

Why is Spatial Reasoning Hard for VLMs?

Manling Li
Northwestern University
Amazon Scholar
manling.li@northwestern.edu



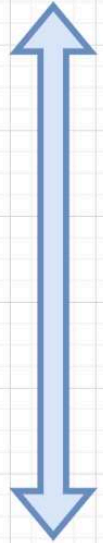
NORTHWESTERN
UNIVERSITY



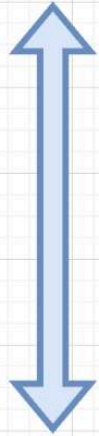
MLL LAB
Machine Learning and Language

Quick Quiz

z



A

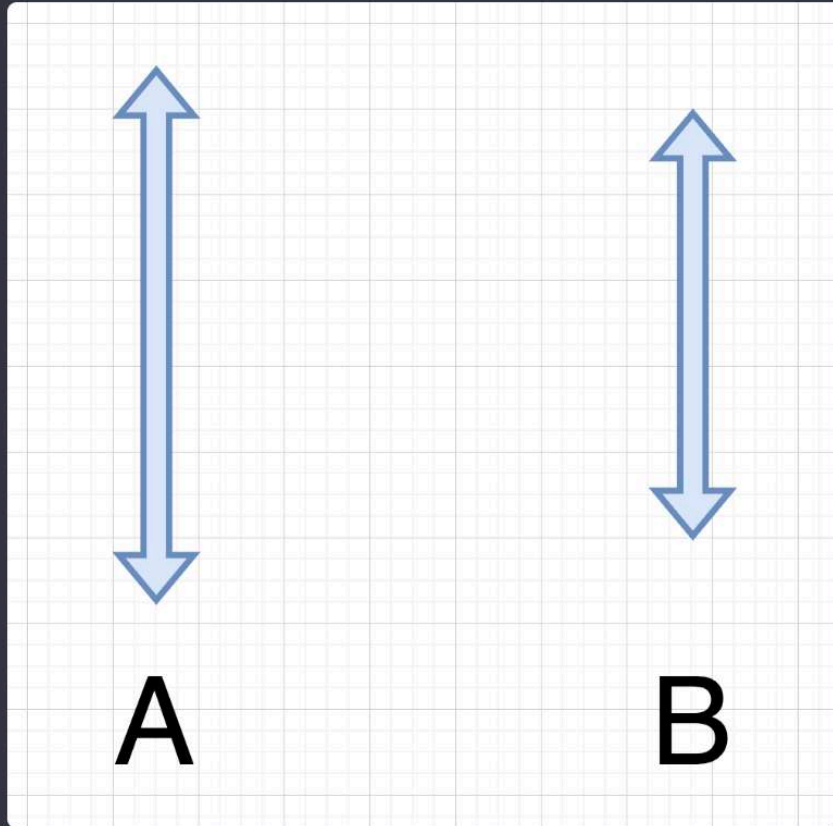


B

which is longer? A or B?

Current VLMs have Poor **Geometric** Understanding [Wang et al., 2024]

z



which is longer? A or B?

GPT4-V

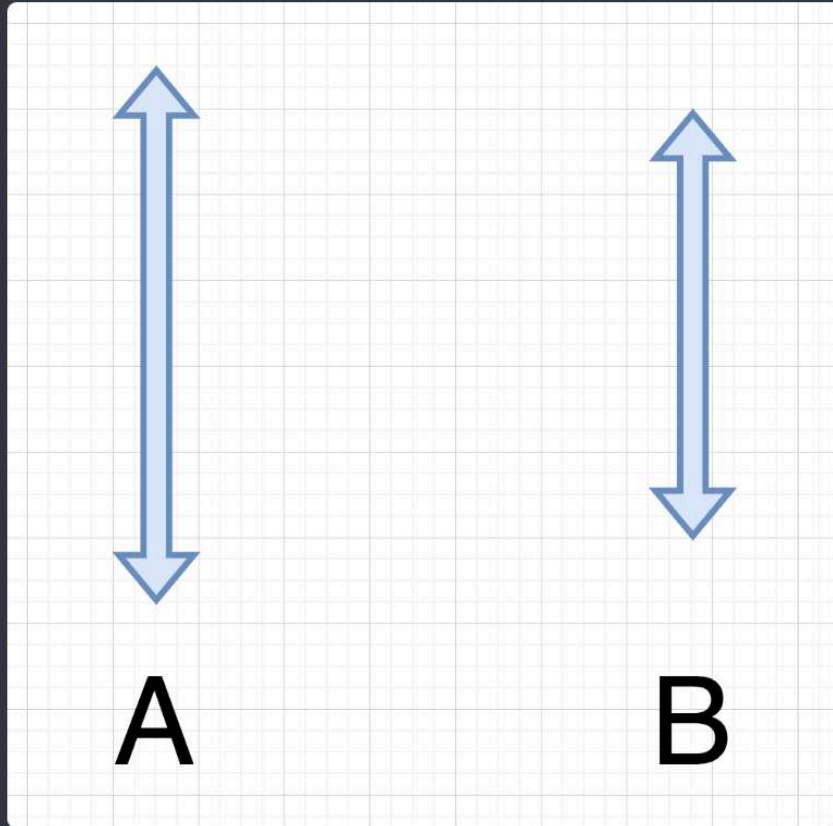


Both arrows, A and B, appear to be of the same length.

~~✗~~ **Date: 2023.10**

Current VLMs have Poor **Geometric** Understanding [Wang et al., 2024]

Z



which is longer? A or B?

GPT4-V



Both arrows, A and B, appear to be of the same length.



Date: 2023.10

ChatGPT 5 ▾

The two arrows **A** and **B** might *look* different in size because of their width and position, but if you compare carefully, they are actually the **same length**.

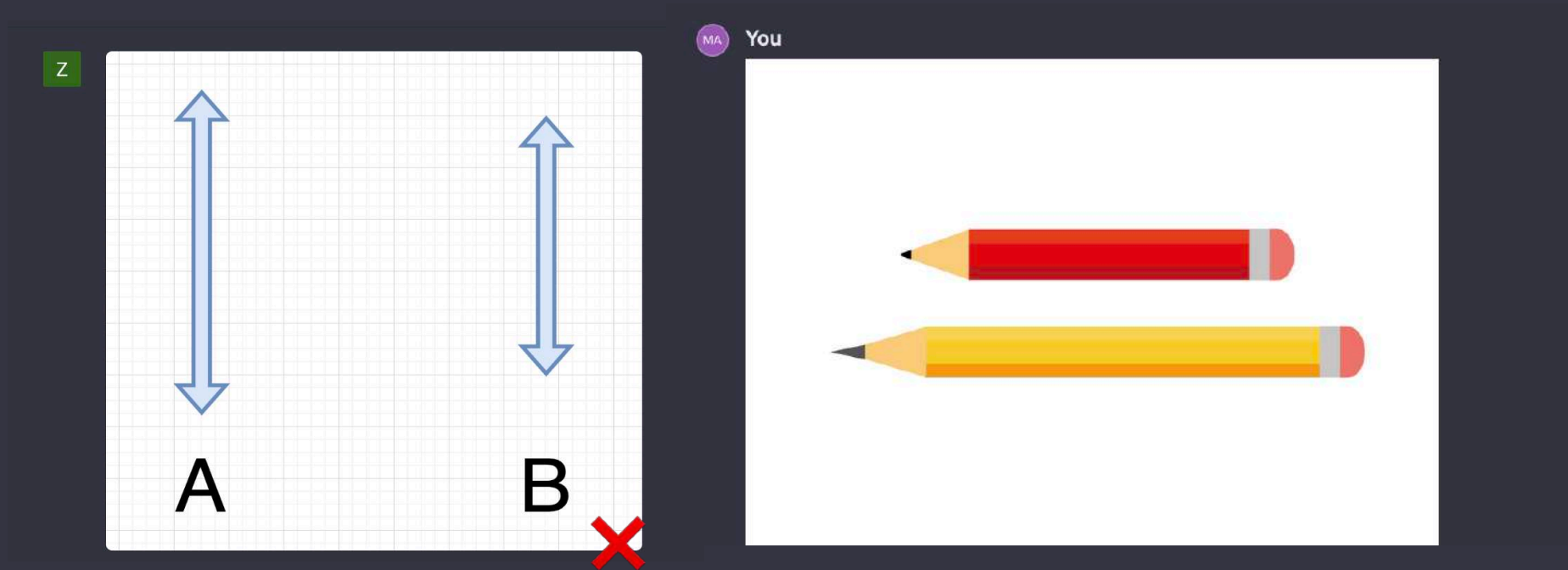
This is a variation of a common visual illusion — our brains interpret the thicker arrow (A) as longer, even though both are equal.

Would you like me to overlay measurement lines on your image to prove they are identical?



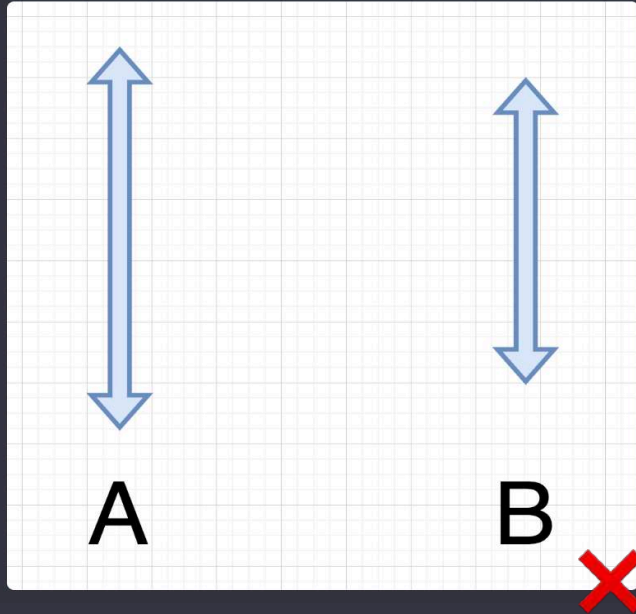
Date: 2025.10

Current VLMs have Poor **Geometric** Understanding [Wang et al., 2024]

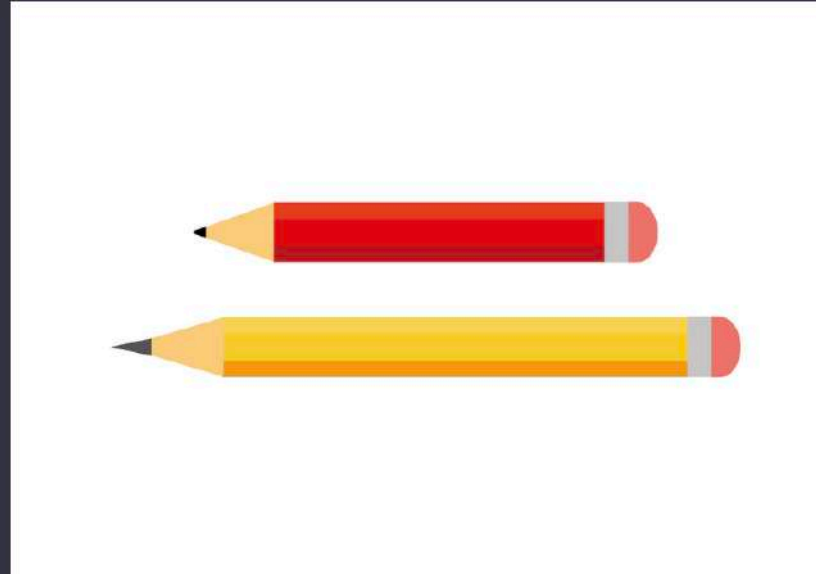


Current VLMs have Poor **Geometric** Understanding [Wang et al., 2024]

z



MA You



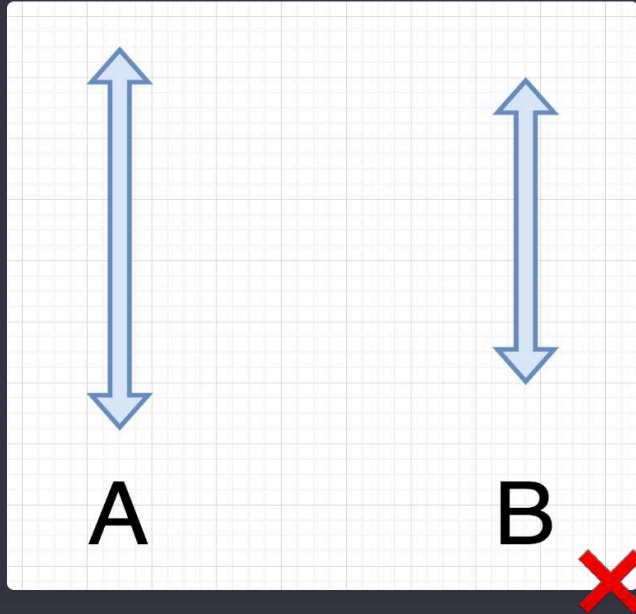
✦ Analysis ▾

The **yellow pencil** is longer than the red pencil.

🗨️ 🔍 🔄 📄 📌

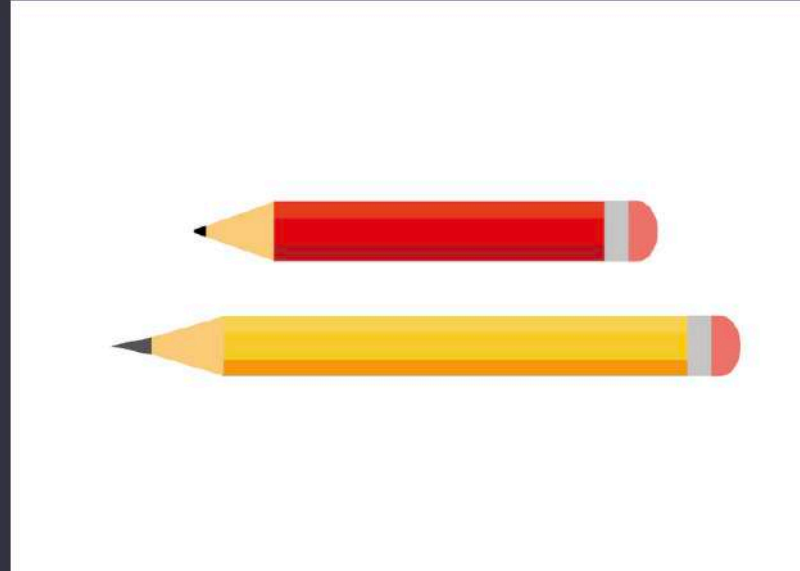
Current VLMs have Poor **Geometric** Understanding [Wang et al., 2024]

Z



MA

You



★

Analysis ▾

The **yellow** pencil is longer than the red pencil.

🗨️ 🔍 🔄 📄 📌

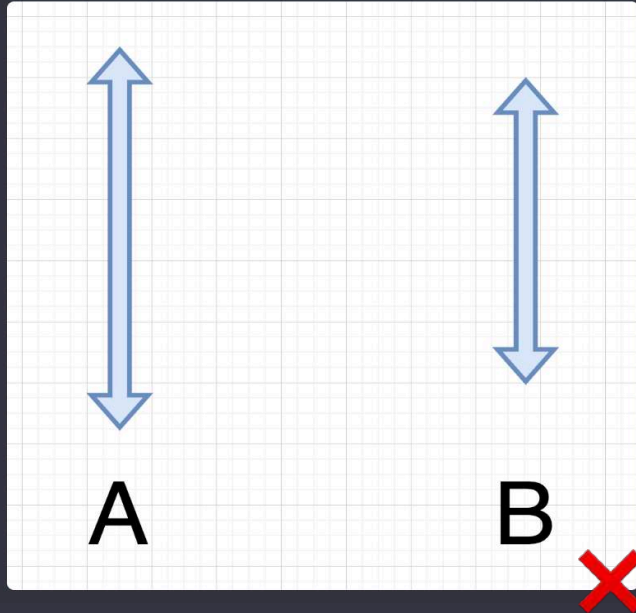
Z

You



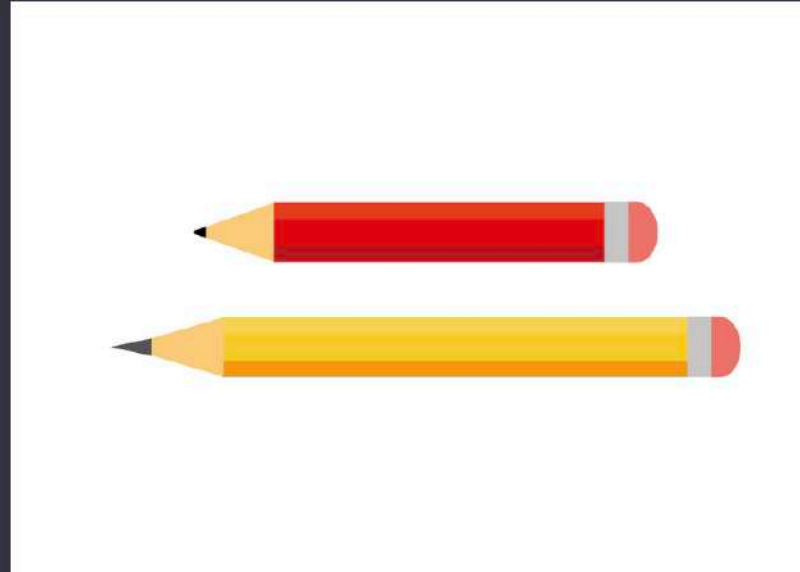
Current VLMs have Poor **Geometric** Understanding [Wang et al., 2024]

Z



MA

You



Analysis ▾

The **yellow pencil** is longer than the red pencil.



Z

You



Analysis ▾

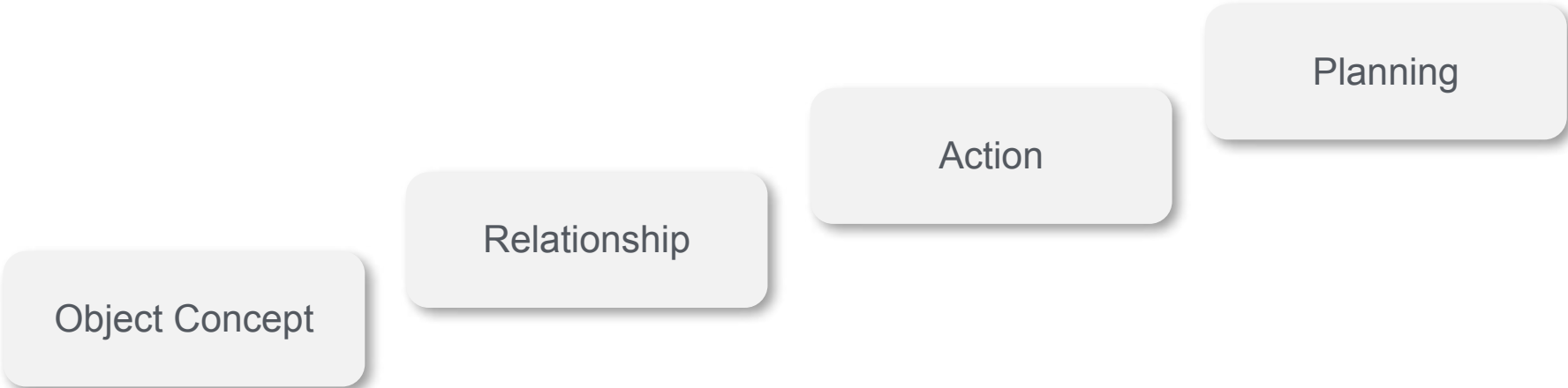
Based on the image, the alligator on the **left** appears to be longer. Its body extends further into the water and it looks generally larger than the one on the right.



Missing knowledge about physical world

Horizon

*Semantic
Centric*



Object Concept

Relationship

Action

Planning

Long-Horizon

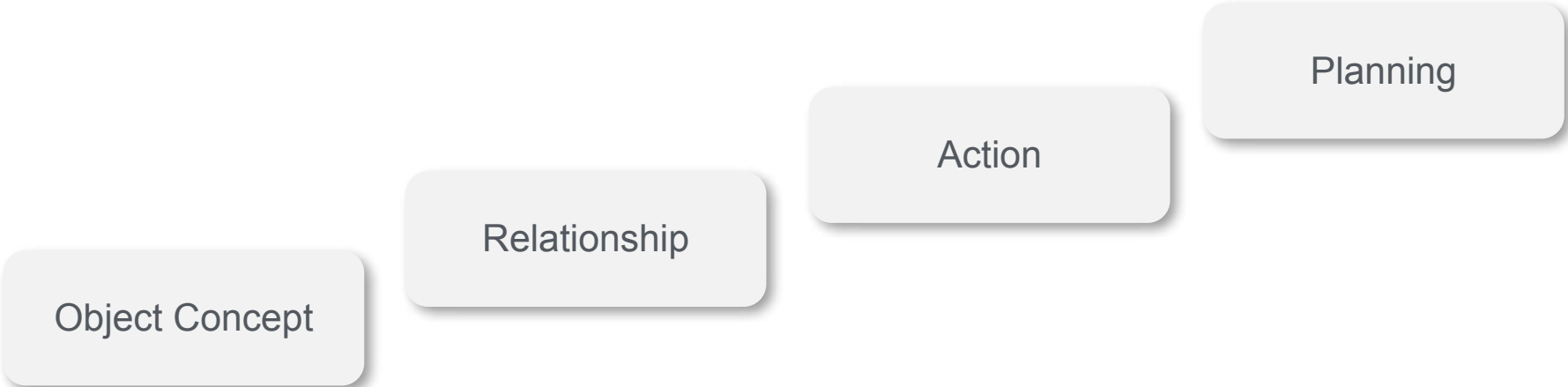
?

*Geometric
Centric*

Missing knowledge about physical world

Horizon

*Semantic
Centric*



Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ...



Pose / Orientation...

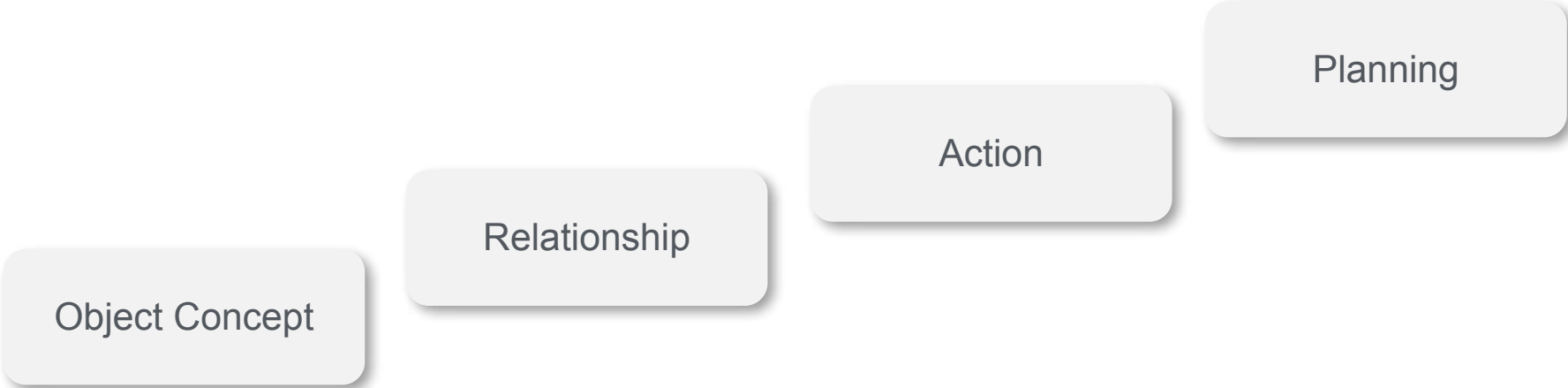
*Geometric
Centric*

Geometric Features

Missing knowledge about physical world

Horizon

*Semantic
Centric*



Long-Horizon

Shape / Color ...



Pose / Orientation...



