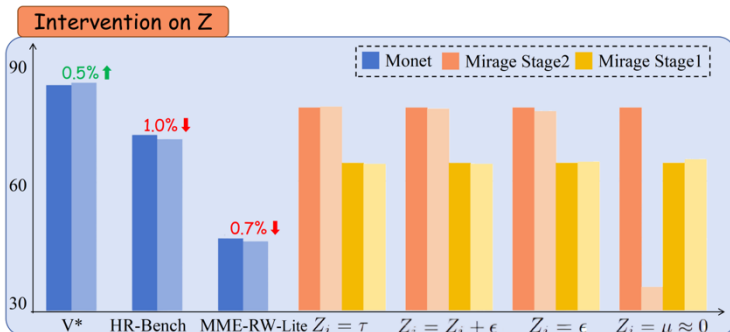
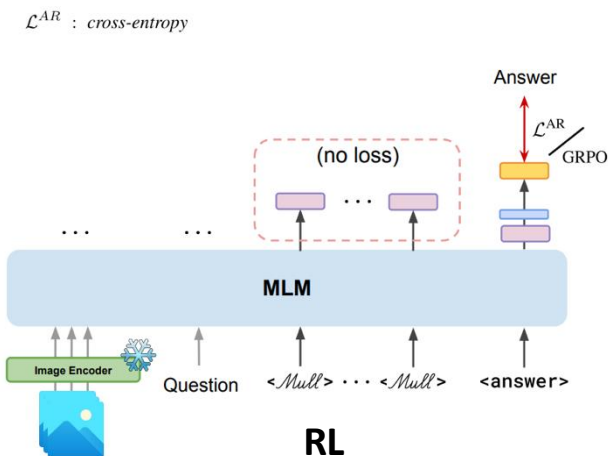
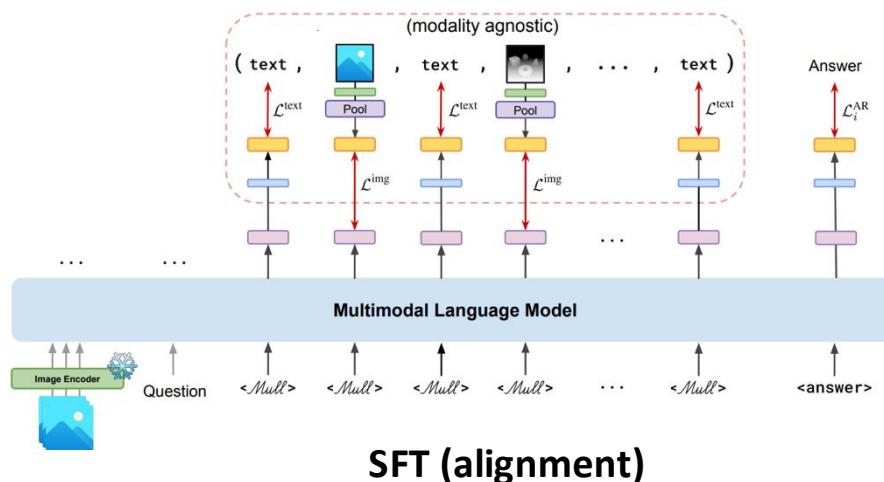
The listed papers include some general-purpose multimodal latent-space reasoning methods that are not specifically designed for spatial reasoning, but are still worth reading.

Latent reasoning for spatial reasoning

Three potential pathways

1 latent alignment + RL

- 1 teach the model to generate token in a semantically rich latent space (e.g. learn to generate VGGT embeddings)
- 2 does not teach the model how to leverage the generated latent information for reasoning; instead, the reasoning capability relies entirely on reinforcement learning



Any way to guarantee that the model **derives its final decision from latent reasoning**, rather than relying solely on the raw image and text inputs?