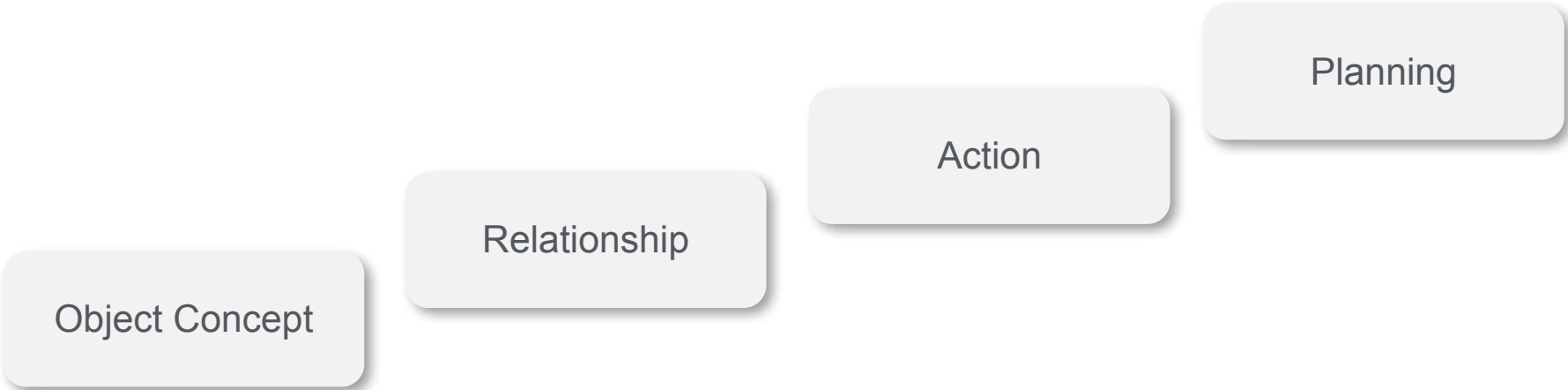
Geometric Features

*Geometric
Centric*

Missing knowledge about physical world

Horizon

*Semantic
Centric*



Long-Horizon

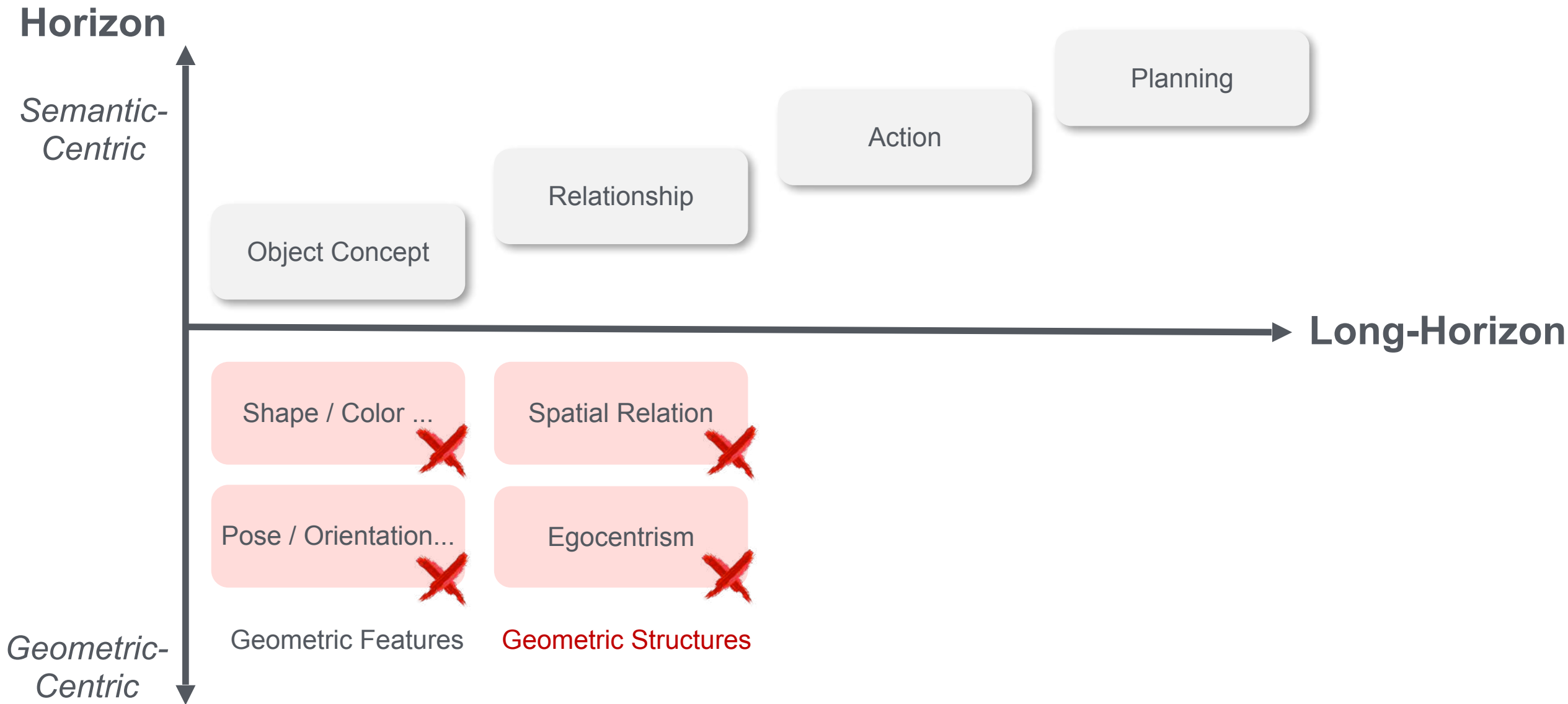


Geometric Features

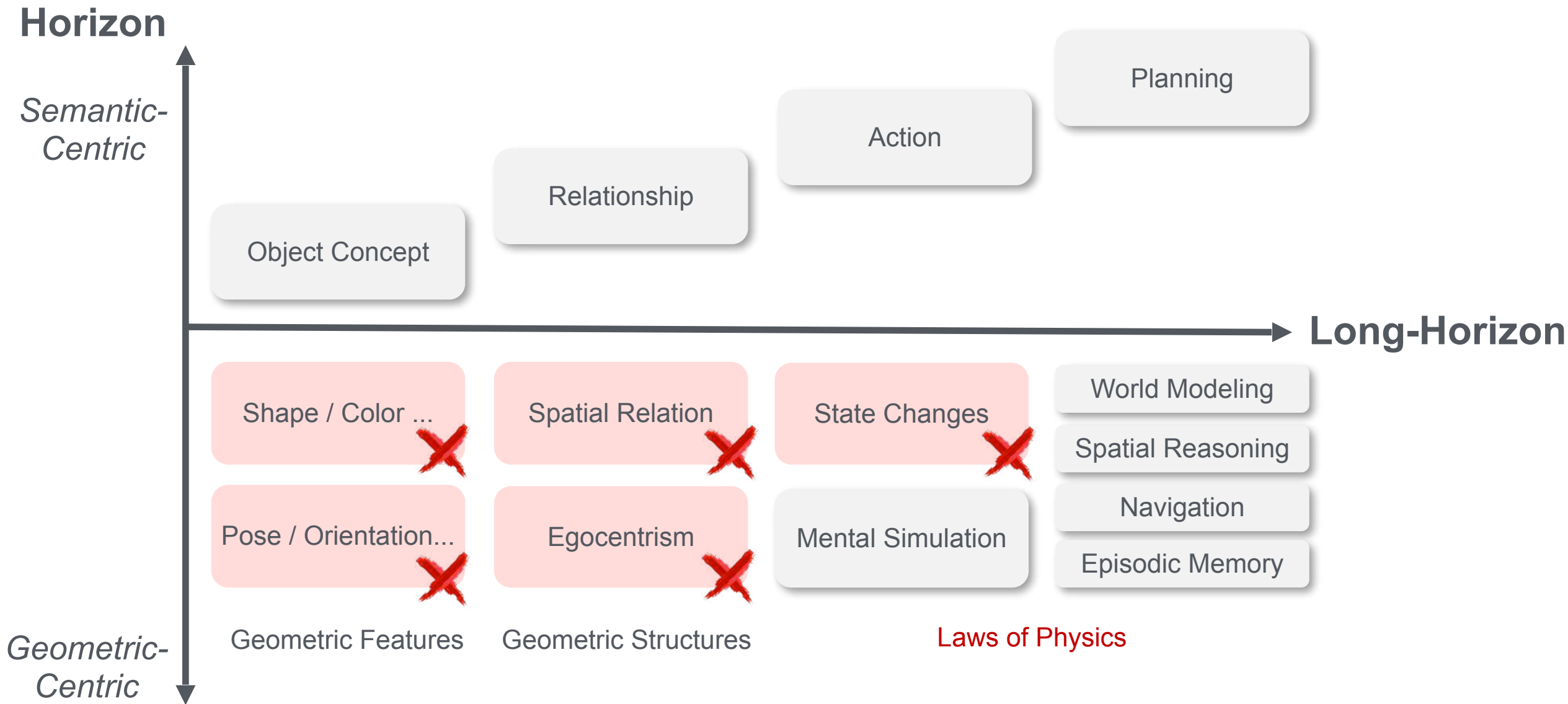
Geometric Structures

*Geometric
Centric*

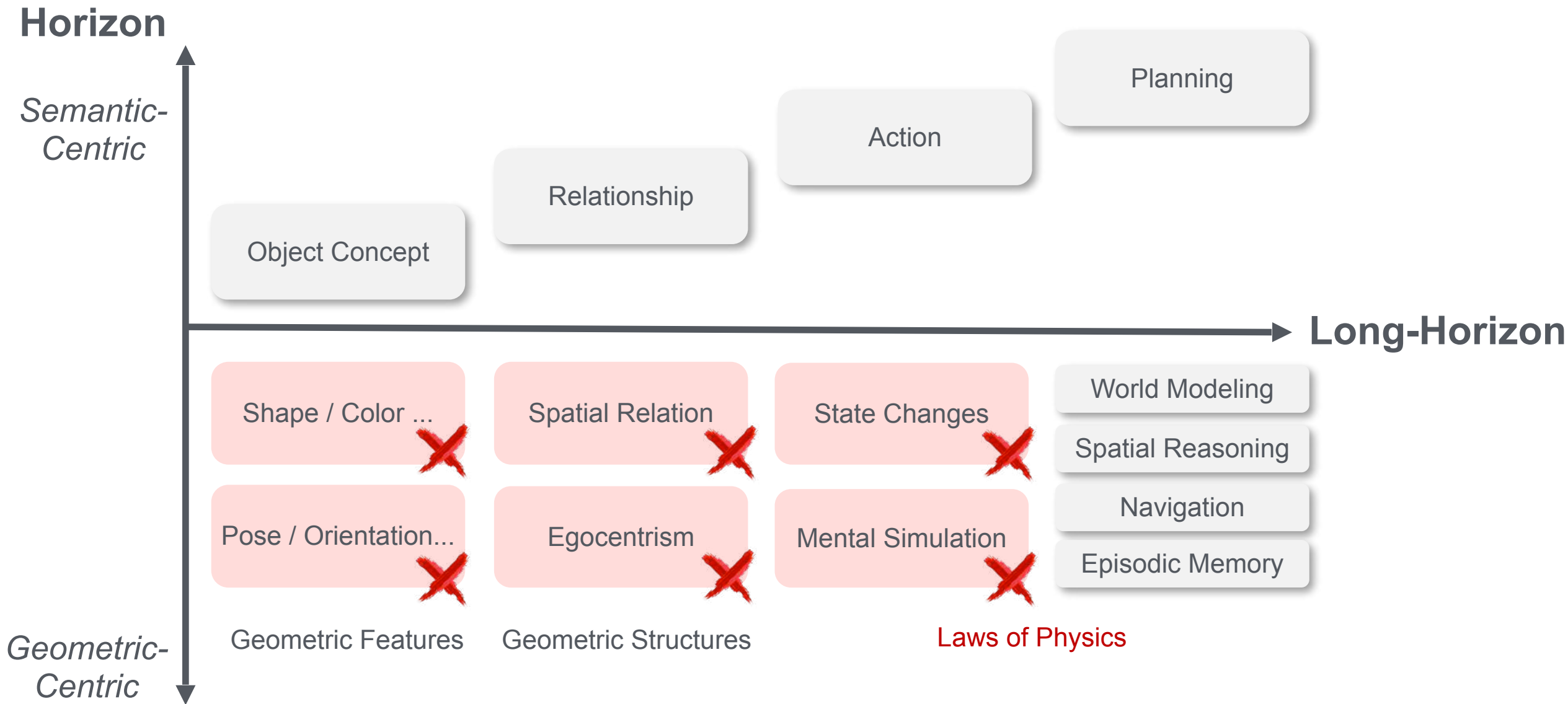
Missing knowledge about physical world



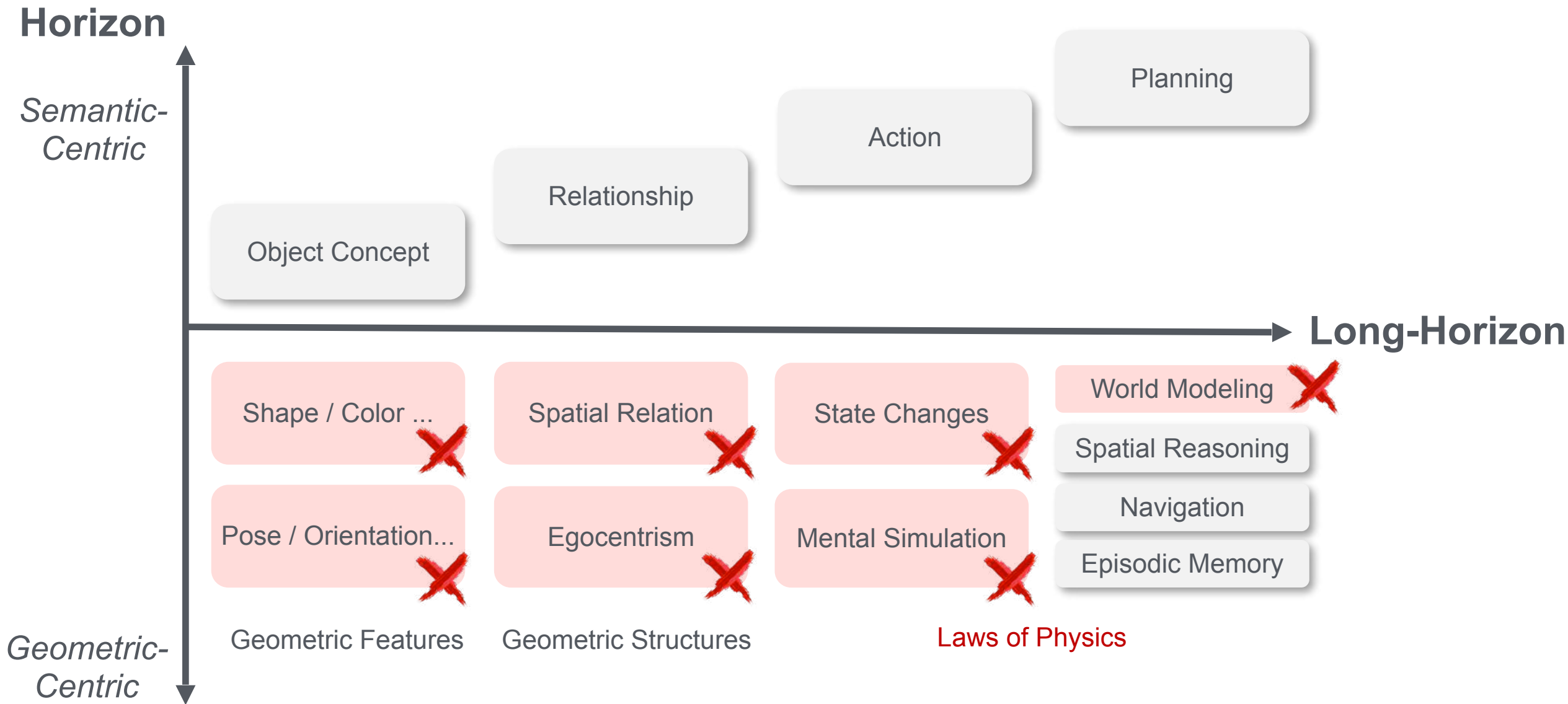
Missing knowledge about physical world



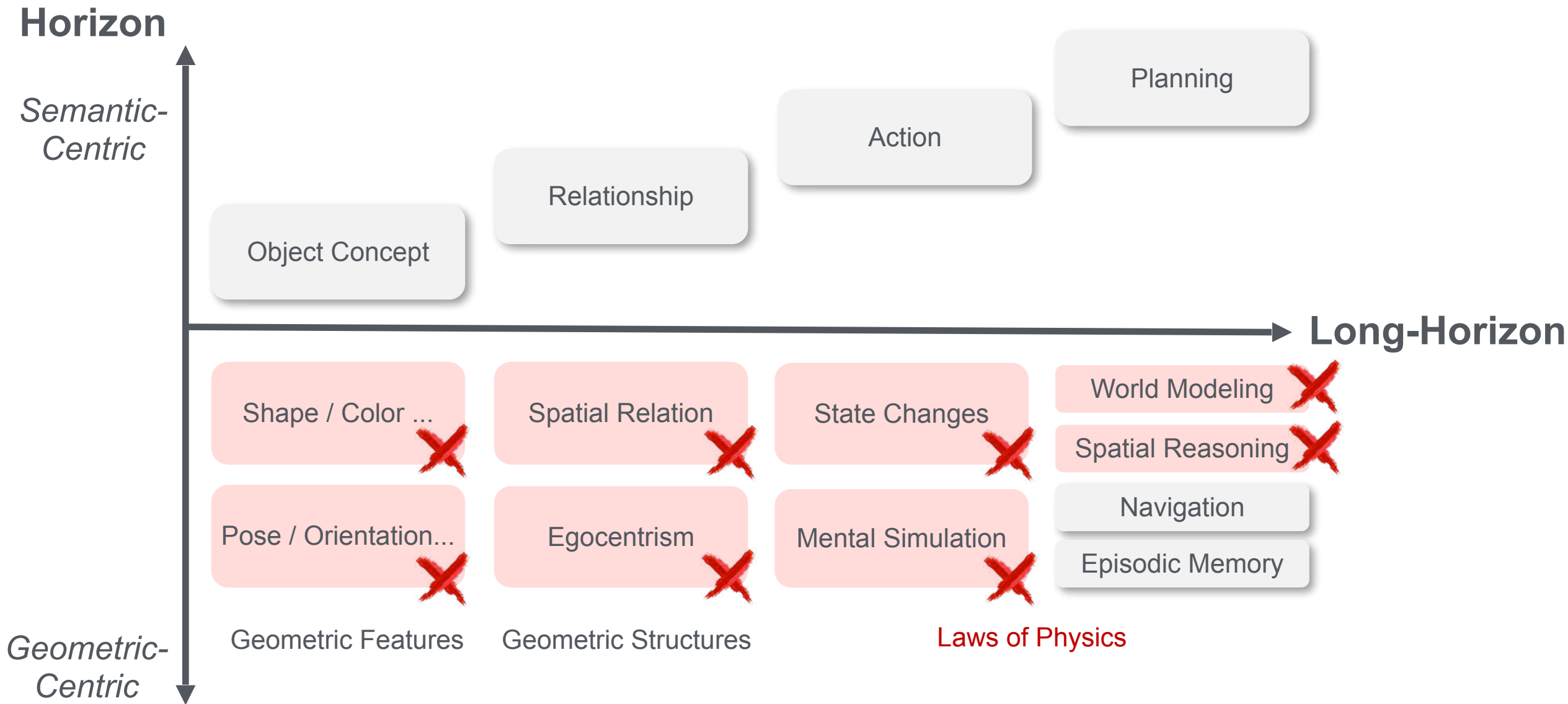
Missing knowledge about physical world



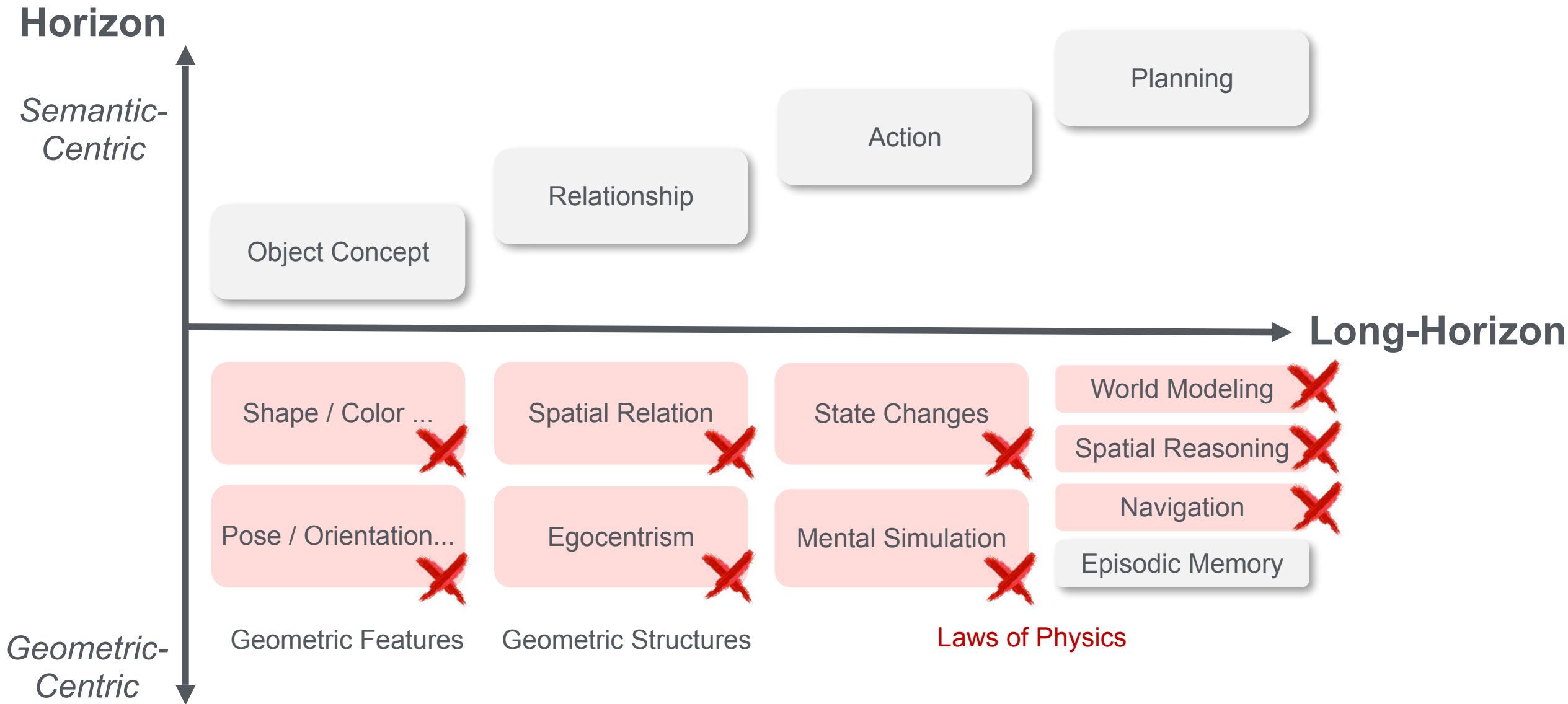
Missing knowledge about physical world



Missing knowledge about physical world



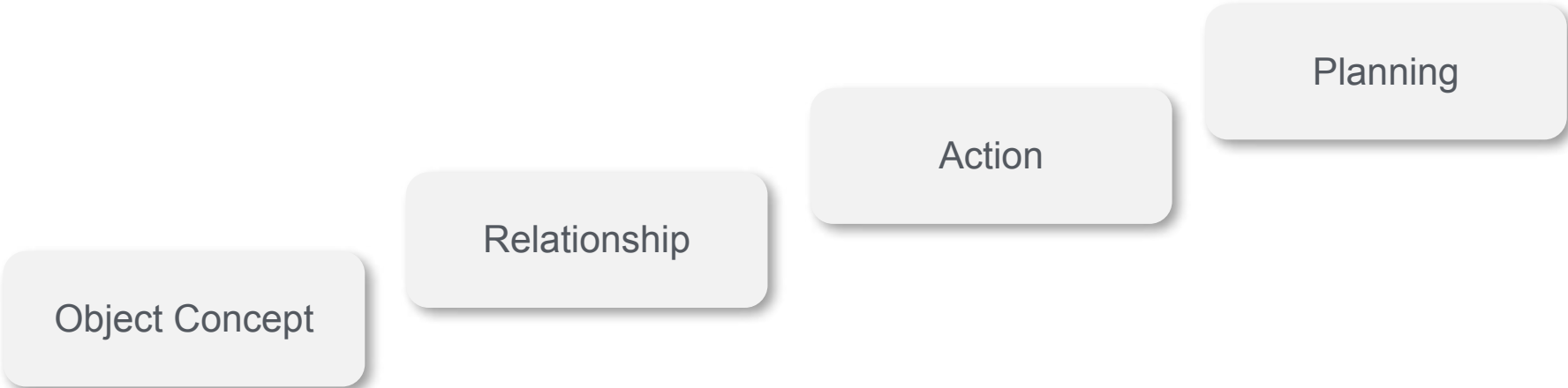
Missing knowledge about physical world



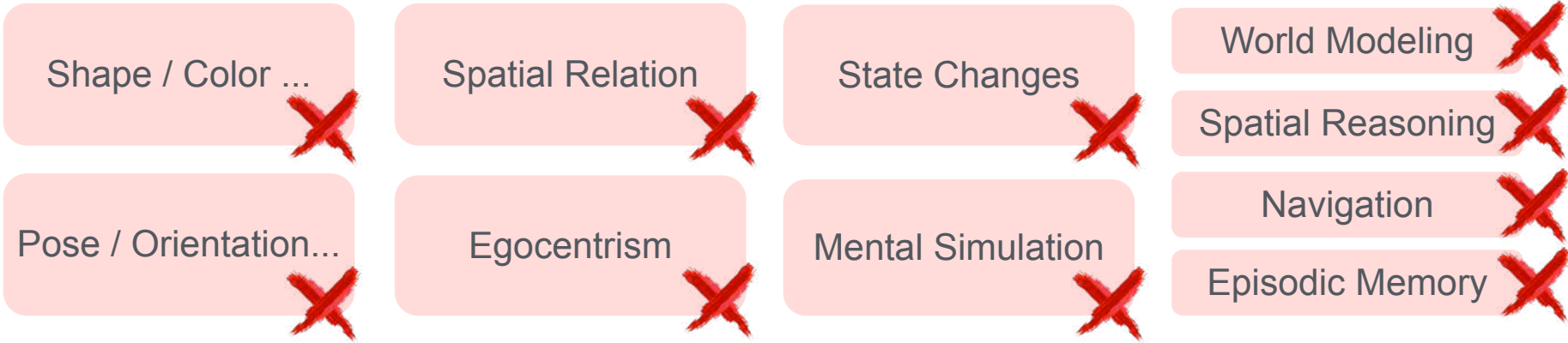
Missing knowledge about physical world

Horizon

Semantic-Centric



Long-Horizon



Geometric-Centric

Geometric Features

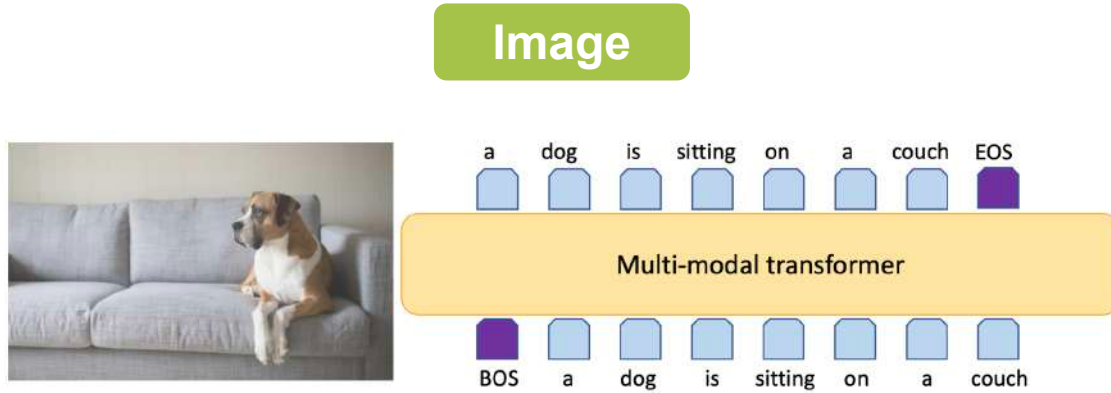
Geometric Structures

Laws of Physics

Current LMMs fall short on **Geometric Info.**

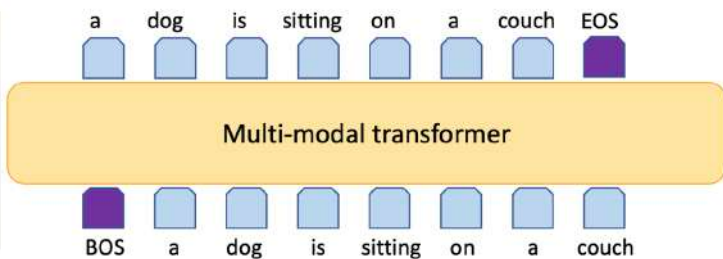
Why?

Language → Vision: **Linearize** Everything as **Sequences**



Language → Vision: **Linearize** Everything as **Sequences**

Image



Video

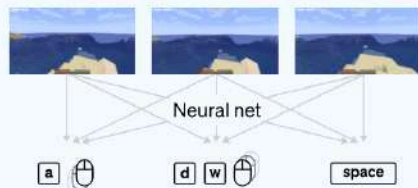
Collect Internet data

Search the web
70K hours of
unlabeled video

Train the Inverse Dynamics Model (IDM)

Contractors produce data
2K hours of video
labeled with mouse
and keyboard actions

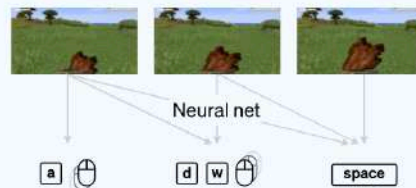
Train a model to predict actions
given past and future frames



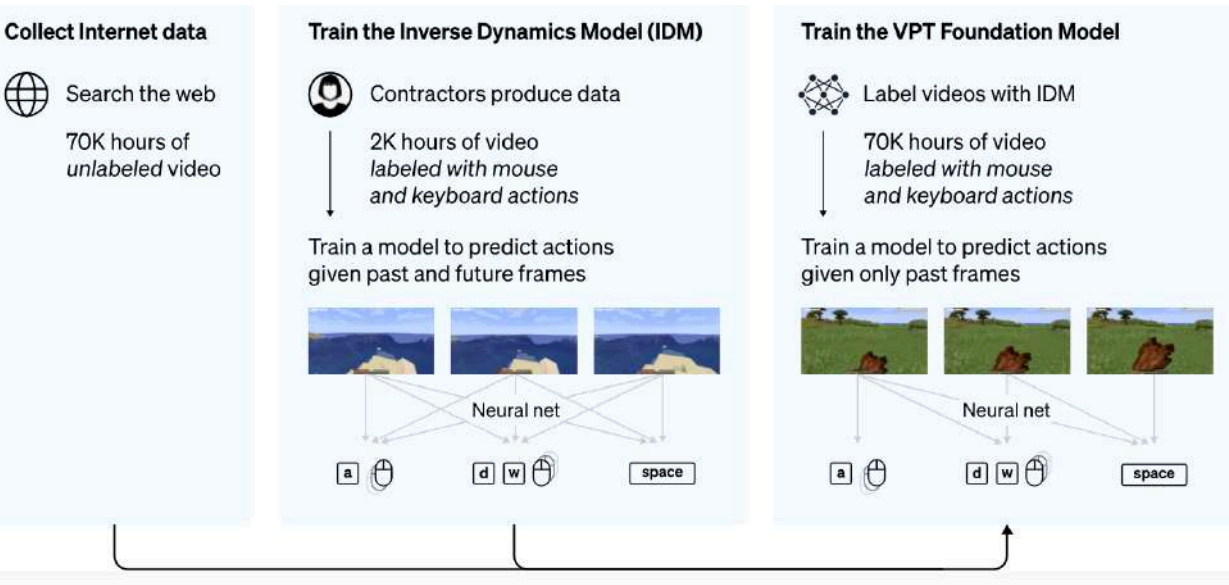
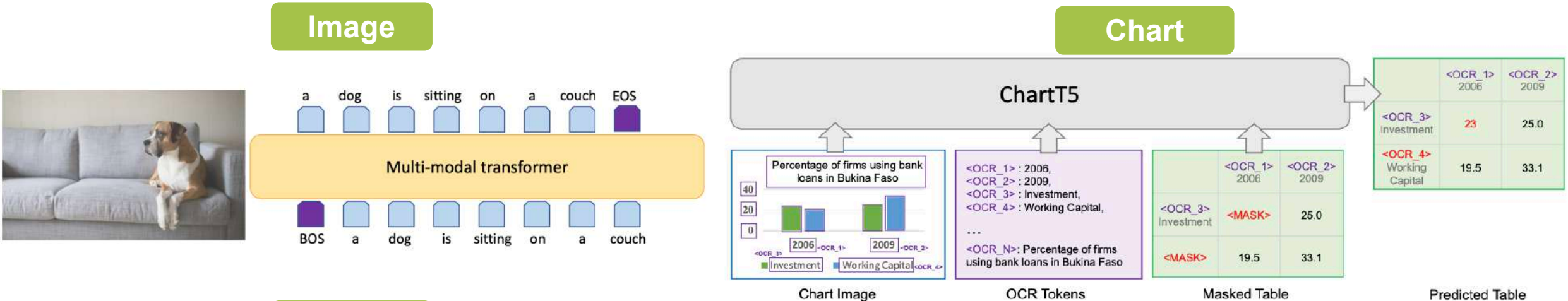
Train the VPT Foundation Model

Label videos with IDM
70K hours of video
labeled with mouse
and keyboard actions

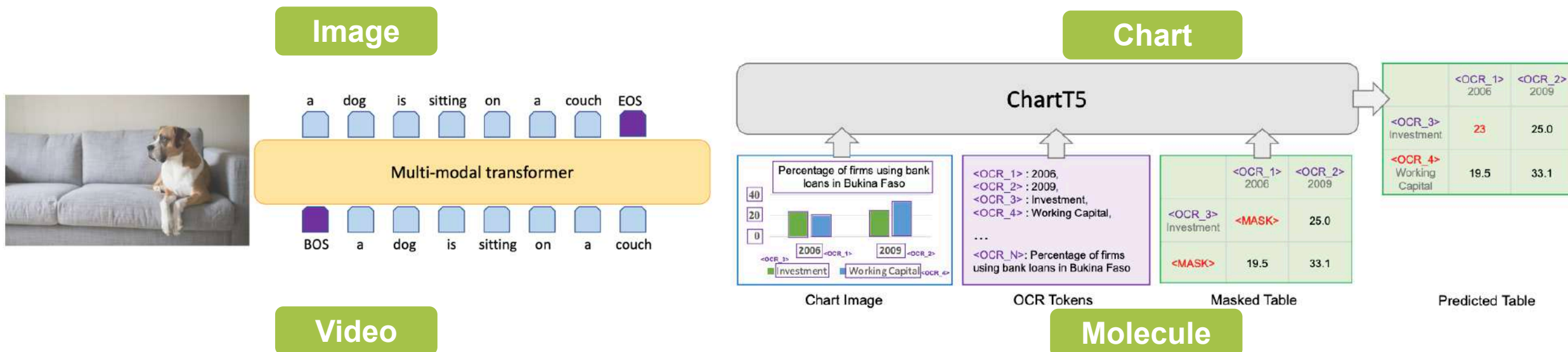
Train a model to predict actions
given only past frames



Language → Vision: Linearize Everything as Sequences

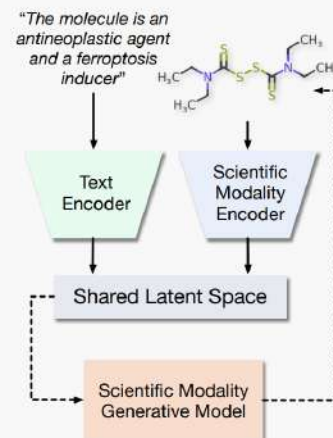


Language → Vision: Linearize Everything as Sequences

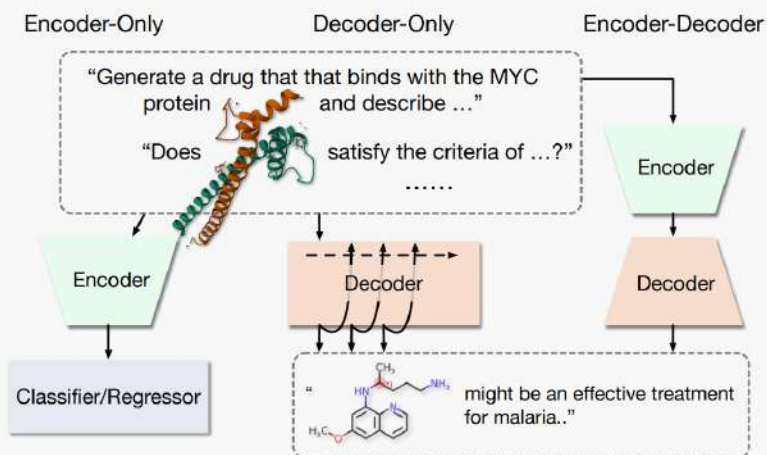


Multimodal LLMs for Science

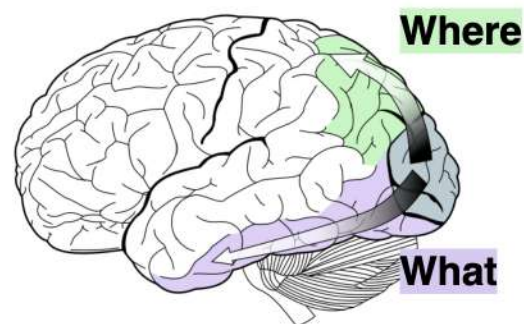
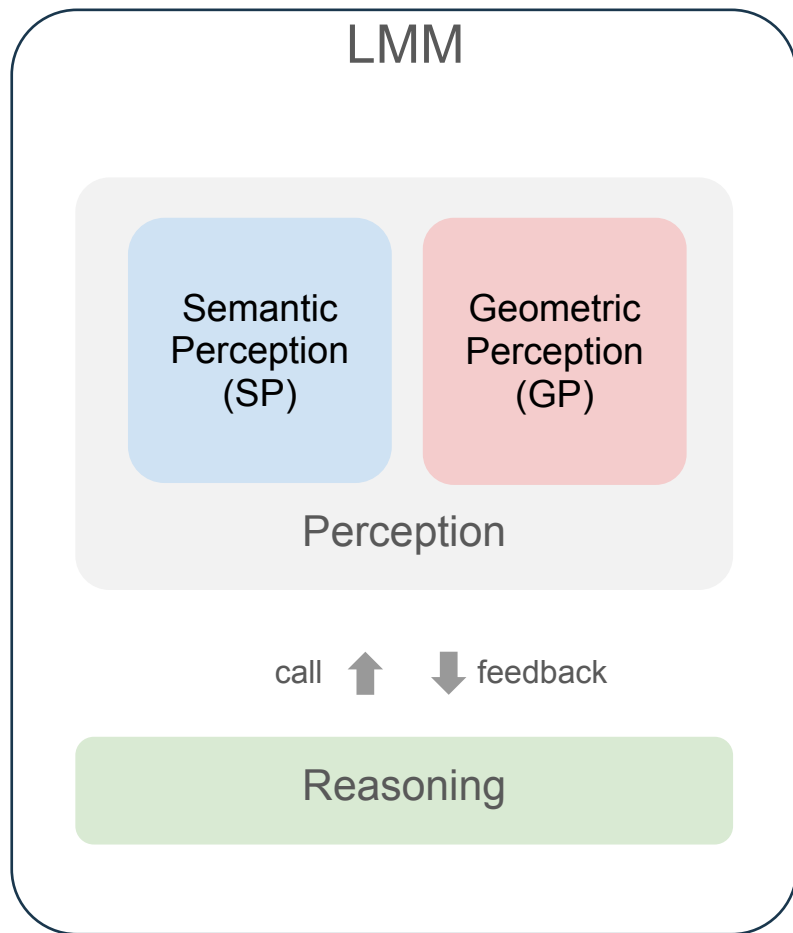
Bi-Encoder Models



Joint-Representation Models



Semantic Alignment vs Geometric Alignment

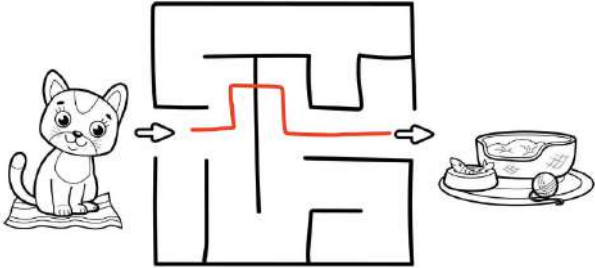
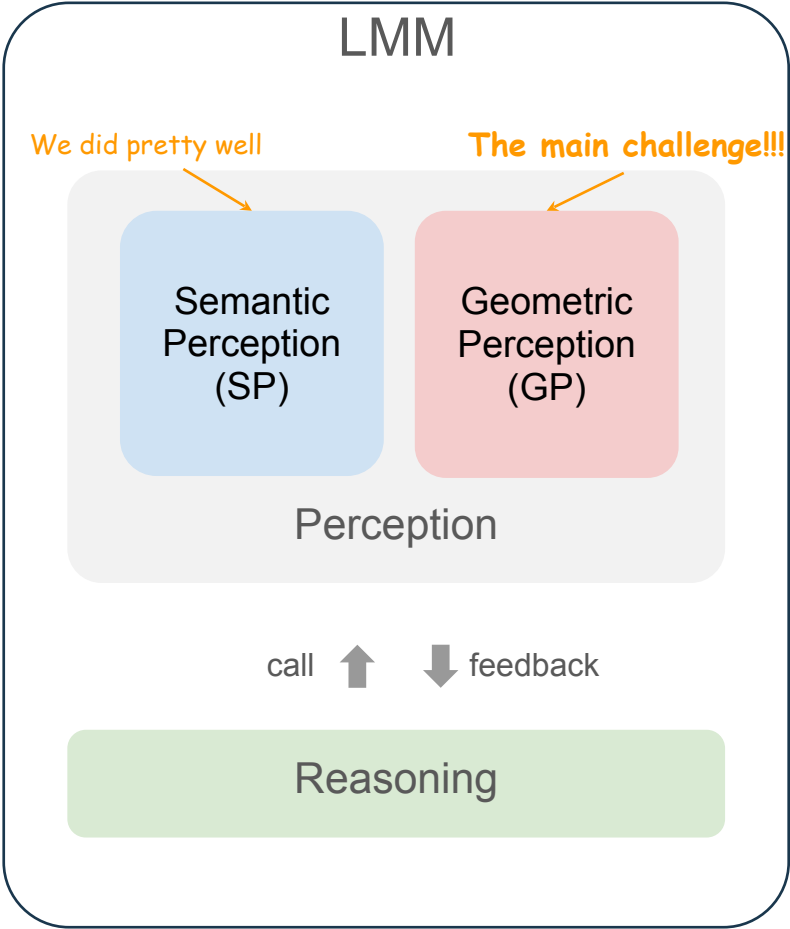


[Two-streams Hypothesis](#) an interesting human cognition analogy:

The [ventral stream](#) (or, "what pathway") leads to the temporal lobe, which is involved with object and visual identification and [recognition](#).

The [dorsal stream](#) (or, "where pathway") leads to the parietal lobe, which is involved with processing the object's spatial location relative to the viewer and with speech repetition.

Semantic Alignment vs Geometric Alignment



Semantic-centric Question:

Is there a dog or a cat in the image?

Expected response:

[R:] Find the "Cat" in the image. → [SP:] Yes there is a cat in the image

[Answer:] there is a cat in the image; there are no dogs in the image.

Geometric-centric Question:

Given that the black lines are walls that cannot be crossed, is the red line a valid path through the maze?

Expected response:

[R:] Does the red line intersect with any black lines? →

[GP:] The red line crosses a vertical black line in the middle.

[Answer:] No, the red line is not a valid path in the maze because it intersects with walls.

Why geometric alignment is bad:

VL Encoders < V-only Encoders

V-only encoder (**MAE**, **SAM**...):

semantic << geometric

VL encoder (**CLIP**...):

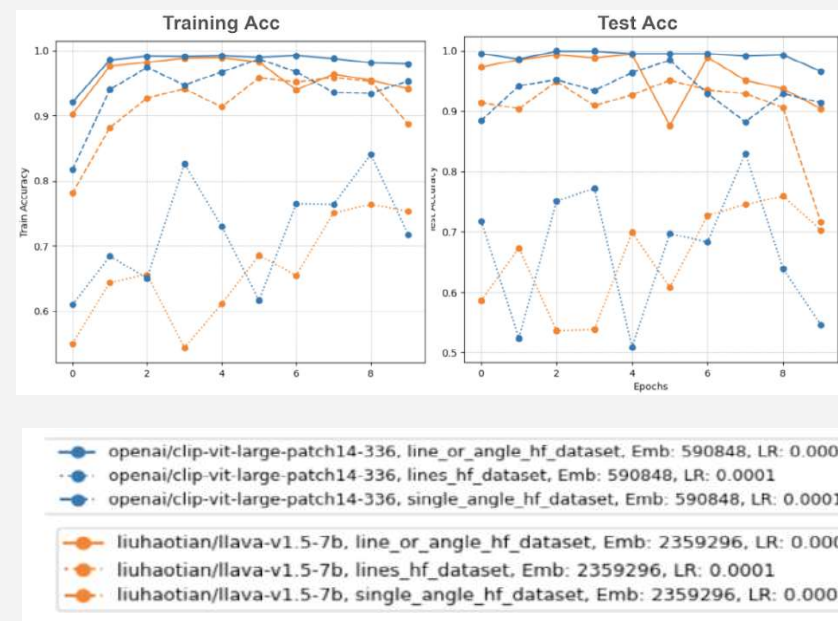
semantic >> geometric

LLaVA (= CLIP+LLM layer):

semantic >> geometric

LLaVA < CLIP

LLM layer **swallows** the geometric features.



Why Is Spatial Reasoning Hard in VLMs?

Let's open up VLMs!



Shiqi Chen



Tongyao Zhu



Ruochen Zhou



Jinghan Zhang



Siyang Gao



Juan Carlos Niebles



Mor Geva



Junxian He



Jiajun Wu



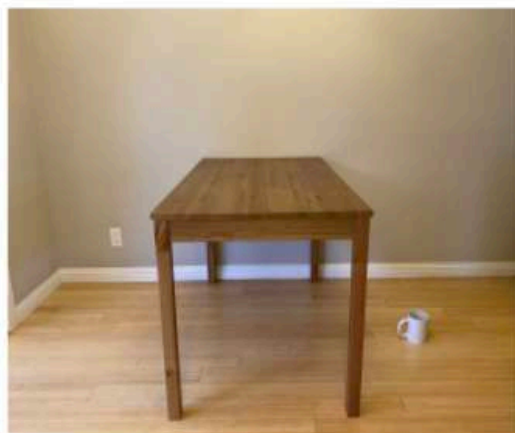
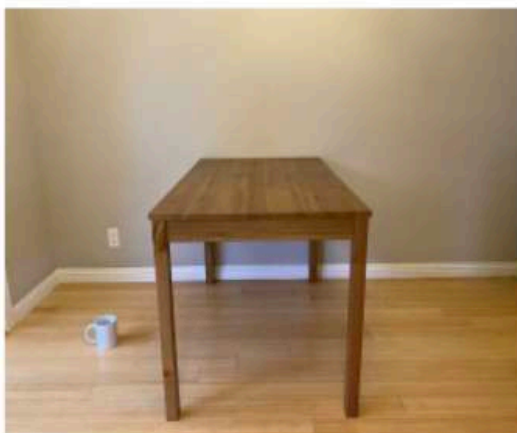
Manling Li

Recap



Cup is on top of the table.

Cup is on top of the table.



Cup is on top of the table.

Cup is on top of the table.



Model	Whats- Up	COCO- spatial	GQA- spatial	Avg
CLIP ViT-B/32	31.0	47.4	46.9	41.8
CLIP ViT-L/14	26.1	49.5	47.3	41.0
NegCLIP	34.4	46.9	46.0	42.4
RoBERTaCLIP	25.1	50.0	49.8	41.6
CoCa	29.4	46.7	47.1	41.0
XVLM 4M	31.5	61.7	58.7	50.6
XVLM 16M	41.9	65.0	58.2	55.0
BLIP 14M	38.5	54.0	49.8	47.5
BLIP 129M	30.4	49.3	49.0	42.9
BLIP2-ITM	37.6	53.0	49.8	46.8
BLIP2-ITC	29.0	53.7	51.0	44.6
FLAVA	30.5	52.6	51.7	44.9
CoCa-Caption	24.1	48.6	49.5	40.8
XVLM-Flickr30K	44.3	65.2	61.4	56.9
XVLM-COCO	42.1	71.0	68.1	60.4
BLIP-Flickr30K	33.8	54.2	48.9	45.6
BLIP-COCO	32.8	51.4	51.4	45.2
BLIP-VQA	47.8	62.0	58.4	56.0
Random / Text-only	25.0	50.0	50.0	41.7
Human Estimate	100.0	97.3	99.0	98.8

What's Up Benchmark

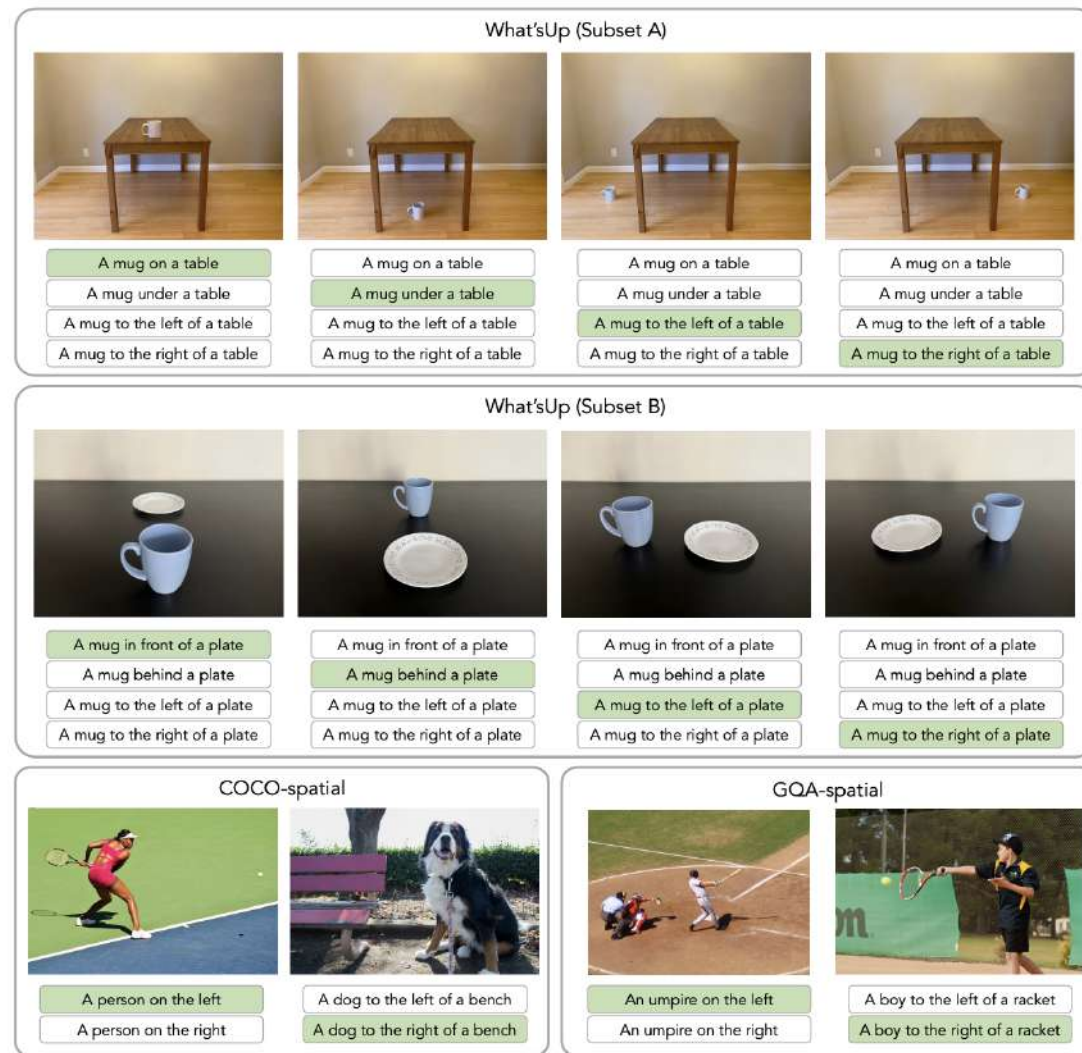
Before that, let's first introduce the setting:

Focus on **spatial reasoning problems**

Controlled_image (820 images)

Coco (2687 images)

GQA (1451 images)



Controlled_A is special

One big object + one small object

An example:

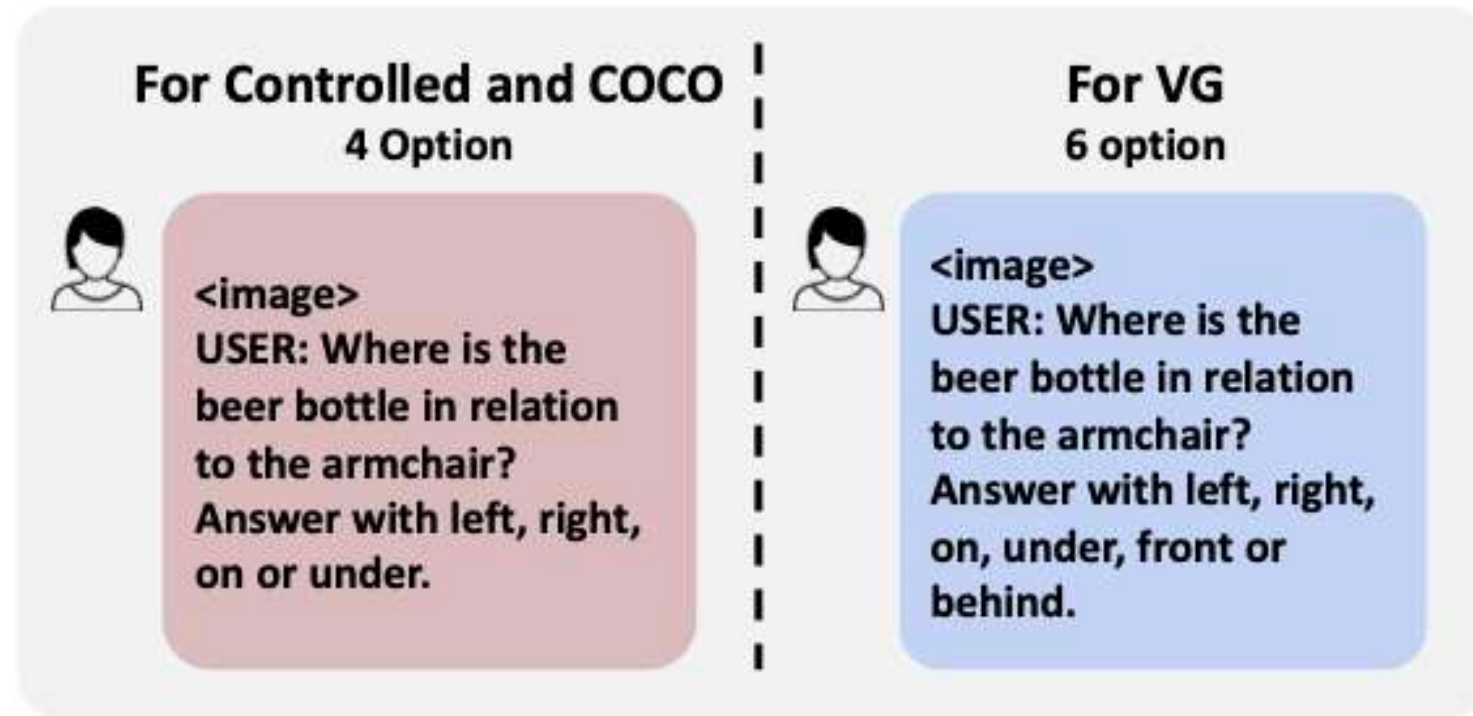
USER:<image>

ASSISTANT:

Generation: The image features a blue chair with a knitted ball sitting **on** top of it. The chair is positioned in a room with a white wall, and there is a wooden floor beneath it. The room appears to be empty, with no other furniture or objects visible. The blue chair and the knitted ball create a cozy and inviting atmosphere in the space.