Latent reasoning for spatial reasoning

Three potential pathways

1 latent alignment + RL

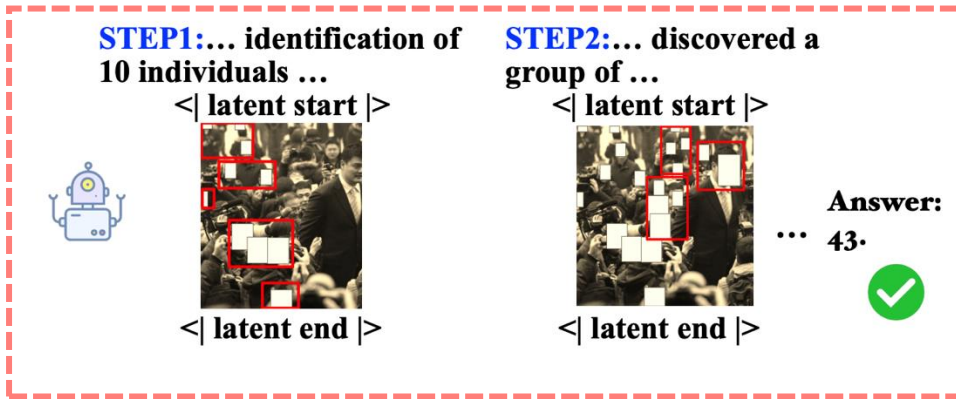
- [17] Think with 3D: Geometric Imagination Grounded Spatial Reasoning from Limited Views
- [18] Monet: Reasoning in Latent Visual Space Beyond Image and Language
- [19] Crystal: Spontaneous Emergence of Visual Latents in MLLMs
- [20] Mull-Tokens: Modality-Agnostic Latent Thinking
- [21] Reasoning Within the Mind: Dynamic Multimodal Interleaving in Latent Space
- [22] Latent Chain-of-Thought for Visual Reasoning
- [23] Latent Reasoning VLA: Latent Thinking and Prediction for Vision-Language-Action Models

Latent reasoning for spatial reasoning

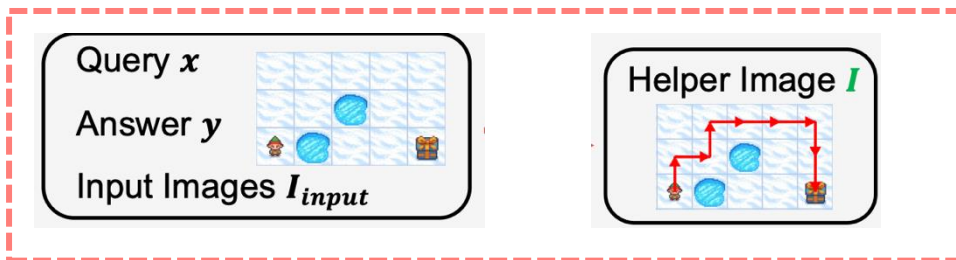
Three potential pathways

2 guide the reasoning with ground truth reasoning traces

Option 1: teach the model to focus on the correct part of input



Option 2: teach the model with mental imaginary operation



Trace Coherence (Local Plausibility): Do the individual, step-by-step sentences make sense? Are there obvious errors between step A and step B?


Trace Validity (Global Validity): Does the entire chain of thought logically sound that leads to the correct final answer

Existing multi-model reasoning data

Highlight task-related region

STEP1:... identification of 10 individuals ...


<| latent start |>



<| latent end |>

STEP2:... discovered a group of ...


<| latent start |>




<| latent end |>

Answer: ... 43. ✓

Caption the input image



In the first image, the teddy bear is rotated sideways ...

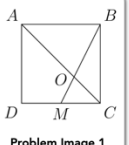


In the second image, the teddy bear's face is oriented toward the camera

2D spatial/geometric reasoning with sketch

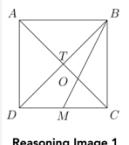
Geometry

Question: In square $ABCD$, AD is 4 centimeters, and M is the midpoint of \overline{CD} . Let O be the intersection of \overline{AC} and \overline{BM} . What is the ratio of OC to OA ?



Problem Image 1

<think>Thought 1: To better visualize the geometric relationships, especially within triangle BCD , let's add the diagonal BD to our diagram. Let T be the intersection of AC and BD . This will help us identify if any segment related to AC can be considered a median of triangle BCD . </think>




Reasoning Image 1

<think>Thought 2: The updated diagram shows the square $ABCD$ with both diagonals AC and BD intersecting at point T . The segment BM intersects AC at O . Now, let's focus on triangle BCD ... </think>

Spatial Planning


Devise an action plan that enables a player to reach the goal.



3D spatial reasoning with toy data


3D Multi Hop Object Counting

Question: Remove zero CAPSULE objects. Eliminate all PYRAMID objects. Remove zero PURPLE objects. Discard all CUBE objects. Count how many blue-colored objects remaining.