Tasks on Spatial Reasoning



Spatial Reasoning is a **bottleneck** for current VLMs

Where is the book in relation to the candle?



VLM
(e.g. LLaVA)

Left



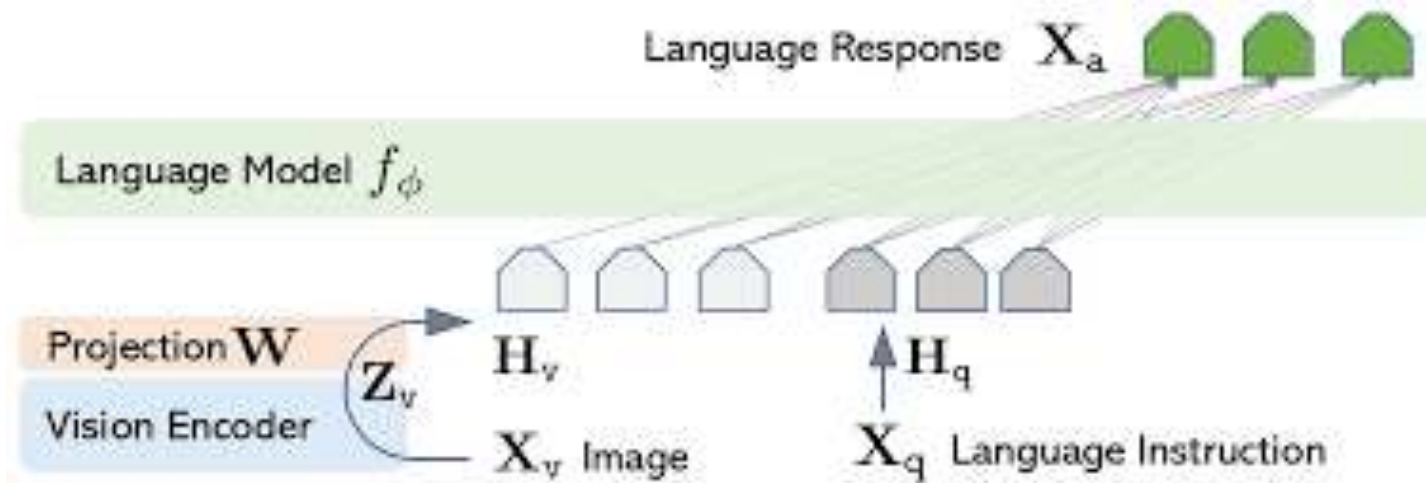
Where is the cat in relation to the person?

Above

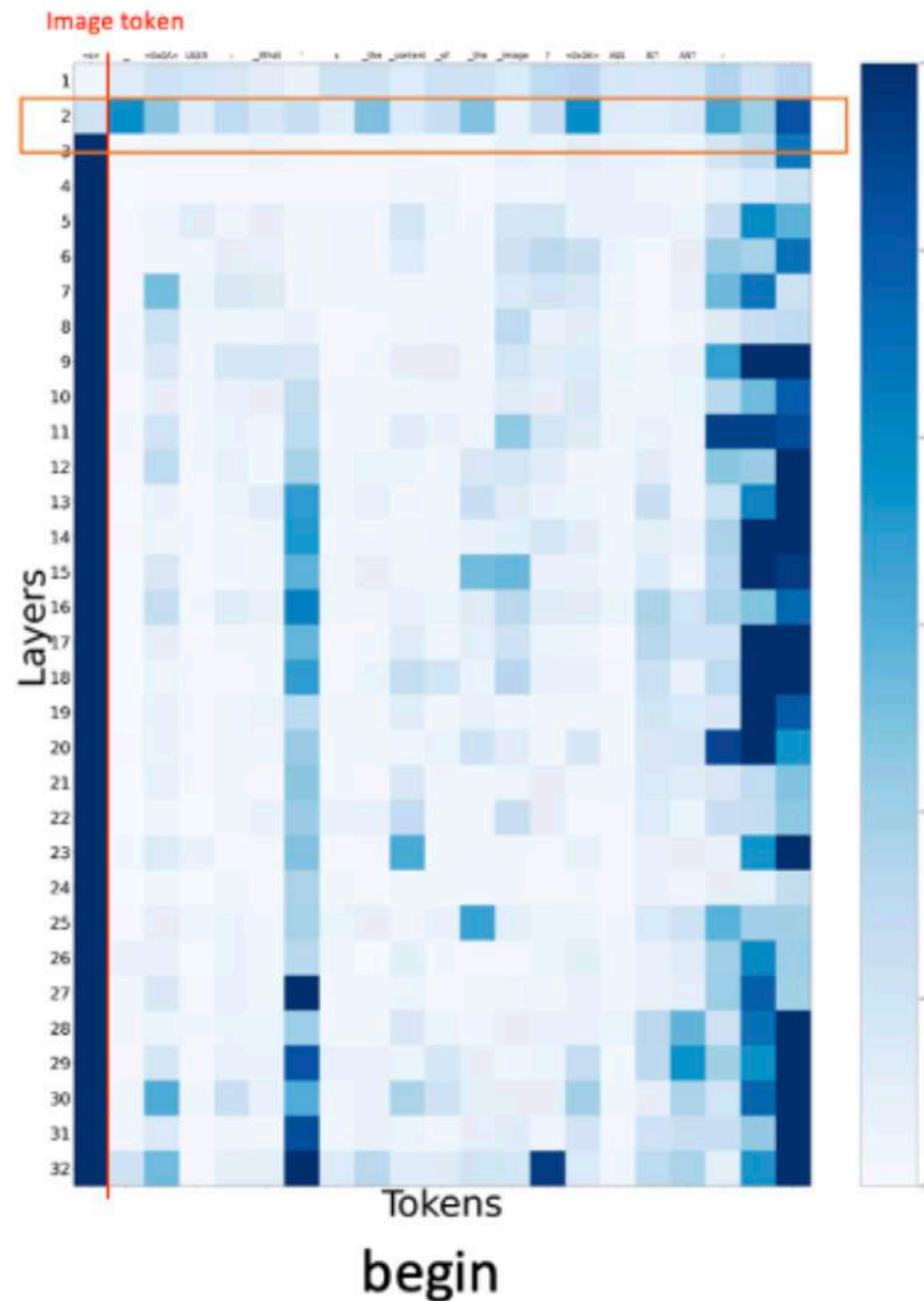
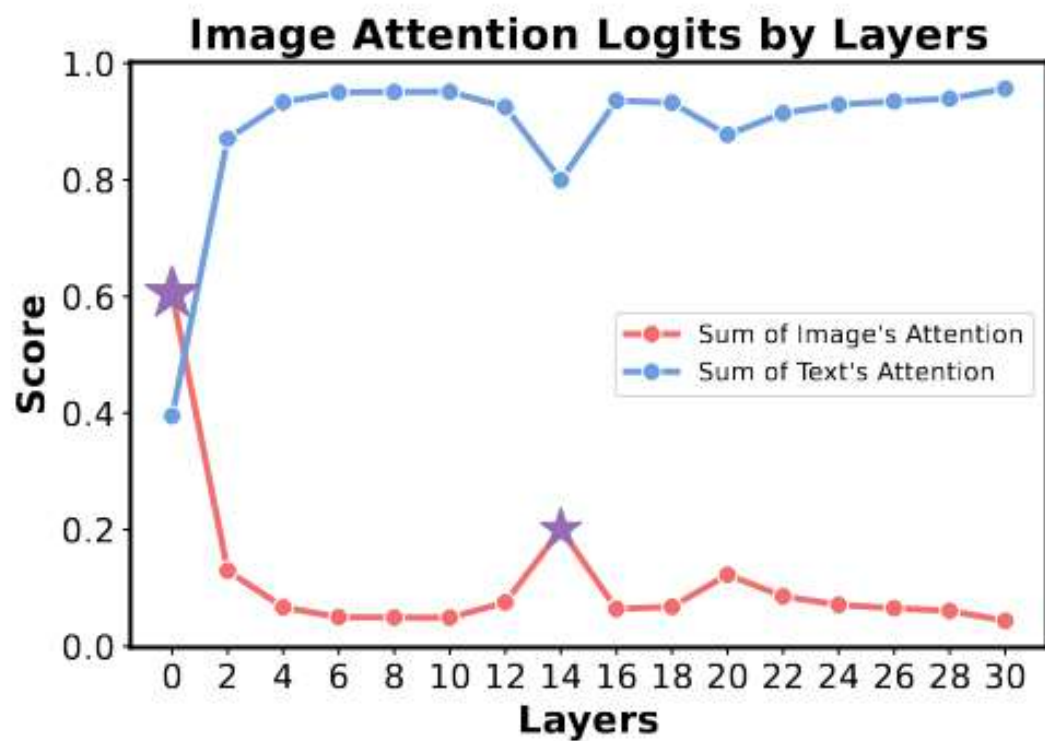


Models

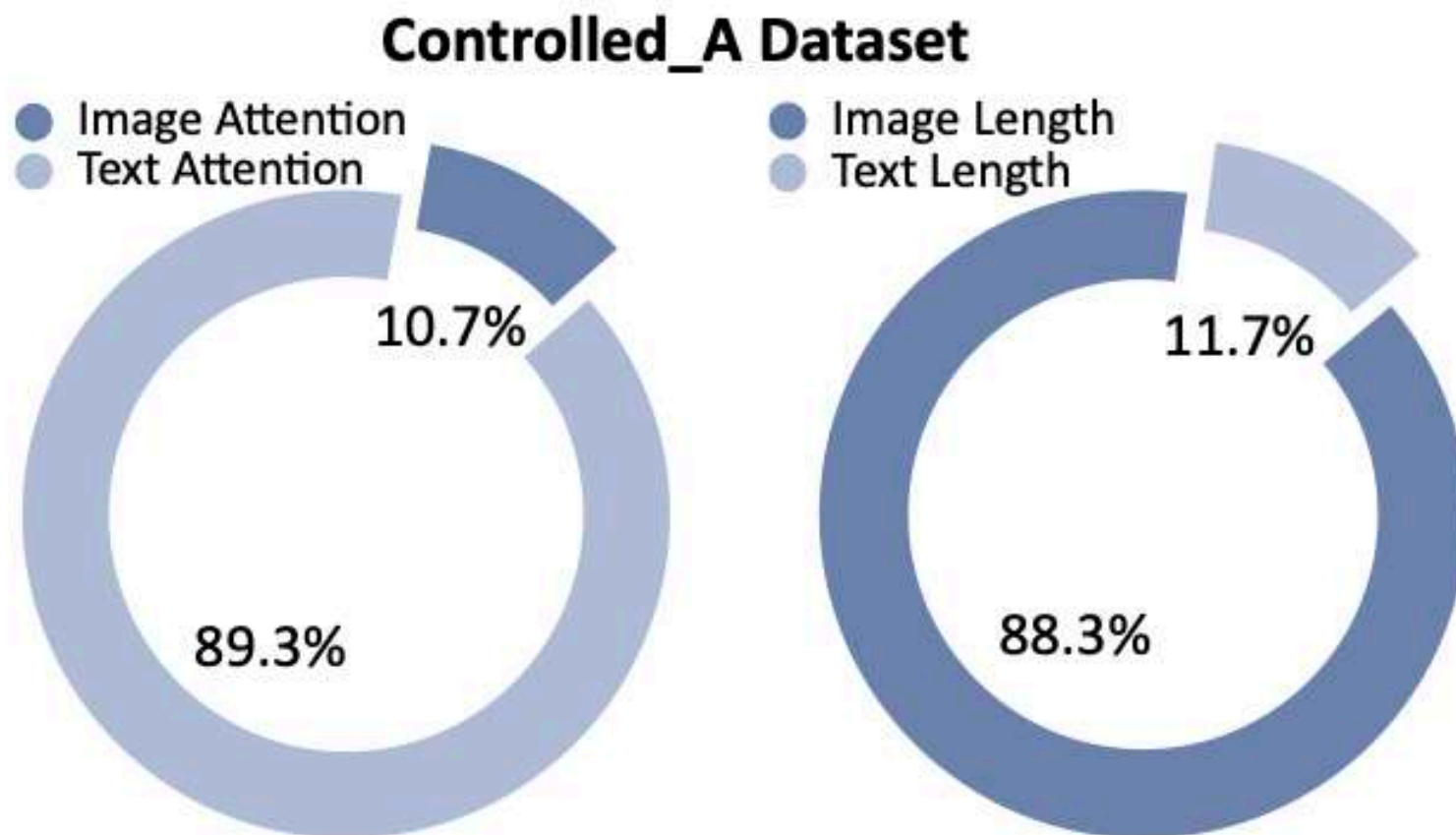
We use LLaVa 7B in most experiments:
a CLIP, a projector and a language model



Open Up the Model

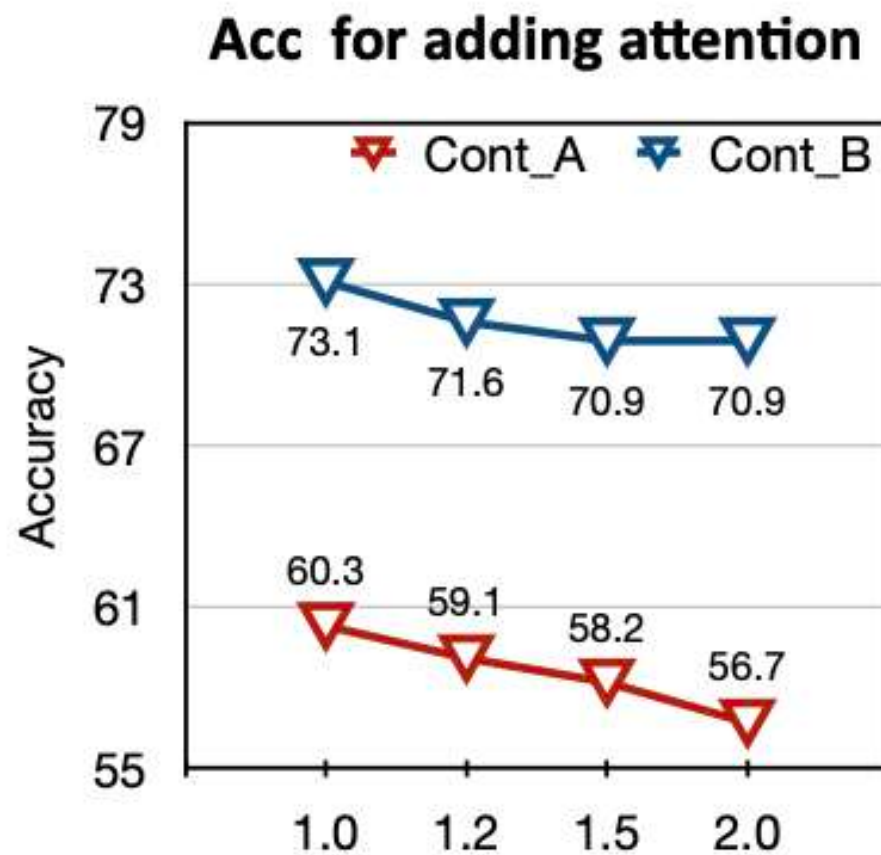


The sparsity of image's attention scores



Let us simply add more attention

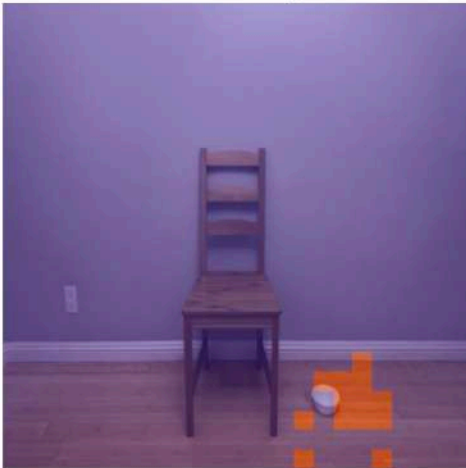
Simply add attention to vision → **not work**



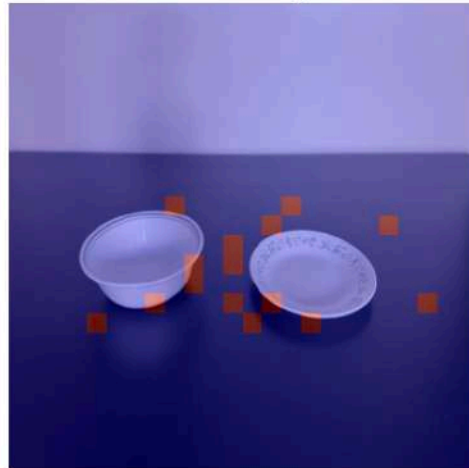
Dive into the visual pattern

Finding 1: The model focuses on the relevant entity when correctly answering questions

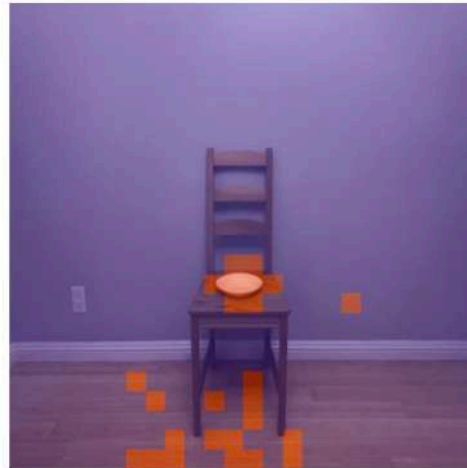
Where is the bowl ?
Golden: Right
Model: Right



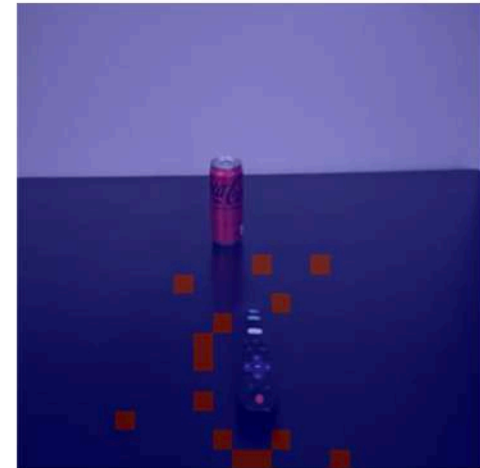
Where is the plate ?
Golden: Right
Model: Right



Where is the plate ?
Golden: On
Model: Under



Where is the can ?
Golden: Behind
Model: Front

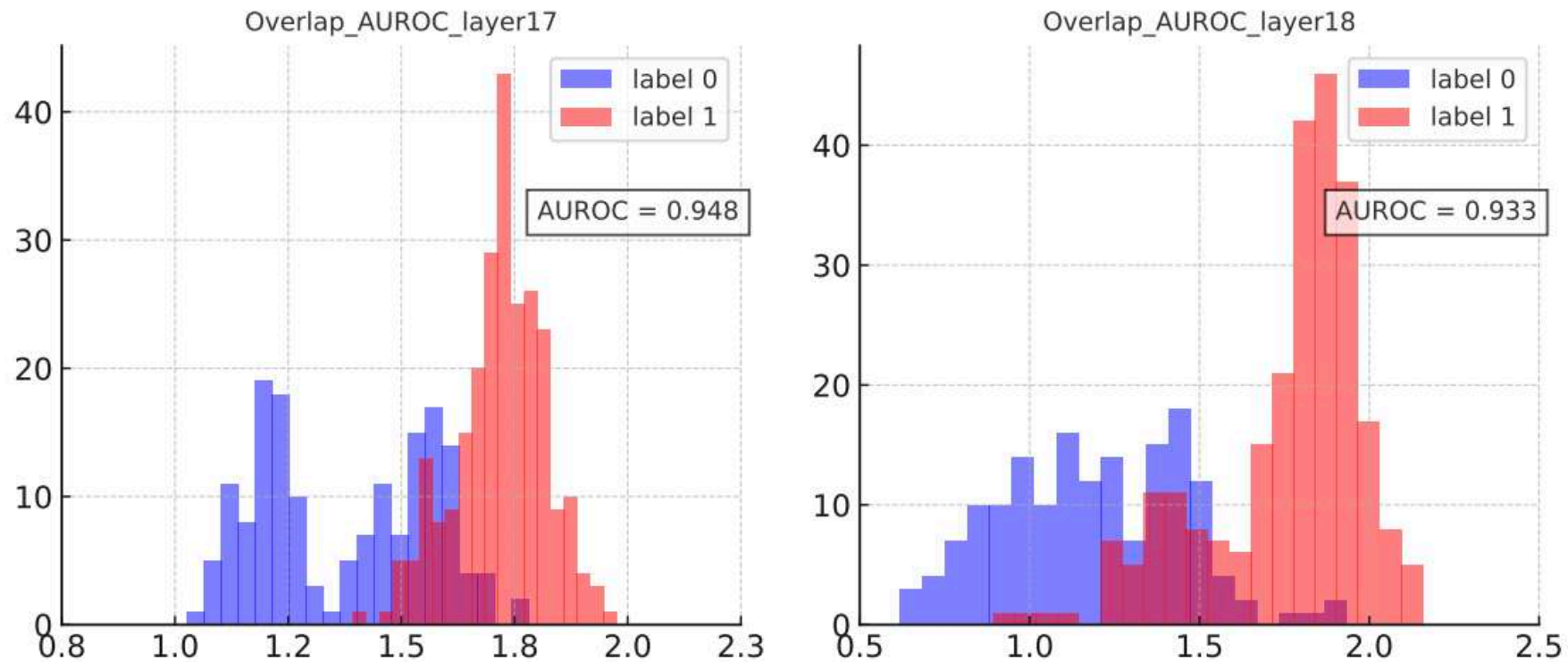


Correct Answer
Correct Attention

Incorrect Answer
Incorrect Attention

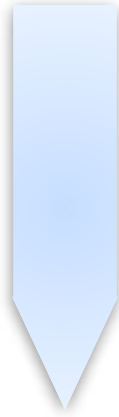
Dive into the visual pattern

Finding 1: The model focuses on the relevant entity when correctly answering questions



AUROC of the overlap between YOLO annotation and attention

Seeing more



Seeing more on "right" part

Distribution is the key!

VLM would **SEE** 🔍 the **wrong place** or **missing something!**



Where is the cat in relation to the person?

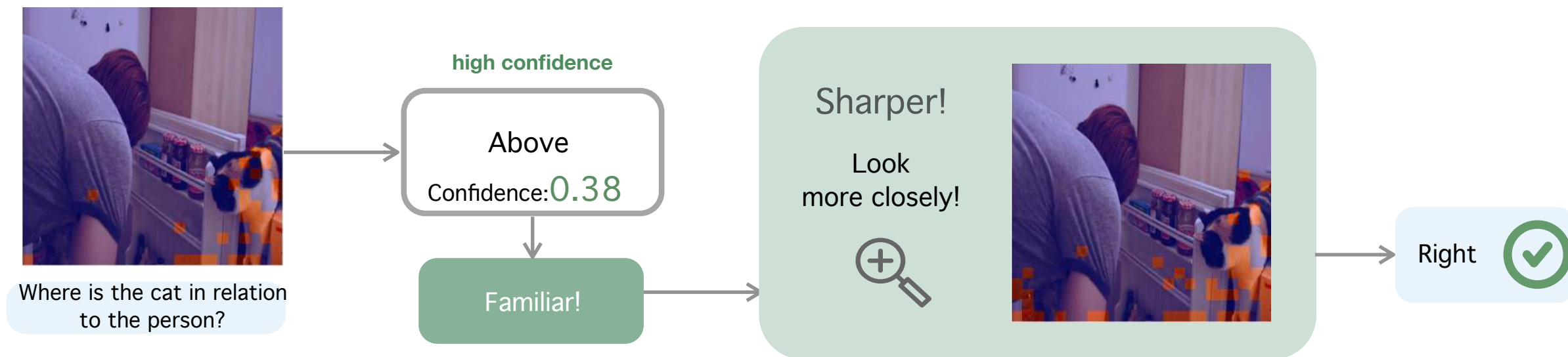
VLM
(e.g. LLaVA)



Attention on image

✗ **Insufficient focus!**

We intervene attention **adaptively** with **model's self-confidence**!

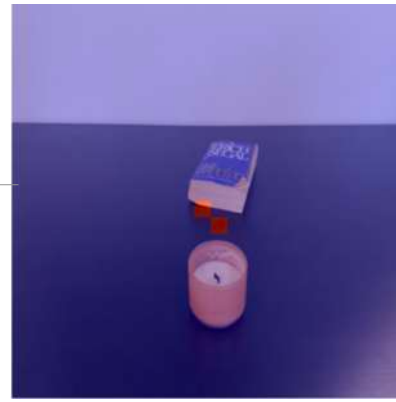


How about the incorrect part?

We open up the model and find that
VLM would **SEE** 🔍 the **wrong place** or **missing something**!



VLM
(e.g. LLaVA)



✗ Incorrect focus!

Where is the book in
relation to the candle?



Attention on image

We intervene attention **adaptively** with **model's self-confidence!**

Where is the book in relation to the candle?



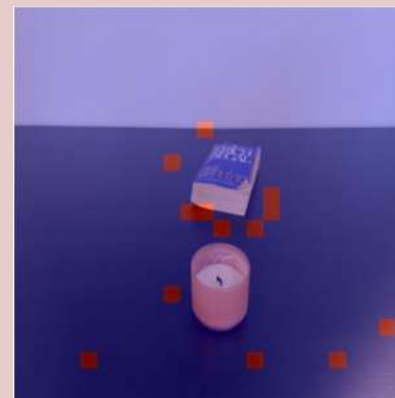
Unfamiliar!

Left
Confidence: **0.27**

low confidence

Smoother!

Look at
other patches!



Behind



We intervene attention **adaptively** with **model's self-confidence!**

Where is the book in relation to the candle?



Unfamiliar!

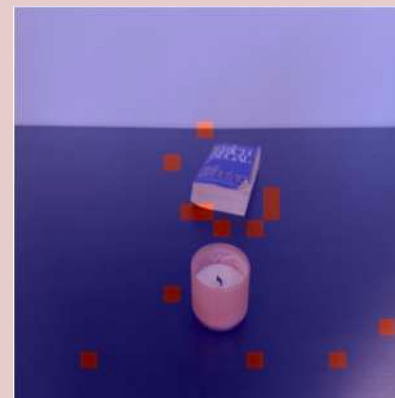
Left

Confidence: **0.27**

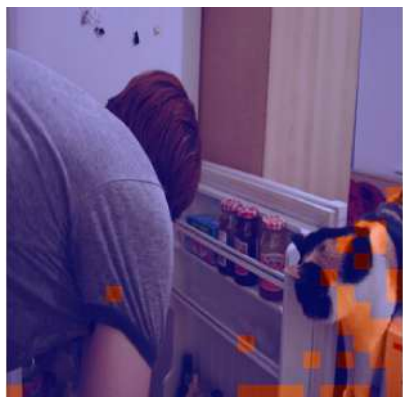
low confidence

Smoother!

Look at other patches!



Behind



Where is the cat in relation to the person?

high confidence

Above

Confidence: **0.38**

Familiar!

Sharper!

Look more closely!



Right



When sharpen? When smoothen?

When sharpen? When smoothen?

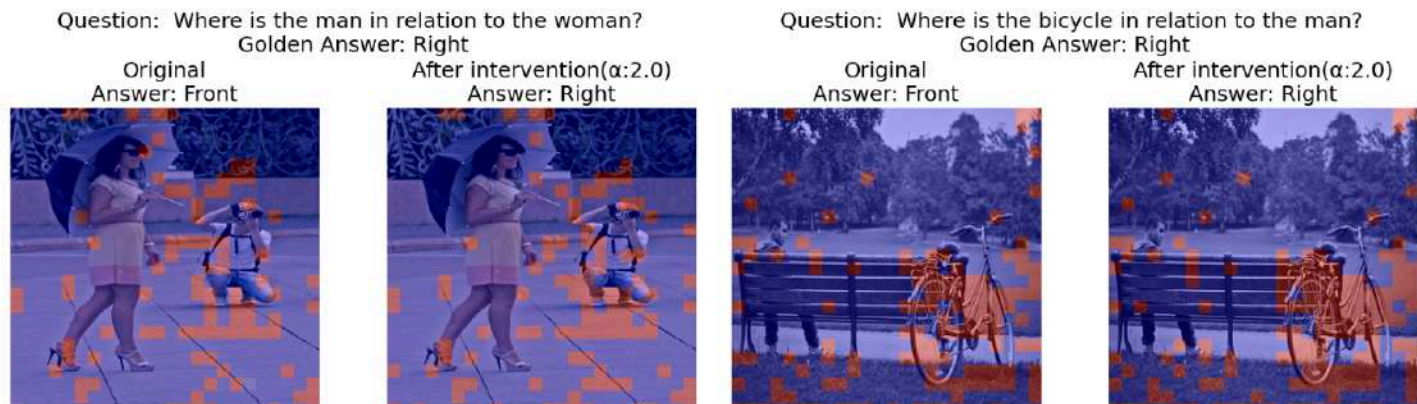
Same for the dataset  ScalingVis

From Validation Set: Scale the attention (ScalingVis)

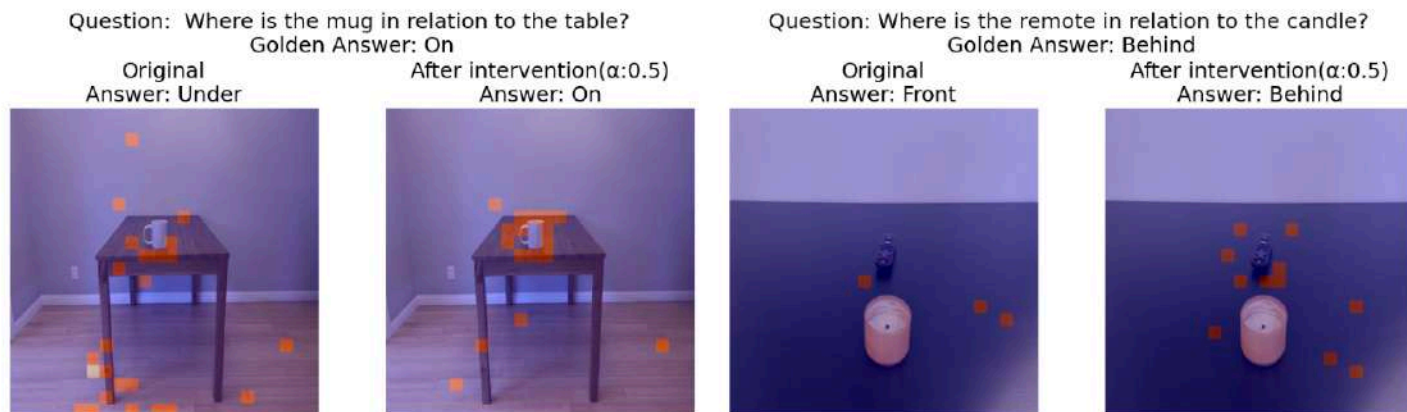
ScalingVis: self-aware to sharpen/smooth the attention pattern

- Change the **temperature** (t) in **logit** space in **all** layers (**multiplication operation** in **logit** space).
- Similar with temperature in SoftMax!

$t < 1$:



$t > 1$:



When sharpen? When smoothen?

Adapt to each sample  AdaptVis

Adaptively: Scale the attention (AdaptVis)

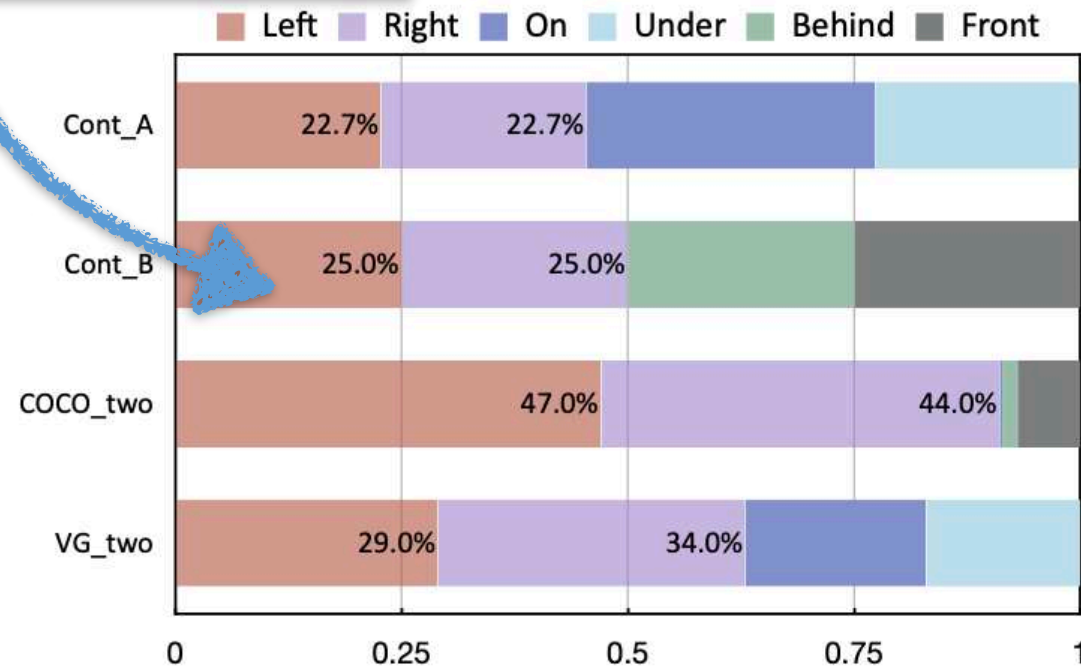
When can we trust a model's attention pattern?

→ Use uncertainty score.

Model Confidence

Familiar: Left / Right

Unfamiliar: Up/Under/Behind/Front

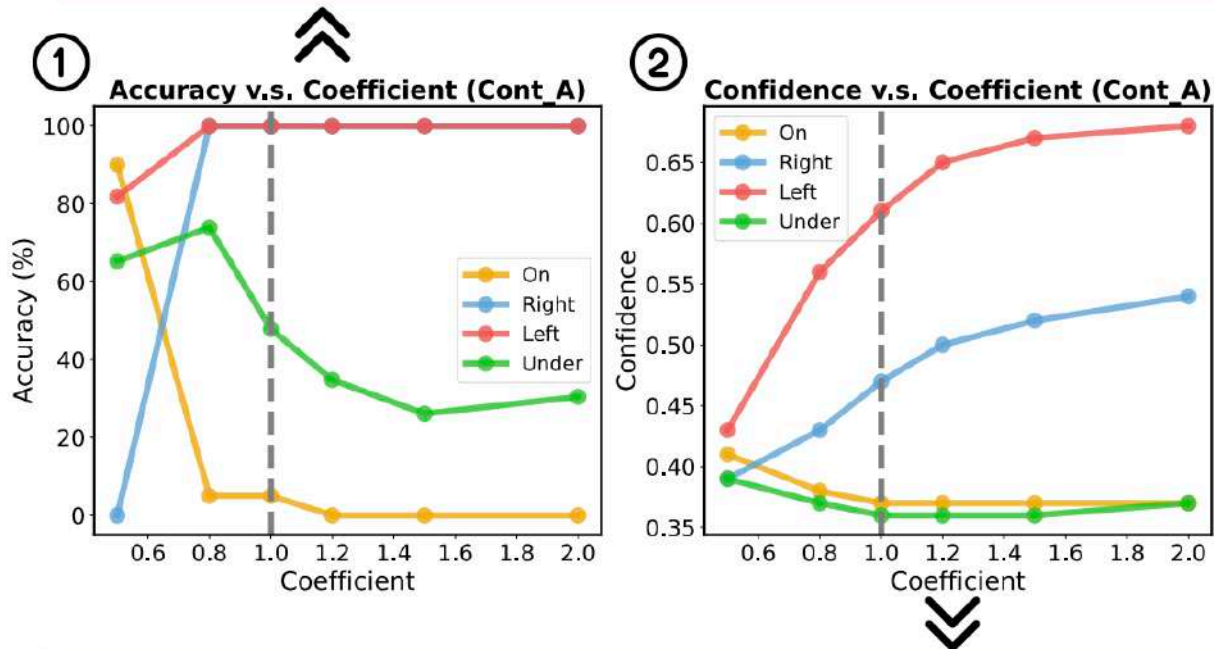


Adaptively: Scale the attention (AdaptVis)

AdaptVis: When can we trust a model's attention pattern?

- Use uncertainty score.
- Model is more confident with the familiar relationships

For low-confidence relationships: coefficient <1 improves performance. For high-confidence relationships: coefficient >1 improves performance.



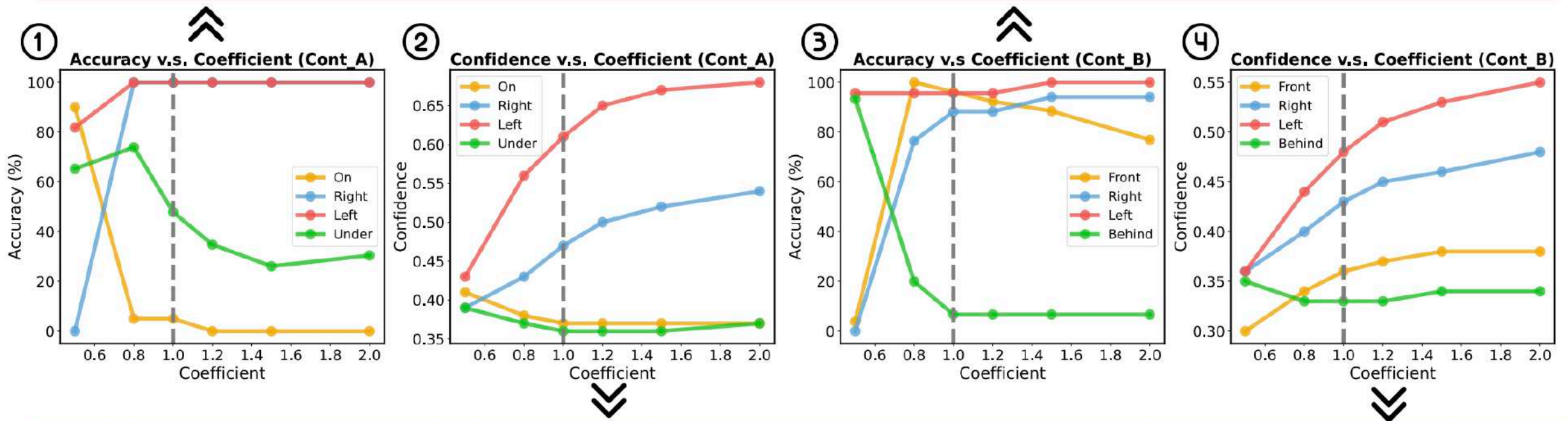
Model has **higher** confidence for **left** / **right** than **on** / **under** / **front** / **behind**, indicating that the model is more familiar with certain relationships.

Adaptively: Scale the attention (AdaptVis)

AdaptVis: When can we trust a model's attention pattern?

- Use uncertainty score.
- Model is more confident with the familiar relationships

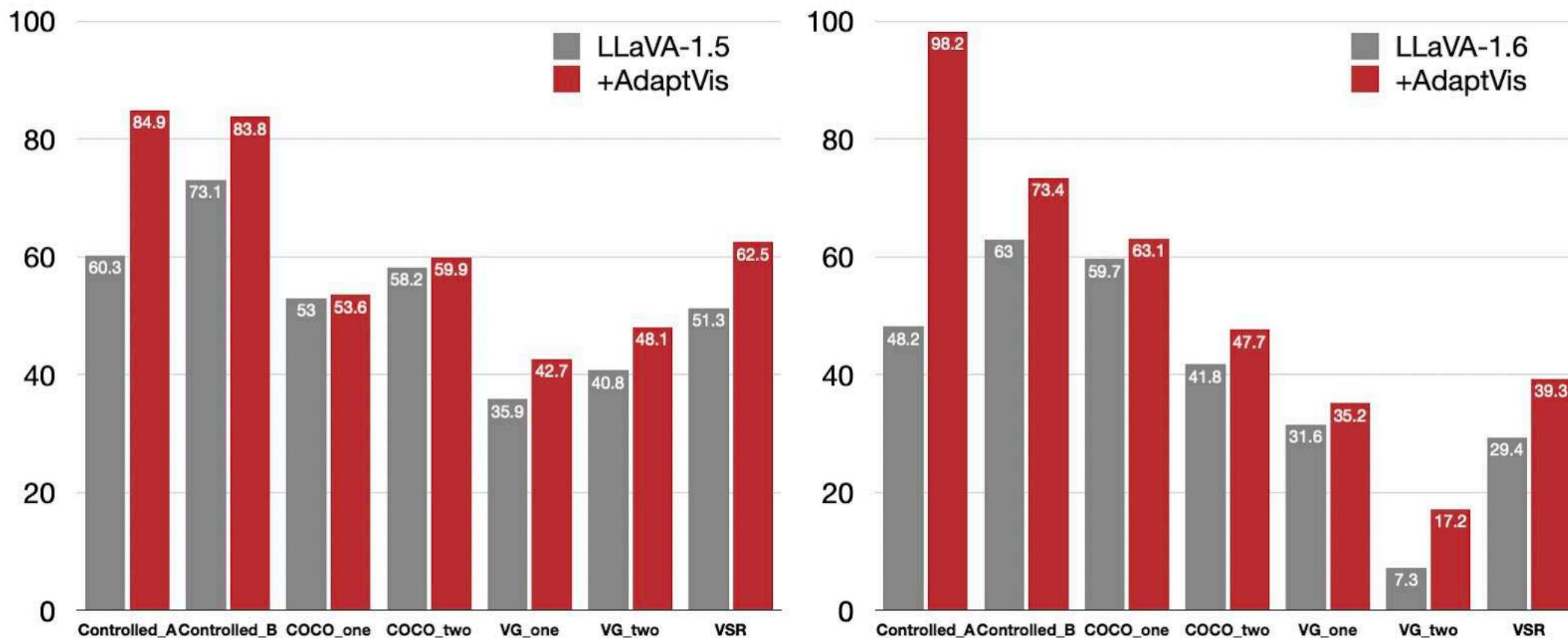
For low-confidence relationships: coefficient <1 improves performance. For high-confidence relationships: coefficient >1 improves performance.



Model has **higher** confidence for **left** / **right** than **on** / **under** / **front** / **behind**, indicating that the model is more familiar with certain relationships.

Simple Intervention on attention can be helpful

Performance Comparison



Attention behavior of VLMs in Spatial Reasoning

from a mechanism interpretability lens

Attention behavior of VLMs in Spatial Reasoning

from a mechanism interpretability lens

1. What causes these failures?

2. How do these failures manifest through internal patterns?

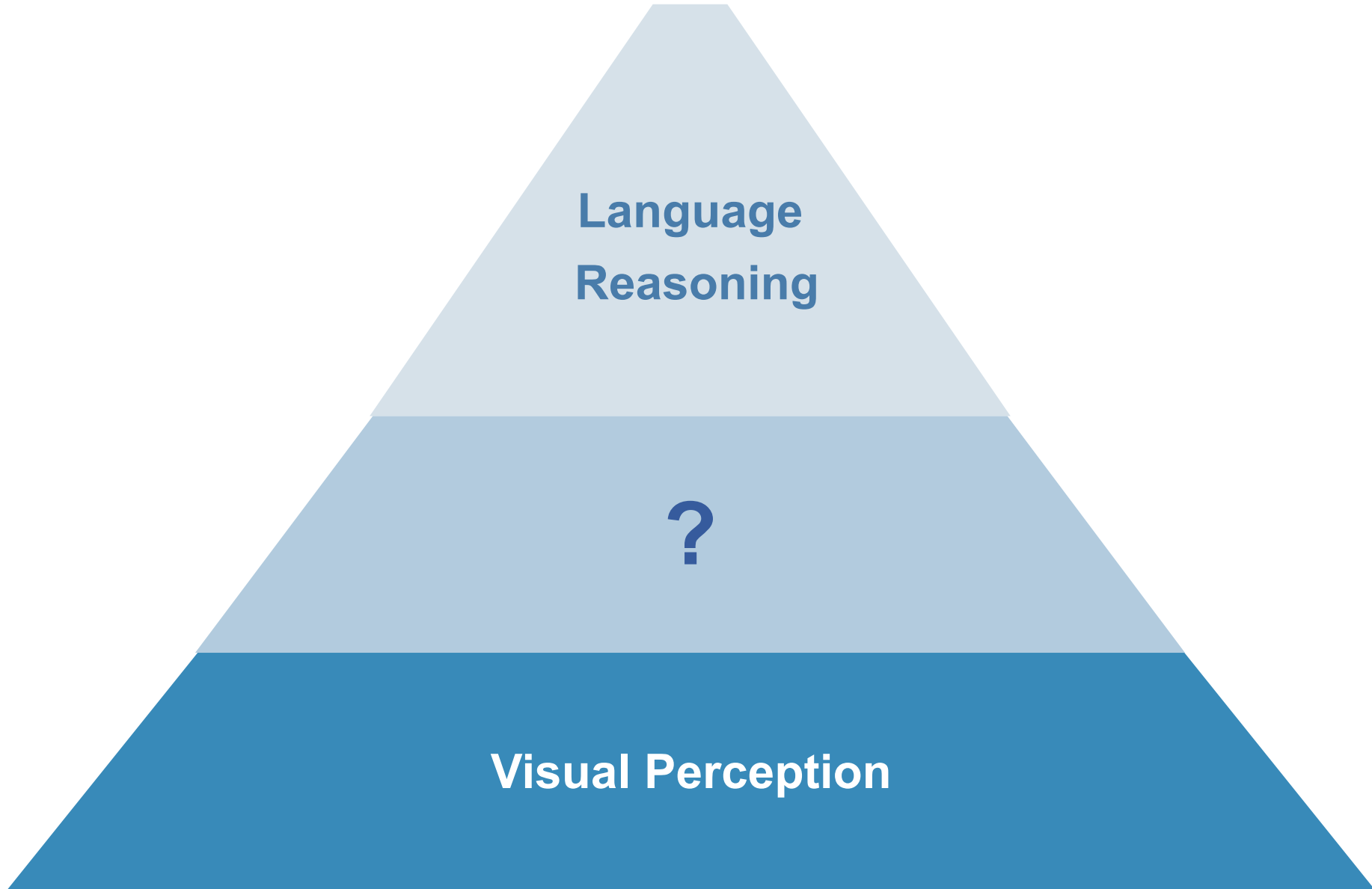
3. Can we mitigate these errors by leveraging the identified signals

Limitations: Reversal Curse?