Problem Image 1

<think>Thought 2: Proceeding to Step 2, the instruction is to "Remove all PYRAMID objects." Upon reviewing the current scene, there is one orange pyramid present. This object must be removed from this removal is depicted in the following image. </think>



Reasoning Image 1

<think>Thought 3: For Step 3, the directive states to "Remove 0 PURPLE object." Similar to Step 1, this operation specifies that no purple objects are to be removed. The objects remaining in the scene after Step 2 are preserved... </think>

Data Source	Problem Domain	Visual Operation Type	Amount
ReFocus [12]	Chart	Drawing bounding boxes, highlighting	0.4K
CogCoM [24]	Real-world, chart	Cropping, drawing auxiliary lines and bounding boxes	0.5K
Visual-CoT [26]	Real-world, documents, chart	Cropping, drawing bounding boxes	118.6K
Zebra-CoT visual search [18]	Real-world, documents, chart	Cropping, drawing bounding boxes	2.7K
Zebra-CoT geometry [18]	Geometry	Drawing auxiliary lines, creating new geometric sketchpads	0.1 K
Zebra-CoT count [18]	3D object counting	Creating new images with 3D objects removed or added	2.9K

Statistics of monet-sft-125k

Latent reasoning for spatial reasoning

Three potential pathways

2 guide the reasoning with ground truth reasoning traces

[24] Sketch-in-Latents: Eliciting Unified Reasoning in MLLMs

[25] Vision-aligned Latent Reasoning for Multi-modal Large Language Model

[26] Interleaved Latent Visual Reasoning with Selective Perceptual Modeling

[27] VisMem: Latent Vision Memory Unlocks Potential of Vision-Language Models

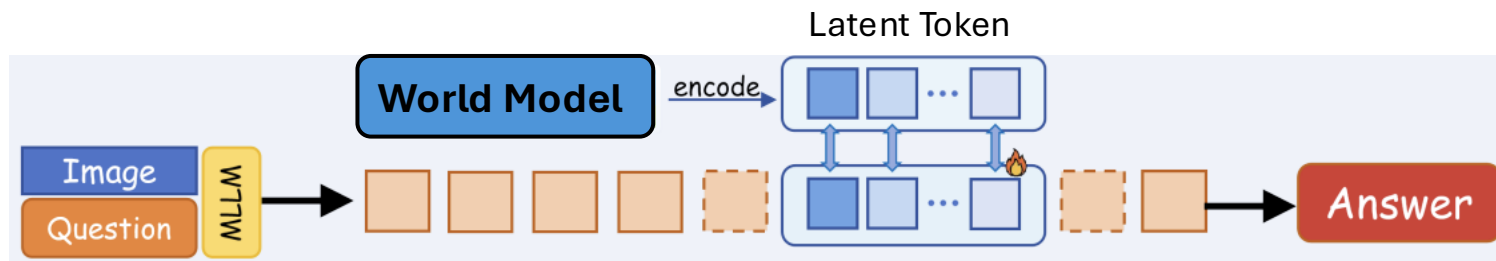
[28] Multimodal Reasoning via Latent Refocusing

Latent reasoning for spatial reasoning

Three potential pathways

3 world model guided latent reasoning

Advantage: easy to get reasoning trace ground truth (predicted world state based on action)



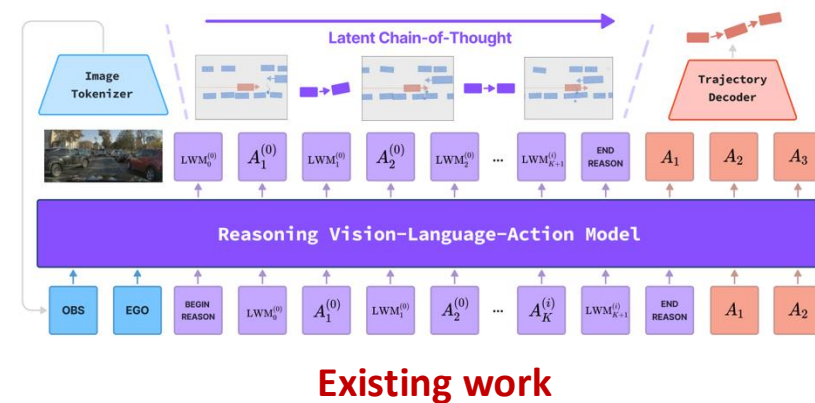
Limitation: not applicable to all tasks

Applicable: Where is object A relative to me if turn right

Future world state
(e.g. image after
turning)

Condition
(action)

Not applicable: What's the spatial relation between object A and B
What's the action?



Latent reasoning for spatial reasoning

Three potential pathways

3 world model guided latent reasoning

[29] Latent Chain-of-Thought World Modeling for End-to-End Autonomous Driving

[30] LaST-VLA: Thinking in Latent Spatio-Temporal Space for Vision-Language-Action in Autonomous Driving

[31] CAN WORLD MODELS BENEFIT VLMS FOR WORLD DYNAMICS?

[32] JEPA-VLA: Video Predictive Embedding is Needed for VLA Models

[33] Policy-Guided World Model Planning for Language-Conditioned Visual Navigation