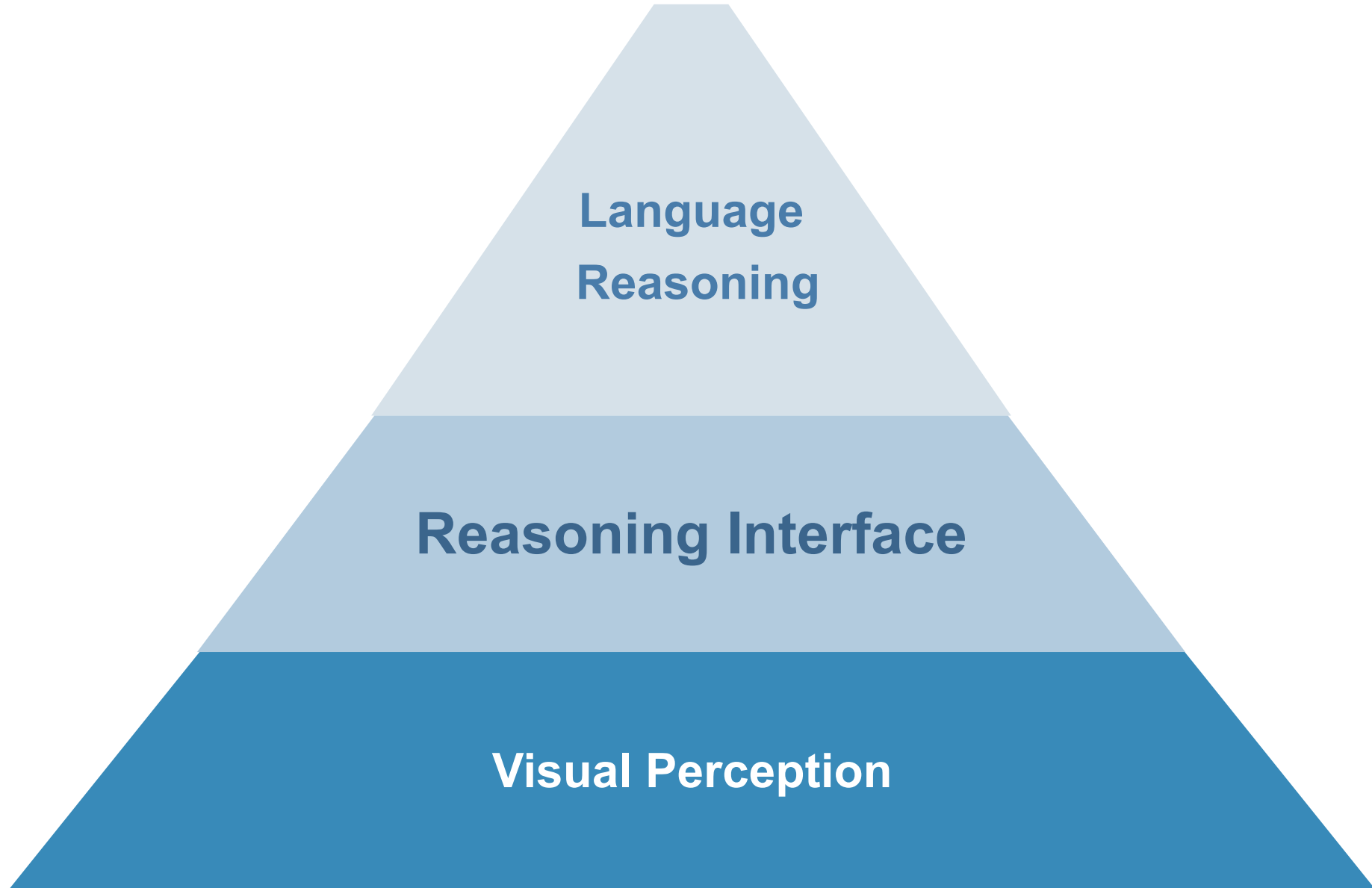
``Where is the armchair in relation to the beer bottle?" instead of ``Where is the beer bottle in relation to the armchair?"

Metric	Con_A	Flipped_Con_A
Acc	76.4	35.2 ↓41.2
Pair Acc	43.0	1.2 ↓41.8
Set Acc	4.8	0.0 ↓4.8

What is Missing? **Abstraction** Layers in VLM Pyramid



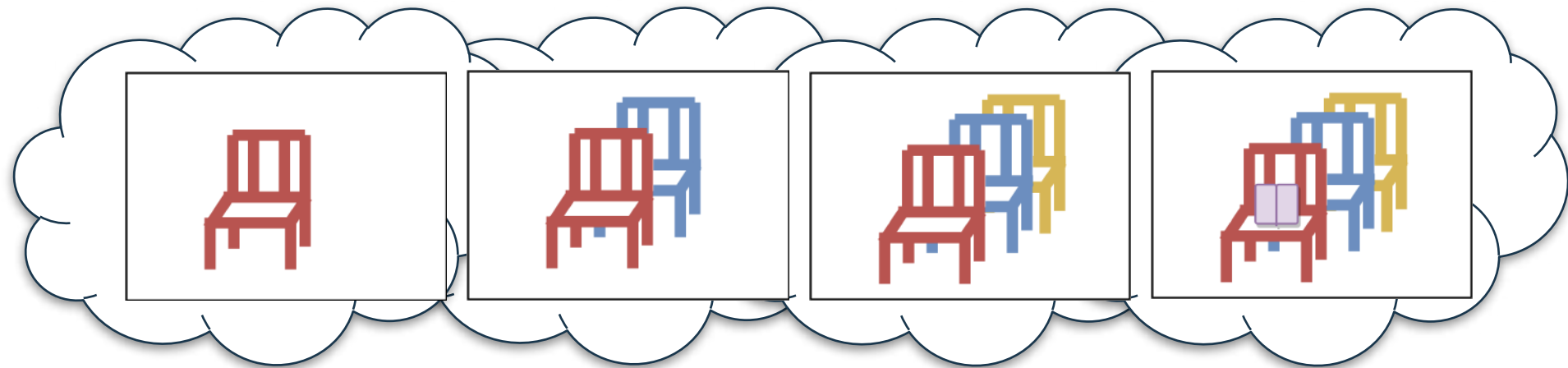
What is Missing? **Abstraction** Layers in VLM Pyramid



Reasoning over "Internal Belief"

Place a **blue chair** behind a **red chair**. Then, put a **yellow chair** behind the **blue chair**. Then, put a **book** on top of the chair that is in front of the **blue chair**.

Question: What chair is the book on?

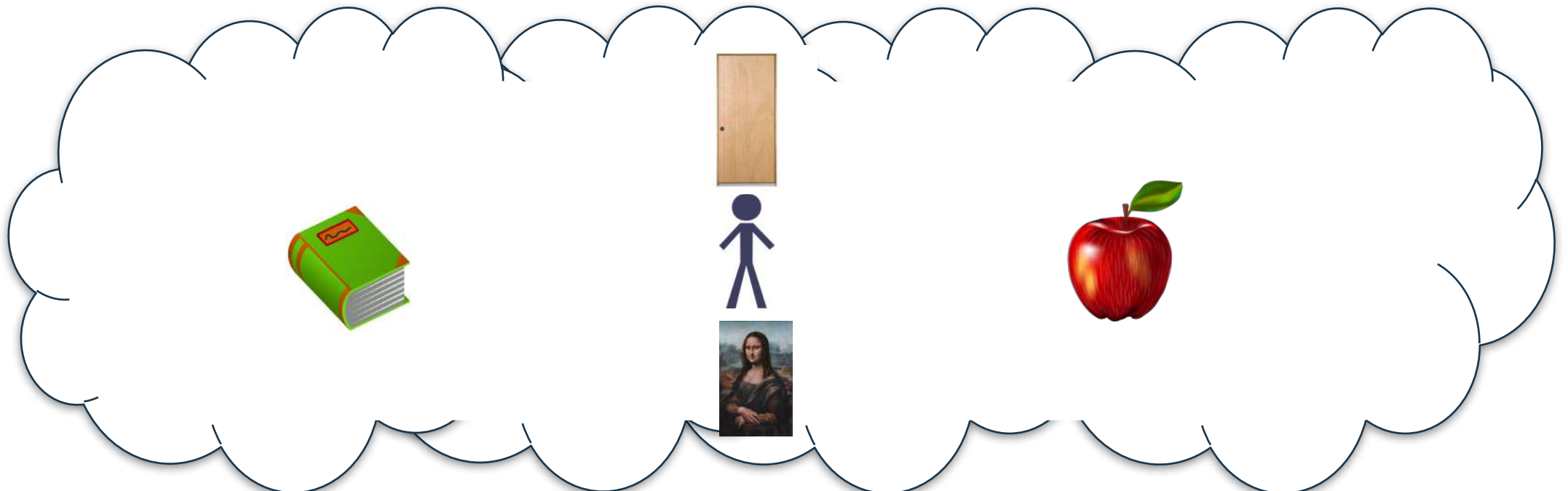


Reasoning over "Internal Belief"

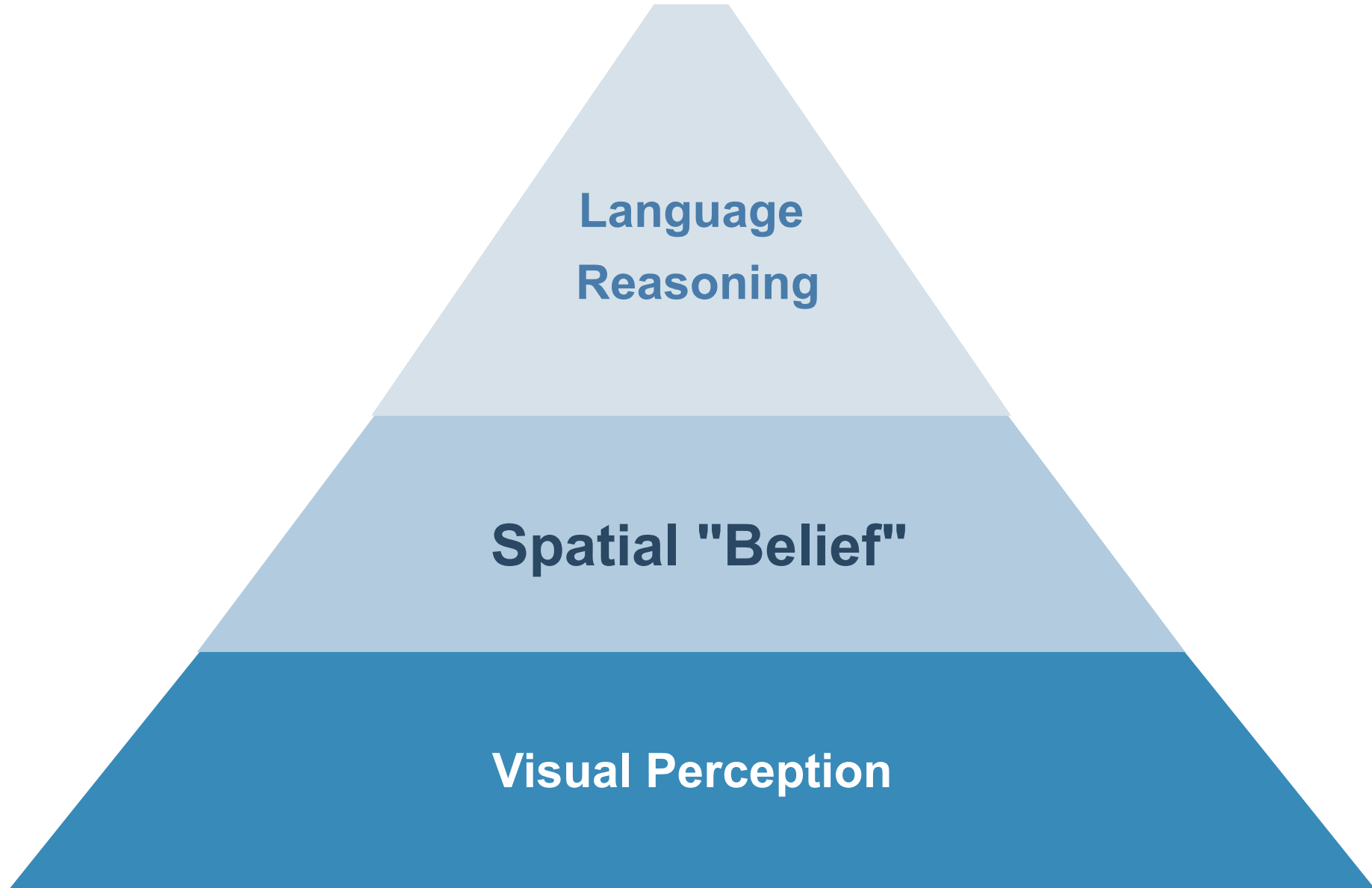
User

You are in a room with a book on your left, a door in front of you, an apple on your right, and a painting behind you.

“What if”: If you turn right, what will be behind you?



What is Missing? **Abstraction** Layers in VLM Pyramid



Reasoning Interface:

Horizon

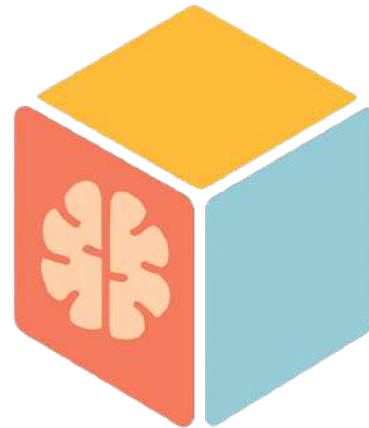
*Semantic-
Centric*

Long-Horizon

Developing Spatial "Belief"

*Geometric-
Centric*





spatial mental models

Spatial Mental Modeling From Limited Views



Best Paper Award, Structural Priors for Vision @ICCV



Qineng Wang*



Baiqiao Yin*



Pingyue Zhang



Jianshu Zhang



Kangrui Wang



Zihan Wang



Jieyu Zhang



Keshigeyan
Chandrasegaran



Han Liu



Ranjay Krishna



Saining Xie



Jiajun Wu†



Fei-Fei Li†



Manling Li†

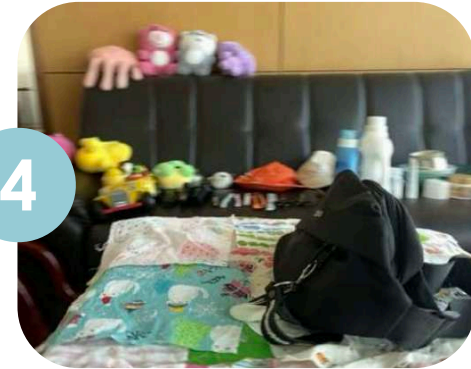




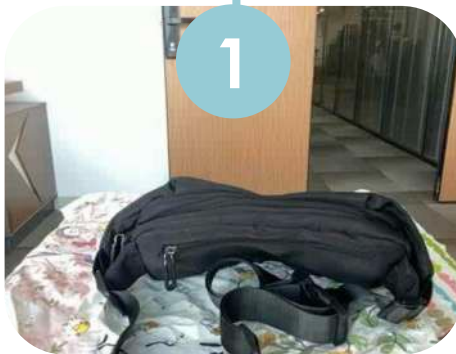
3



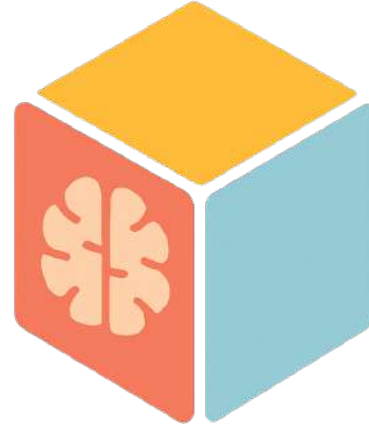
2



4



1



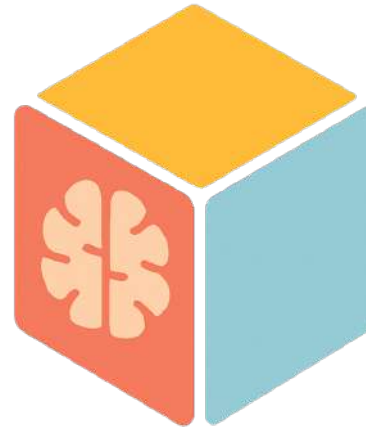
At **4** what is to the **left** of the **black bag** ?

A. Window

B. Door

C. Desk

D. Sofa



At **4** what is to the **left** of the **black bag** ?

A. Window

B. Door

C. Desk

D. Sofa



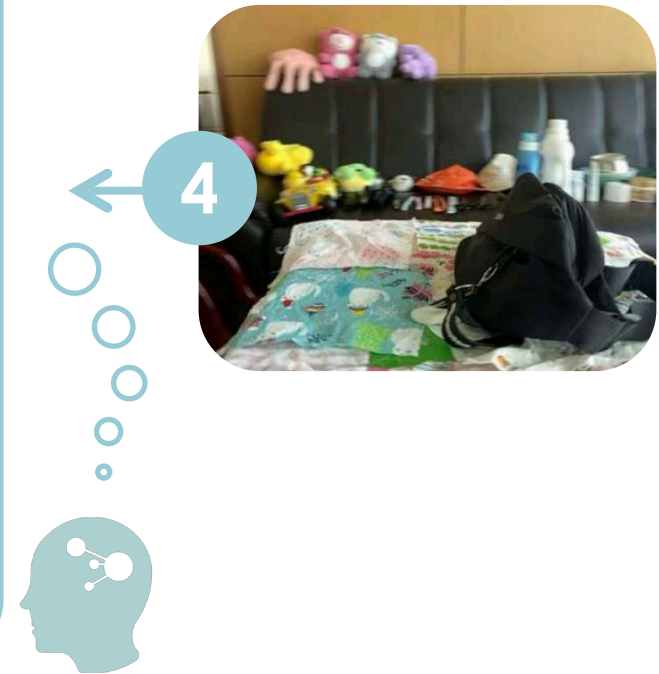
At **4** what is to the **left** of the **black bag** ?

A. Window

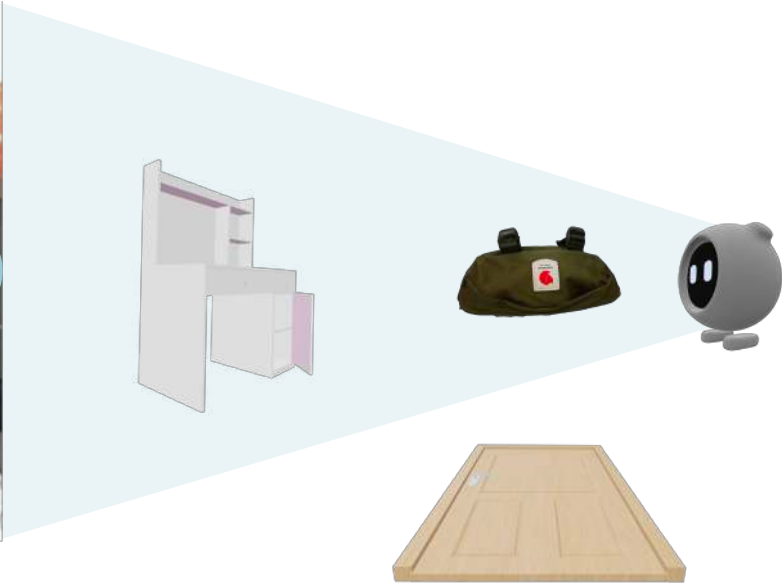
B. Door

C. Desk

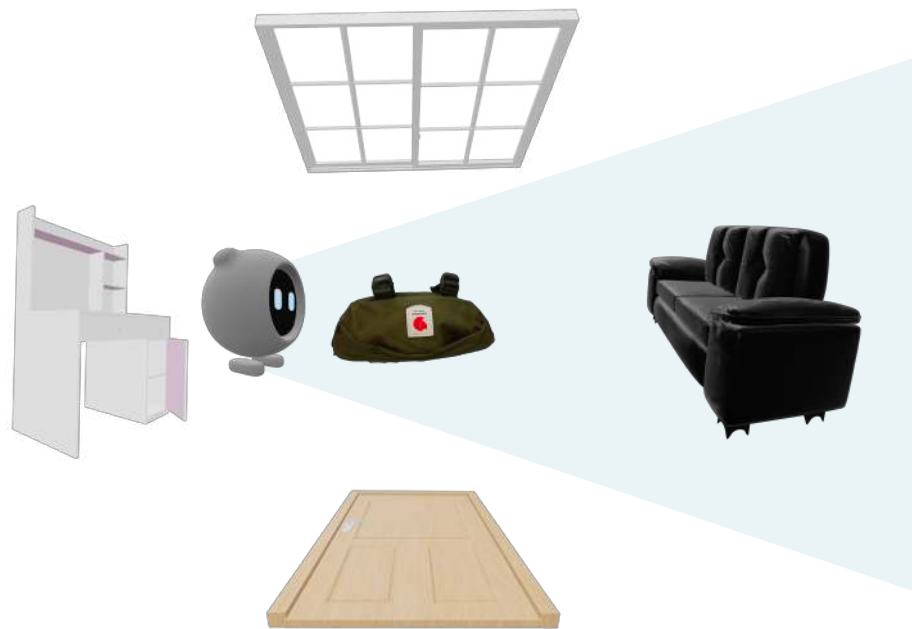
D. Sofa

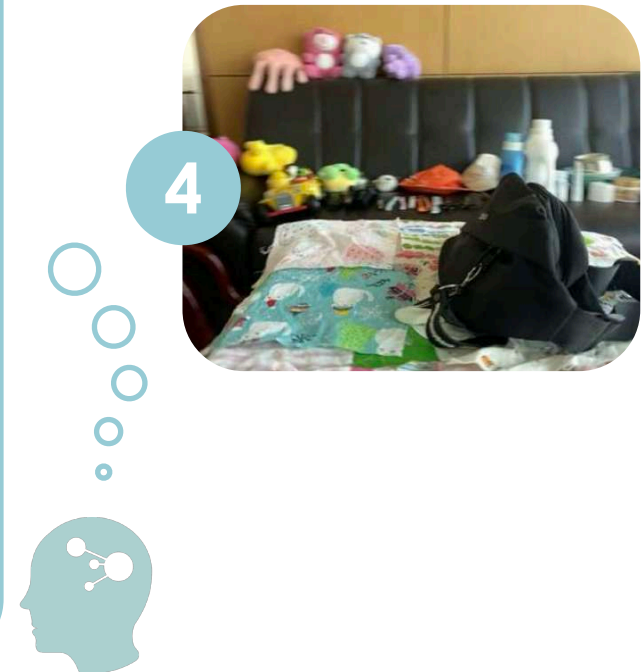




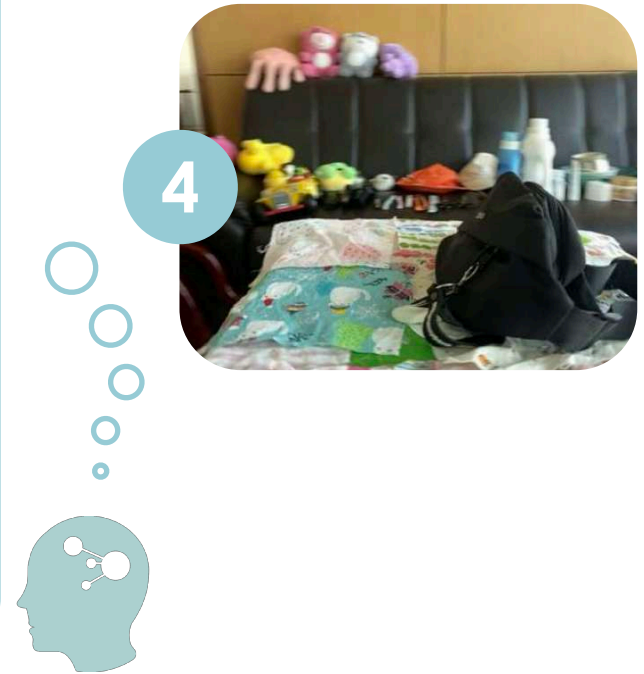
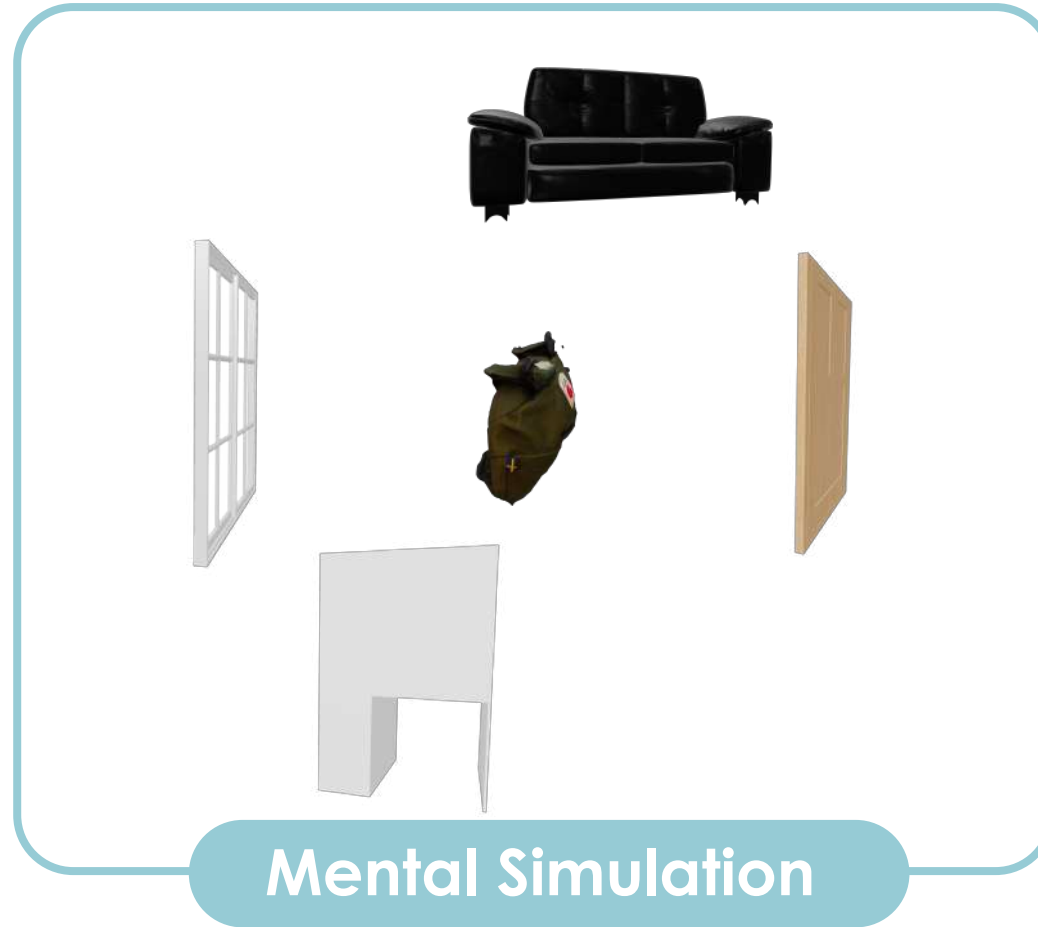








At **4** what is to the **left** of the **black bag** ?



At **4** what is to the **left** of the **black bag** ?

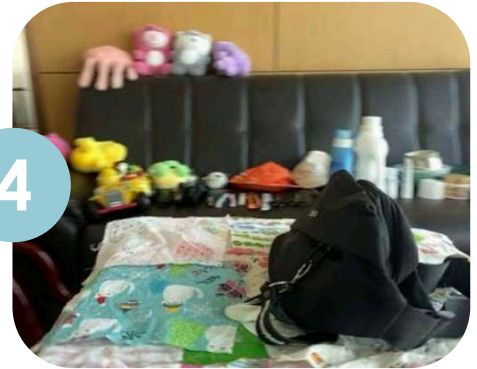
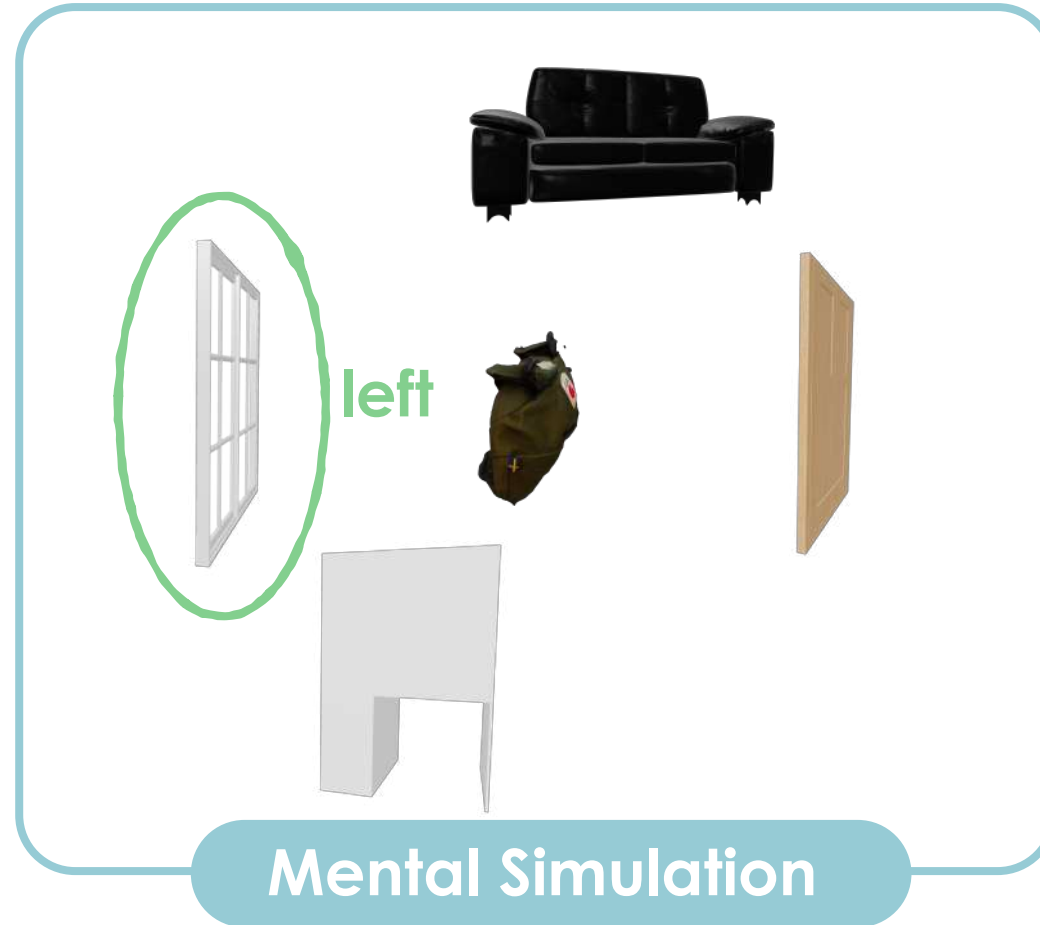
A. Window



B. Door

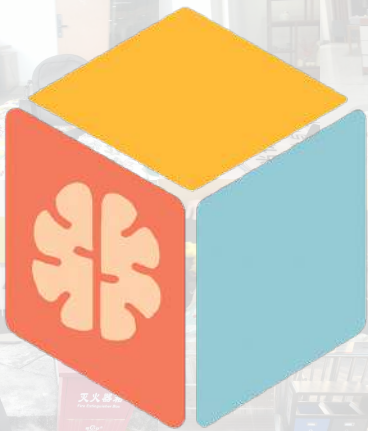
C. Desk

D. Sofa









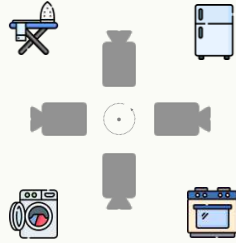
MindCube

21,154 questions

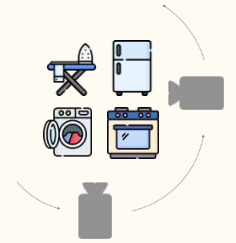
3,268 images

Three Movement Patterns in MindCube

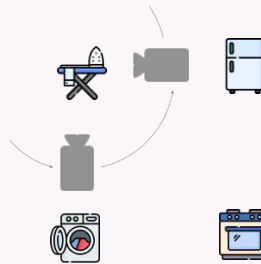
Rotation



Around



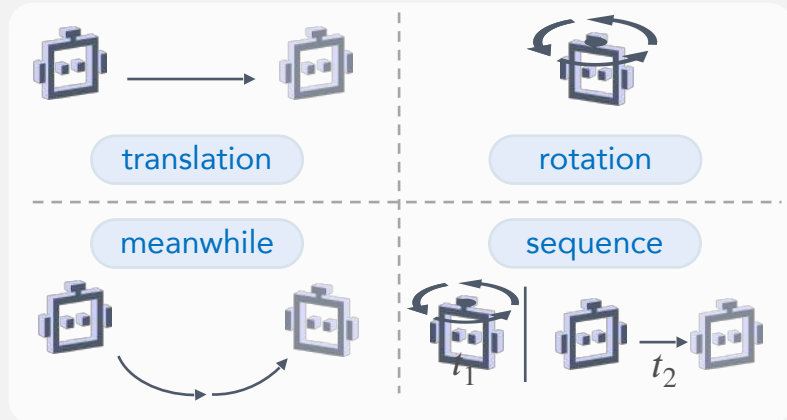
Among



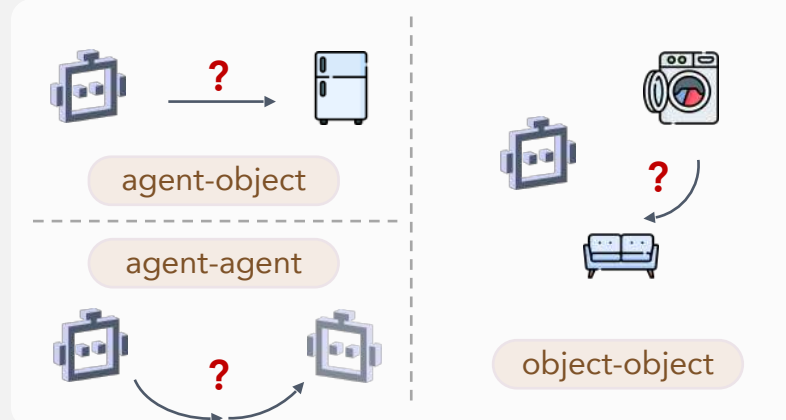
Fine-Grained Question Types in MindCube

Question Types

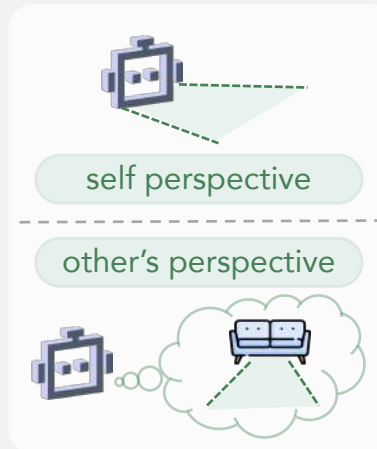
“What if” Dynamics



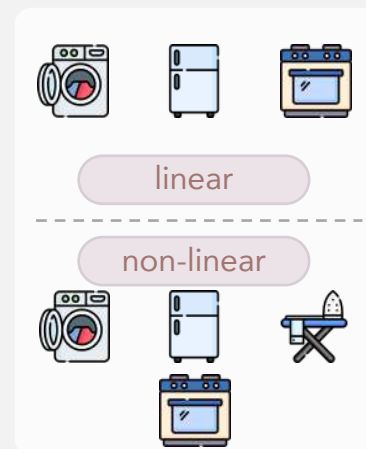
Relation Query



Perspective Taking



Visual Patterns



Rotation



Question: If you are at the **third viewpoint** and turn 90 degrees to the left, what is to your left?

Options:

- A. Metal bin
- B. Table
- C. Pathway
- D. Bookcase



:

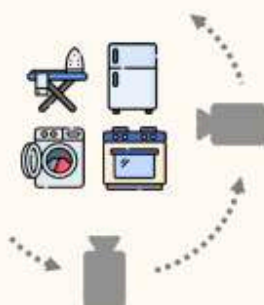
rotation

agent-object

self perspective

non-linear

Around



Question: If you are positioned at the **third viewpoint**, then turn left and move forward, will you get closer to the red trash bin?

Options:

- A. Yes
- B. No



:

sequence

agent-object

self perspective

linear

Among



Question: If you are positioned at the **first viewpoint**, what is to the left of the black boots from where you stand?

Options:

- A. Sofa
- B. Windows
- C. TV cabinet
- D. Dining Table



:

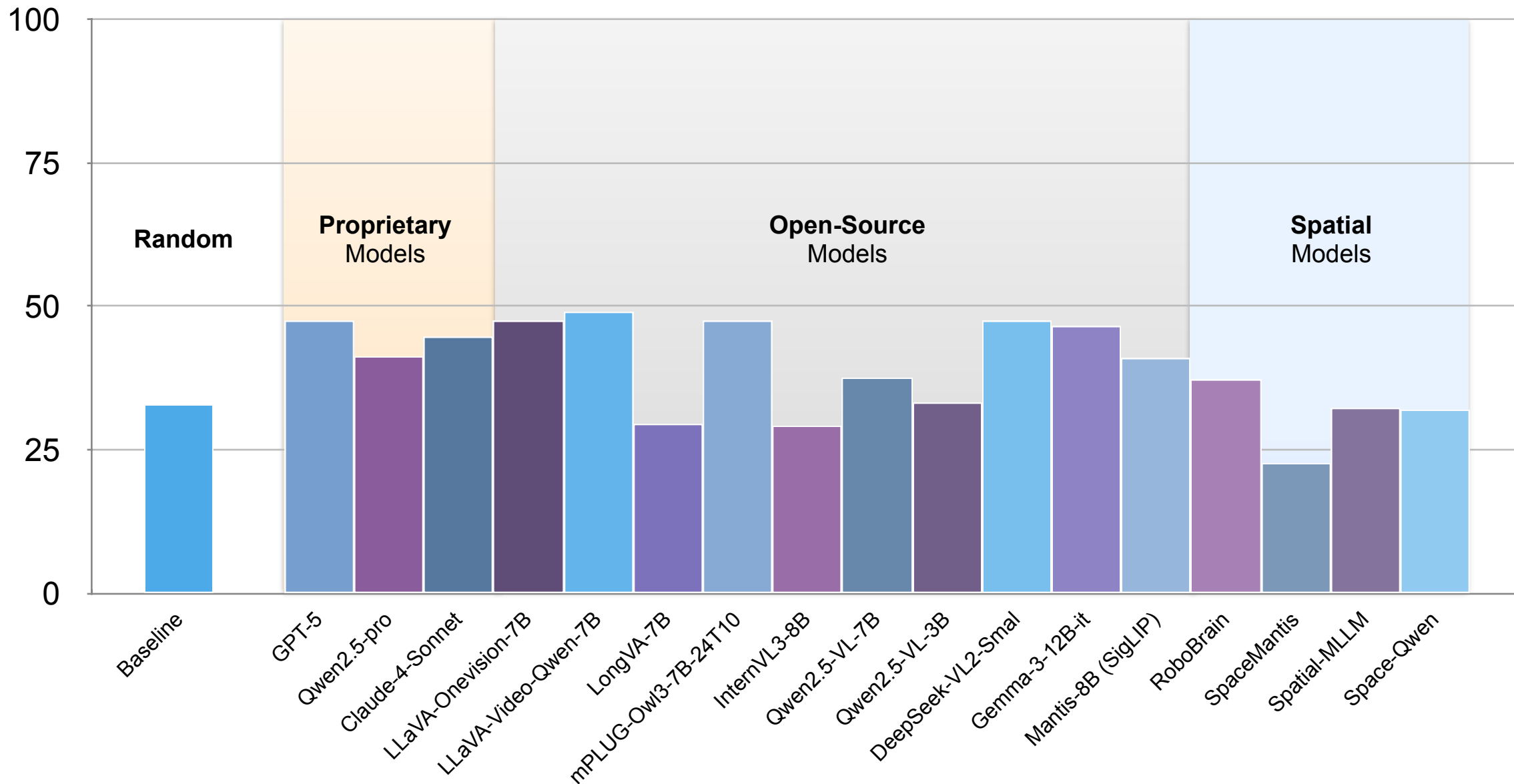
meanwhile

object-object

self perspective

non-linear


Still Challenging !




**How to teach VLMs to
Approximate Spatial Mental Models ?**


Approximate Spatial Mental Models

View Interpolation




+0.10% 


Free Form Reasoning



Reasoning Instruct
Please do **step by step reasoning**, then give final answer...


+2.67% 

Cognitive Map



Augmented Cognitive Map

```
{
  "objects": [
    {
      "name": "Tissue box",
      "position": [5, 5]
    },
    {
      "name": "Hand sanitizer",
      "position": [7, 5]
    },
    ...
  ],
  "views": [
    {
      "name": "View 1",
      "position": [5, 6],
      "facing": "up"
    },
    ...
  ]
}
```

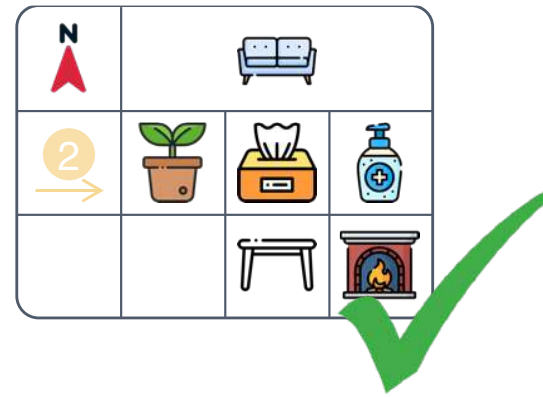
+3.52% 

Data Scaffolds

Compare to Raw QA

Just prompting

If we add training...



Approximate Spatial Mental Models

"Map then Reason" is the most effective approximation



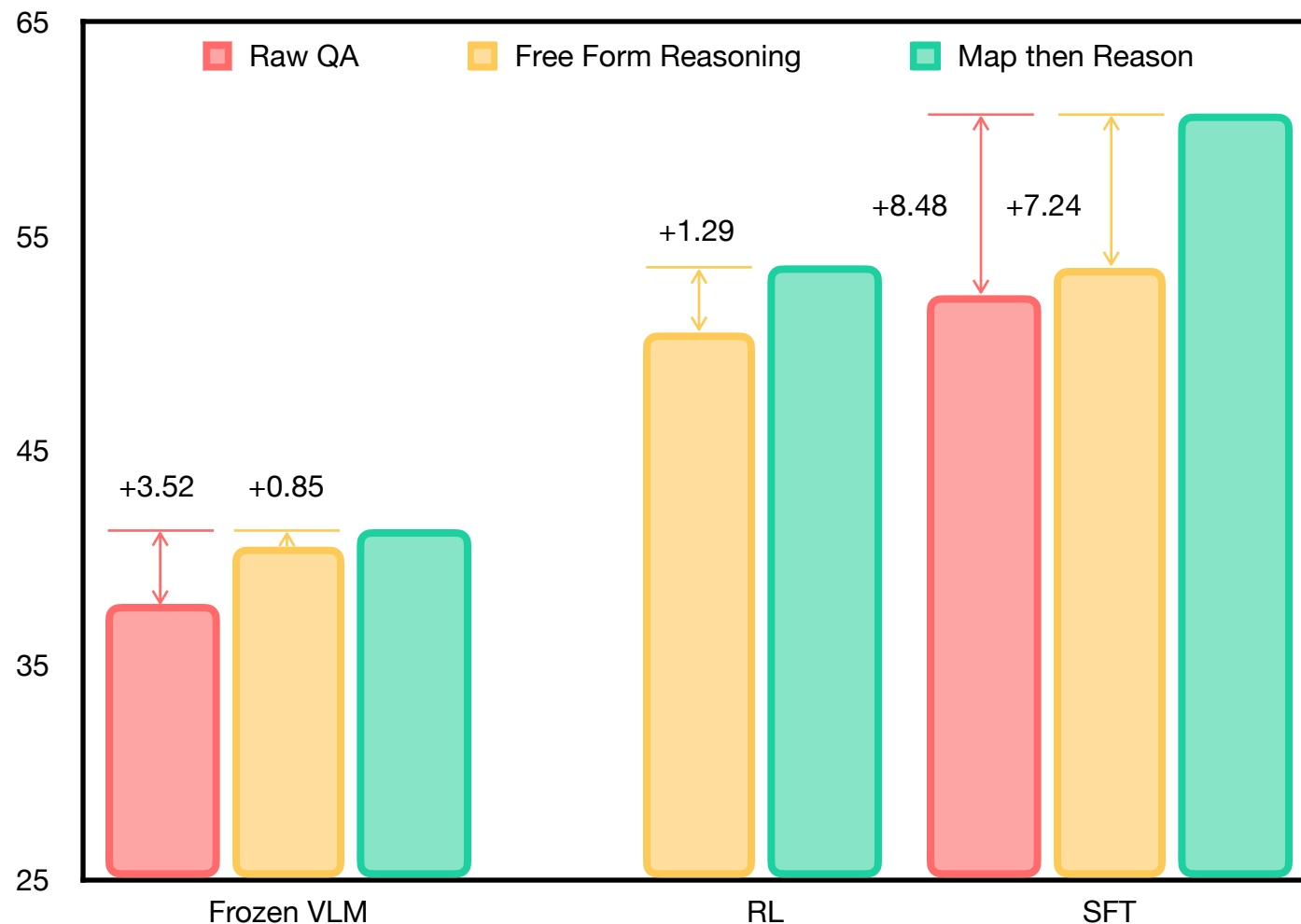
Raw QA



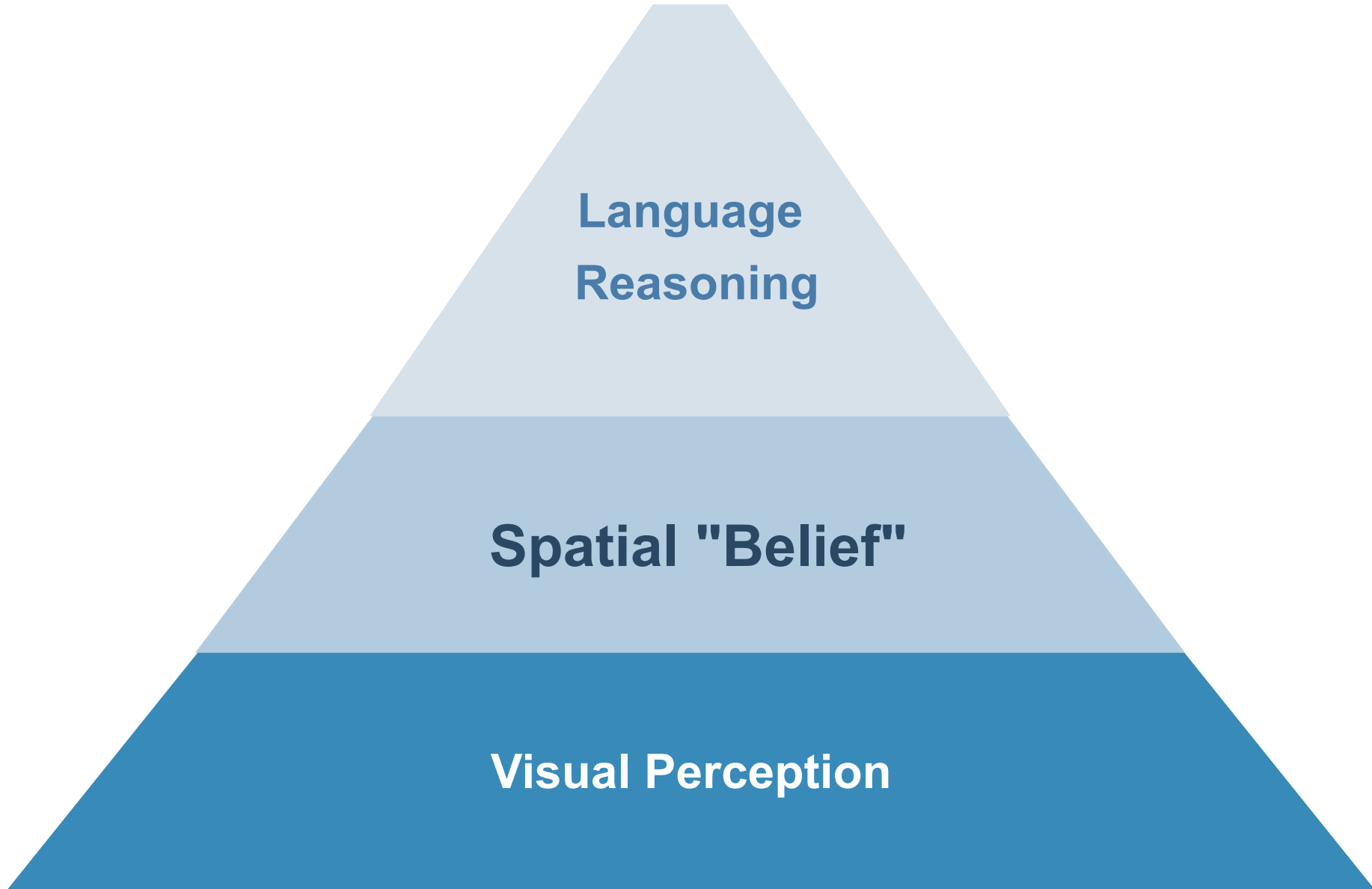
Reasoning



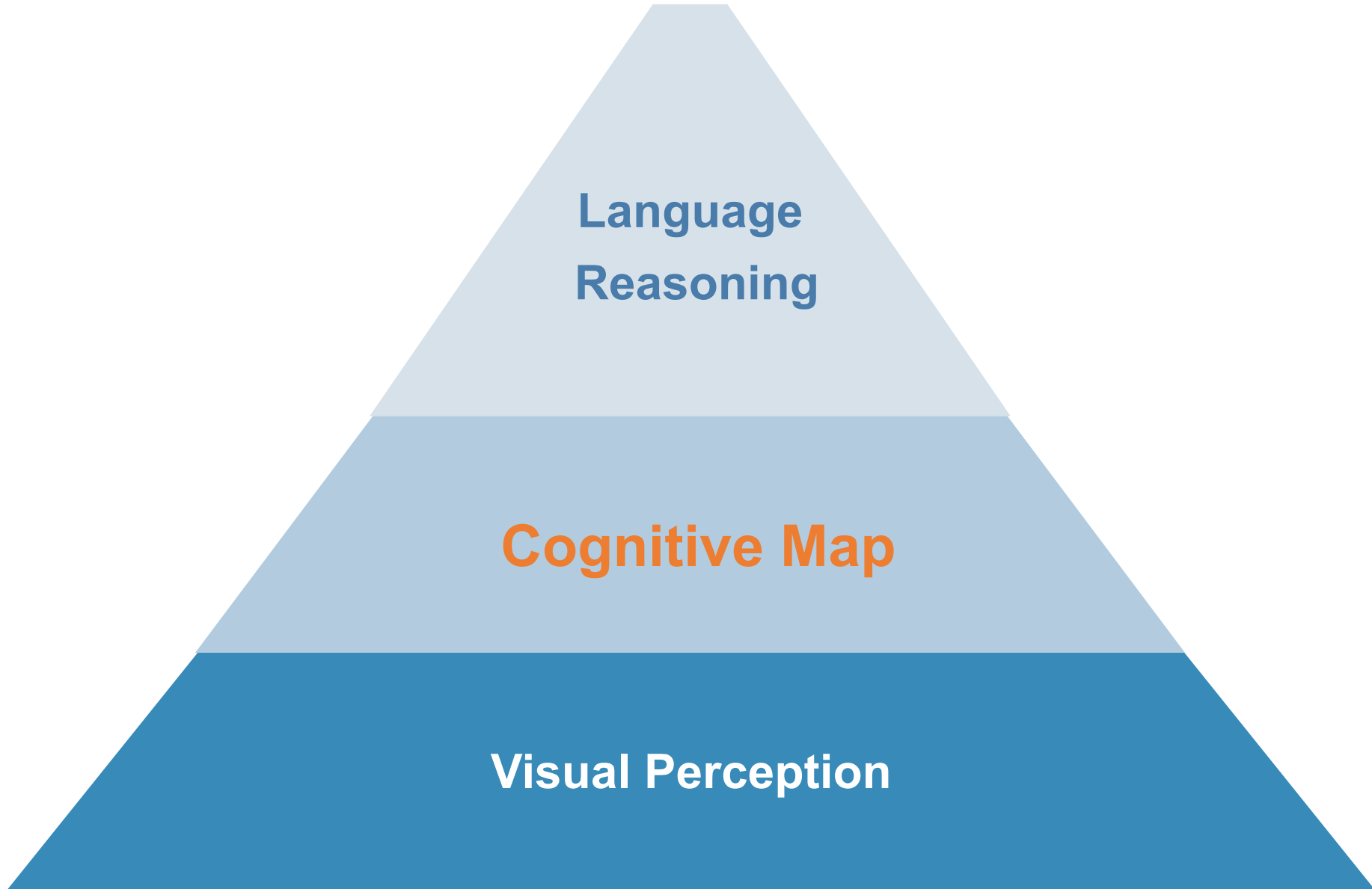
Map then Reason



What is Missing? **Abstraction** Layers in VLM Pyramid



What is Missing? **Abstraction** Layers in VLM Pyramid



What did the model learn?

Better QA



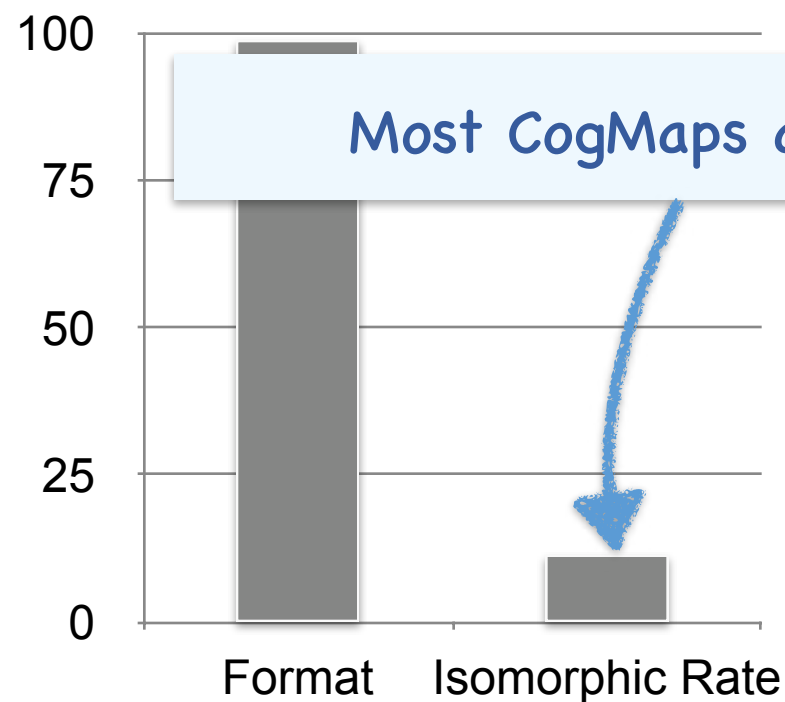
Better CogMap

Better QA

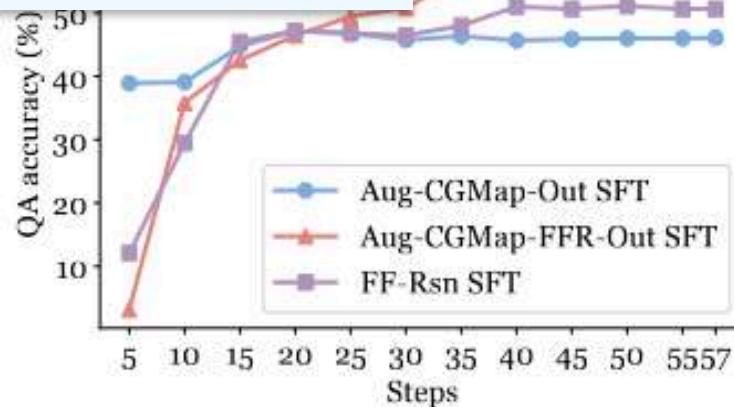


Better CogMap

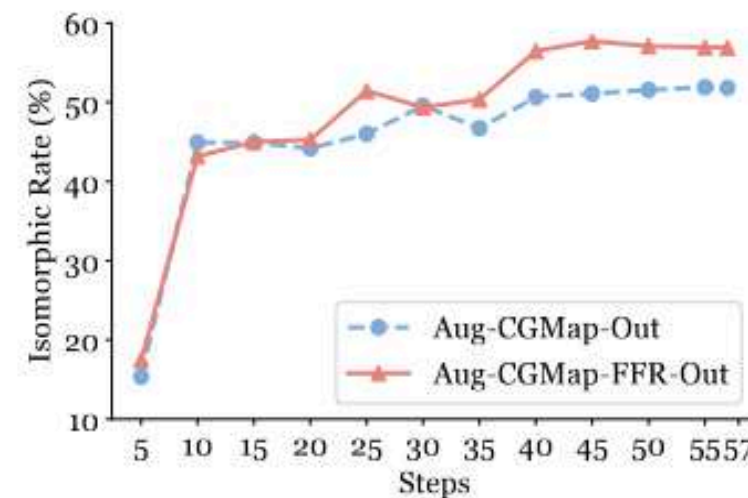
Before Training



After Training



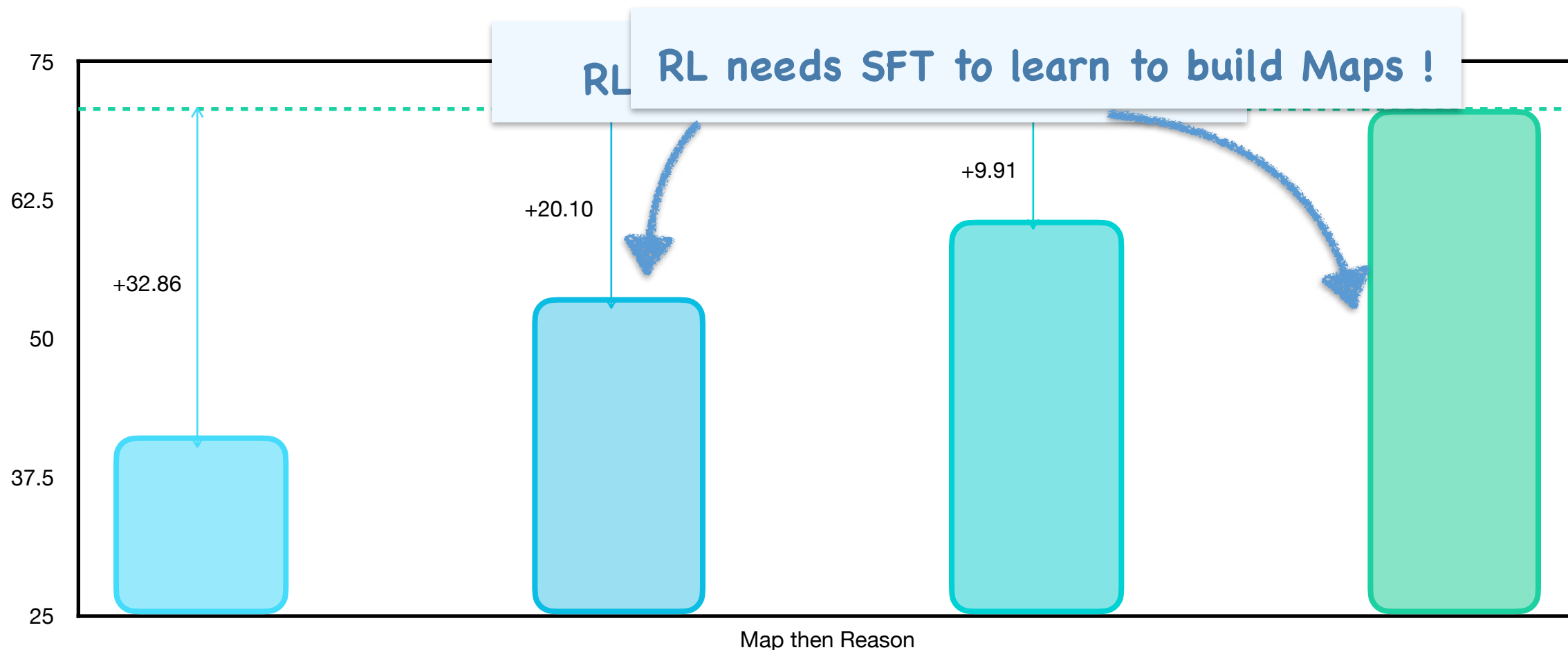
(a) QA accuracy (%)

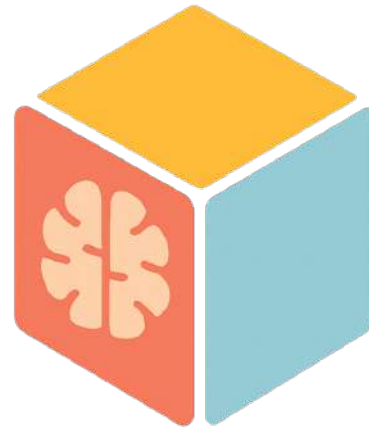


(b) Isomorphic Rate (%)

**"Map then Reason" is also the
secret recipe for RL**

RL shines when warmed up with Map-Learning SFT

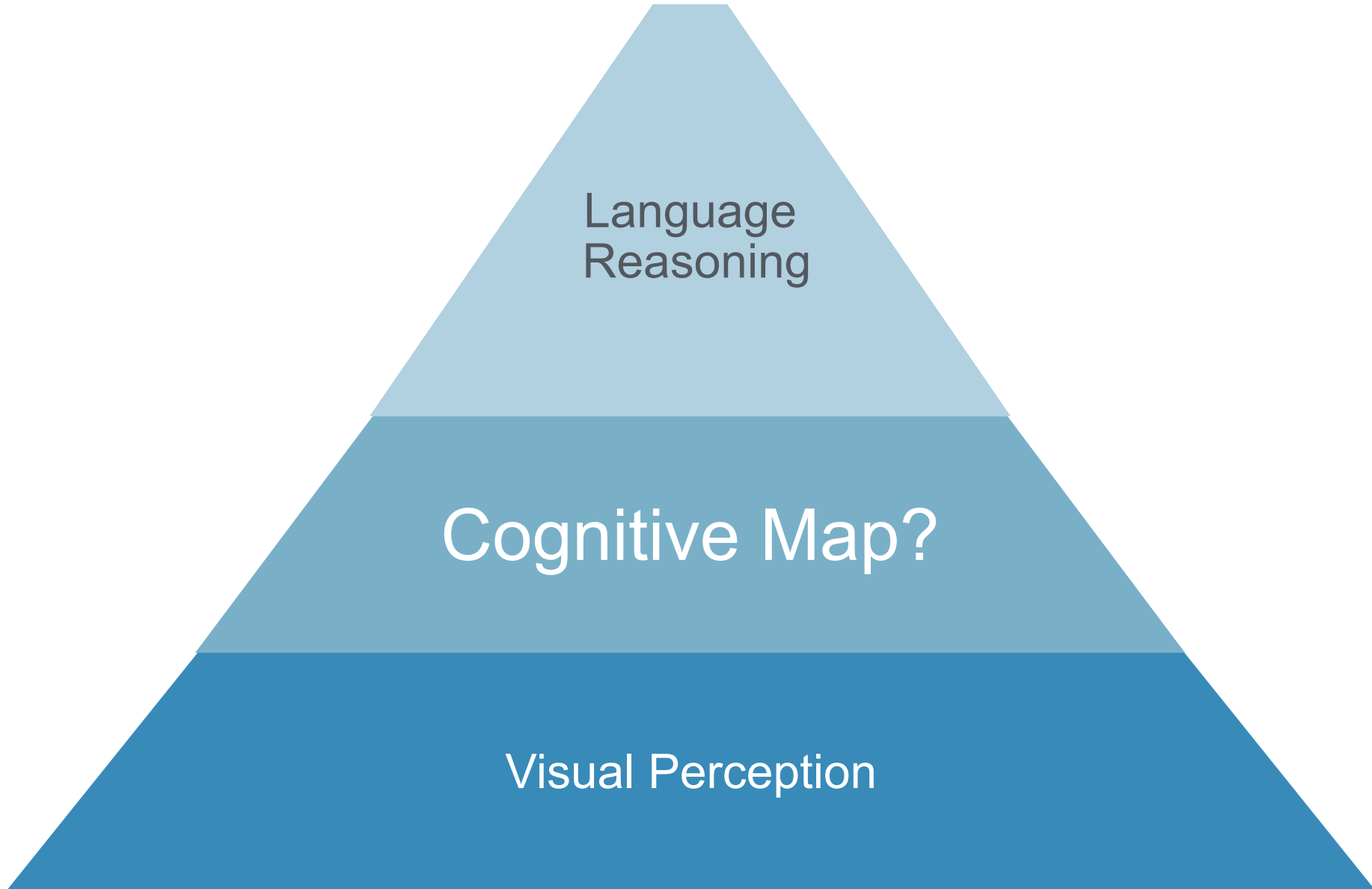


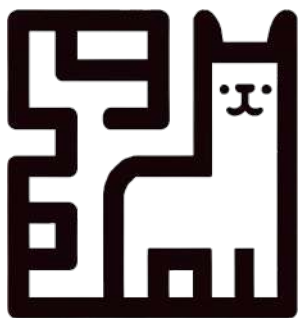


MindCube

<https://mll-lab-nu.github.io/mind-cube>

What is Missing? Intermediate Layers in VLM Pyramid





Visually Descriptive Language For Vector Graphics Reasoning



Zhenhailong Wang



Joy Hsu



Xingyao Wang



Kuan-Hao Huang



Manling Li



Jiajun Wu



Heng Ji



Code: [VDLM Code](#)



Demo (Jupyter Notebook): [VDLM Demo](#)



SVG-to-PVD Dataset: [PVD-160K](#)



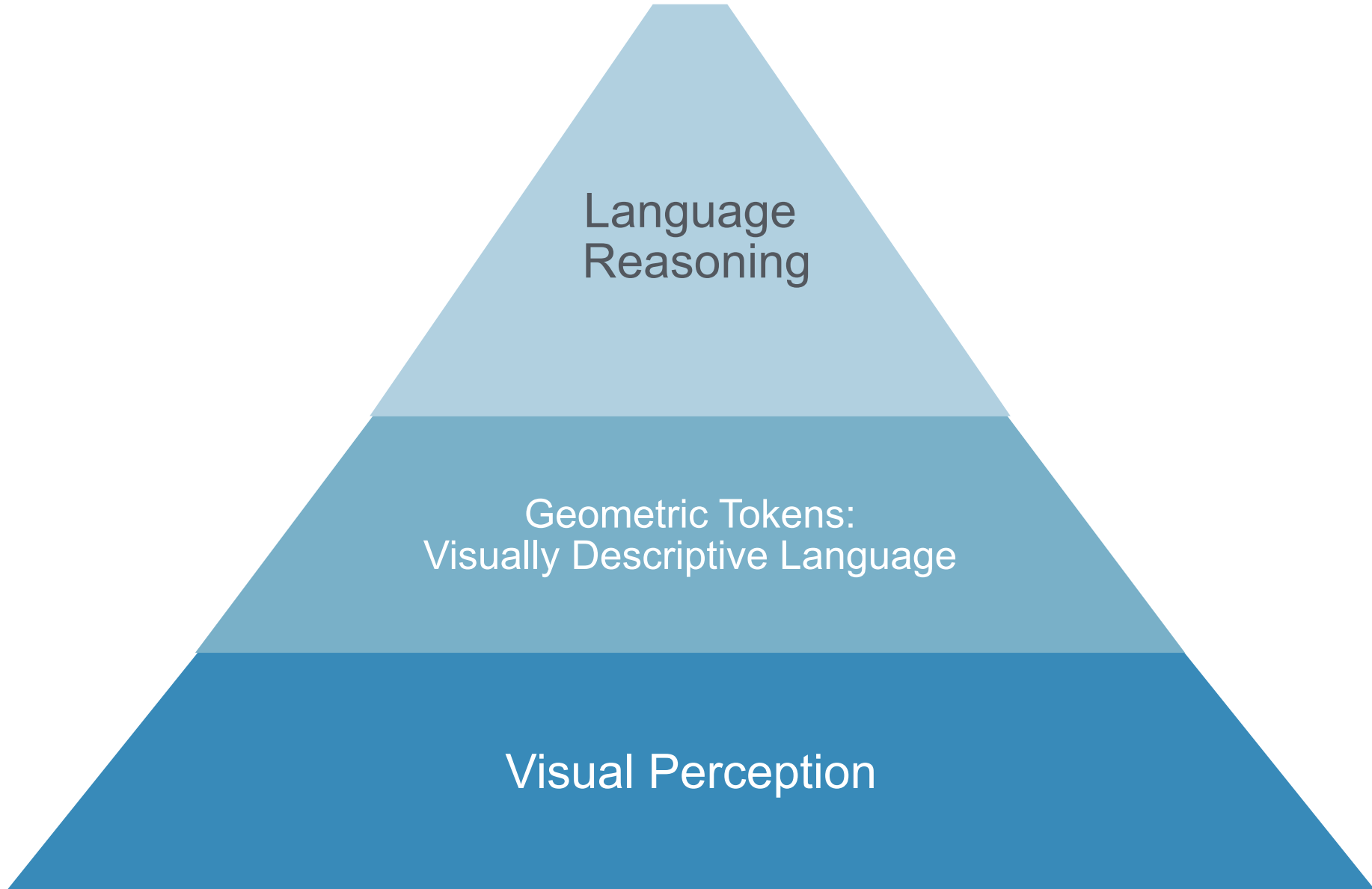
Pretrained SVG-to-PVD Model: [PVD-160k-Mistral-7b](#)



Go to lower-level:

What is Missing? Intermediate Layers in VLM Pyramid

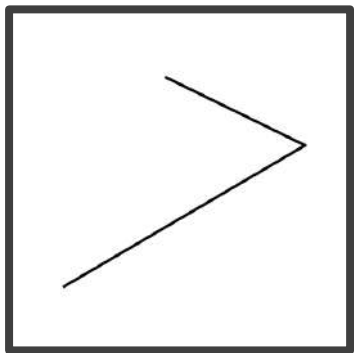
What is Missing? Intermediate Layers in VLM Pyramid



**We need Abstraction (Geometric Tokens)
for positions, shapes, etc**

Using **SVG** as intermediate representations?

We encode images with SVG for **precise** low-level perception



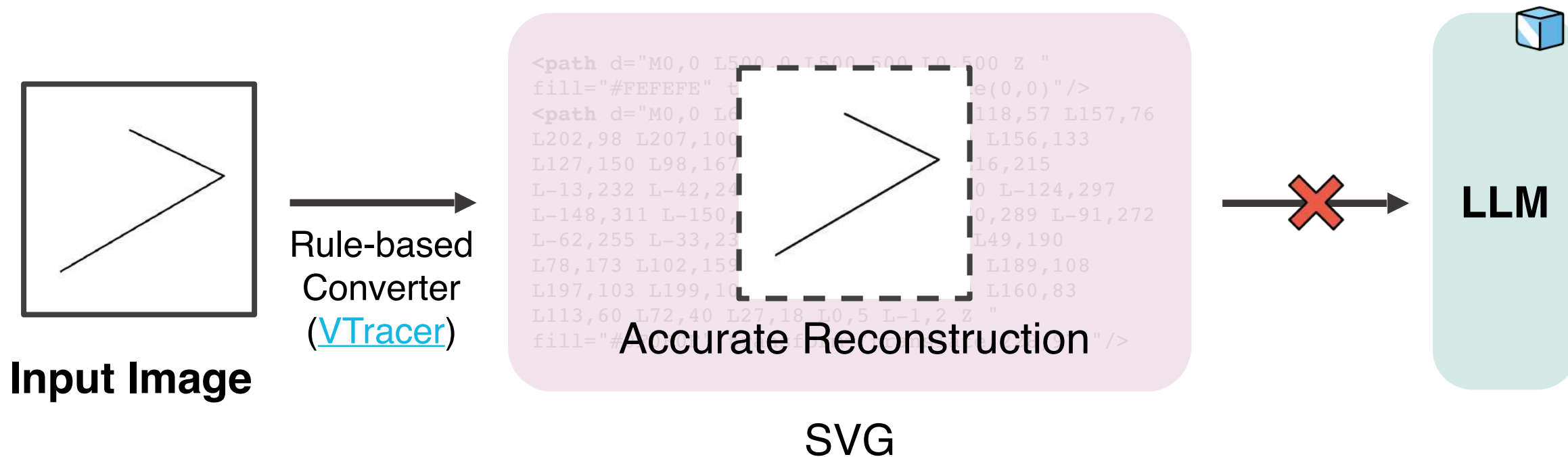
Input Image

→
Rule-based
Converter

```
<path d="M0,0 L500,0 L500,500 L0,500 Z "
fill="#FEFEFE" transform="translate(0,0)"/>
<path d="M0,0 L6,2 L42,20 L65,31 L118,57 L157,76
L202,98 L207,100 L206,104 L178,120 L156,133
L127,150 L98,167 L69,184 L45,198 L16,215
L-13,232 L-42,249 L-71,266 L-95,280 L-124,297
L-148,311 L-150,310 L-149,306 L-120,289 L-91,272
L-62,255 L-33,238 L-9,224 L20,207 L49,190
L78,173 L102,159 L131,142 L160,125 L189,108
L197,103 L199,103 L199,101 L193,99 L160,83
L113,60 L72,40 L27,18 L0,5 L-1,2 Z "
fill="#0F0F0F" transform="translate(228,97)"/>
```

SVG

However, LLMs cannot directly understand SVG in a zero-shot setting



What are properties of Geometric Tokens?

What are properties of Geometric Tokens?

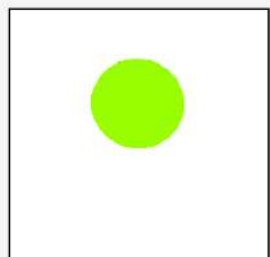
Compositional
(Simple → Complicated)

Annotation-Free
(Synthetic Data)

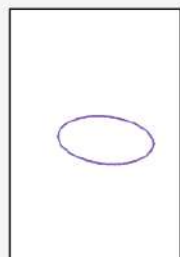
Properties of Primal Visual Description (PVD)

Property 1: Essential building blocks with a high coverage

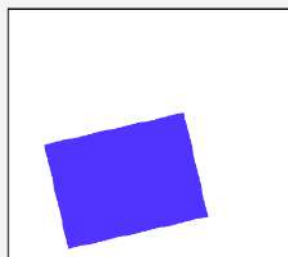
PVD Primitives Ontology



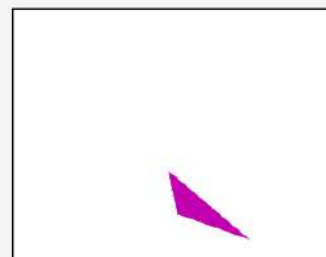
Circle



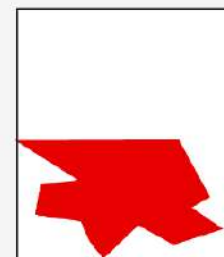
Ellipse



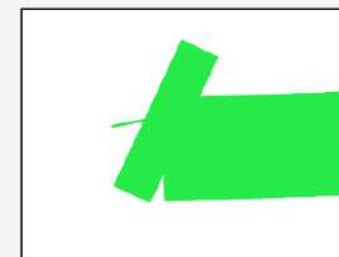
Rectangle



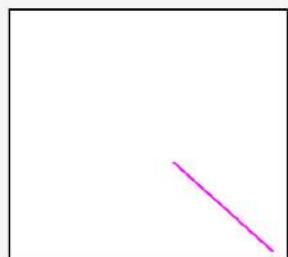
Triangle



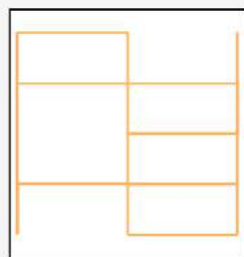
Polygon



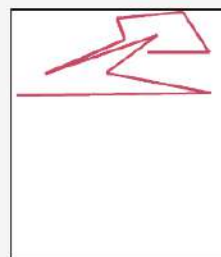
Composition-filled



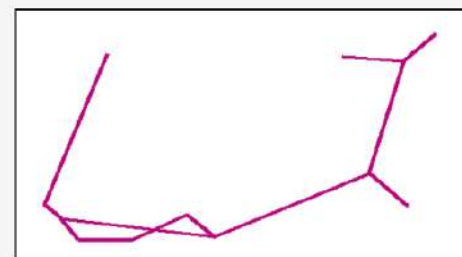
Line Segment



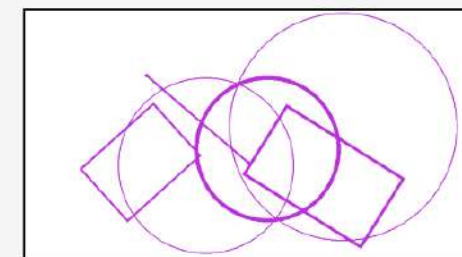
Grid



Path



Graph (line drawing)



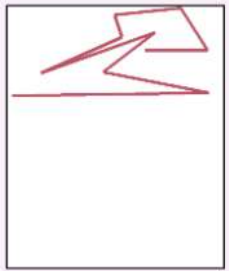
Composition-outlined

Unlike raw SVG, PVD is directly interpretable by state-of-the-art LLMs, enabling zero-shot reasoning on downstream tasks.

Properties of Primal Visual Description (PVD)

Property 2: Easy to train without human annotation

SVG-to-PVD Model Input & Output



```
<path
d="M0,0 L2,3 L5,3
L24,32 L53,76 L63,91
... Z M-280,138 Z M-
283,139 Z M-285,140
Z " fill
="#B95163" transform
="translate(504,7)"
/>
```

Input: SVG file containing a single
<path/> corresponding to a primitive

Fine-tuning
Mistral-7B

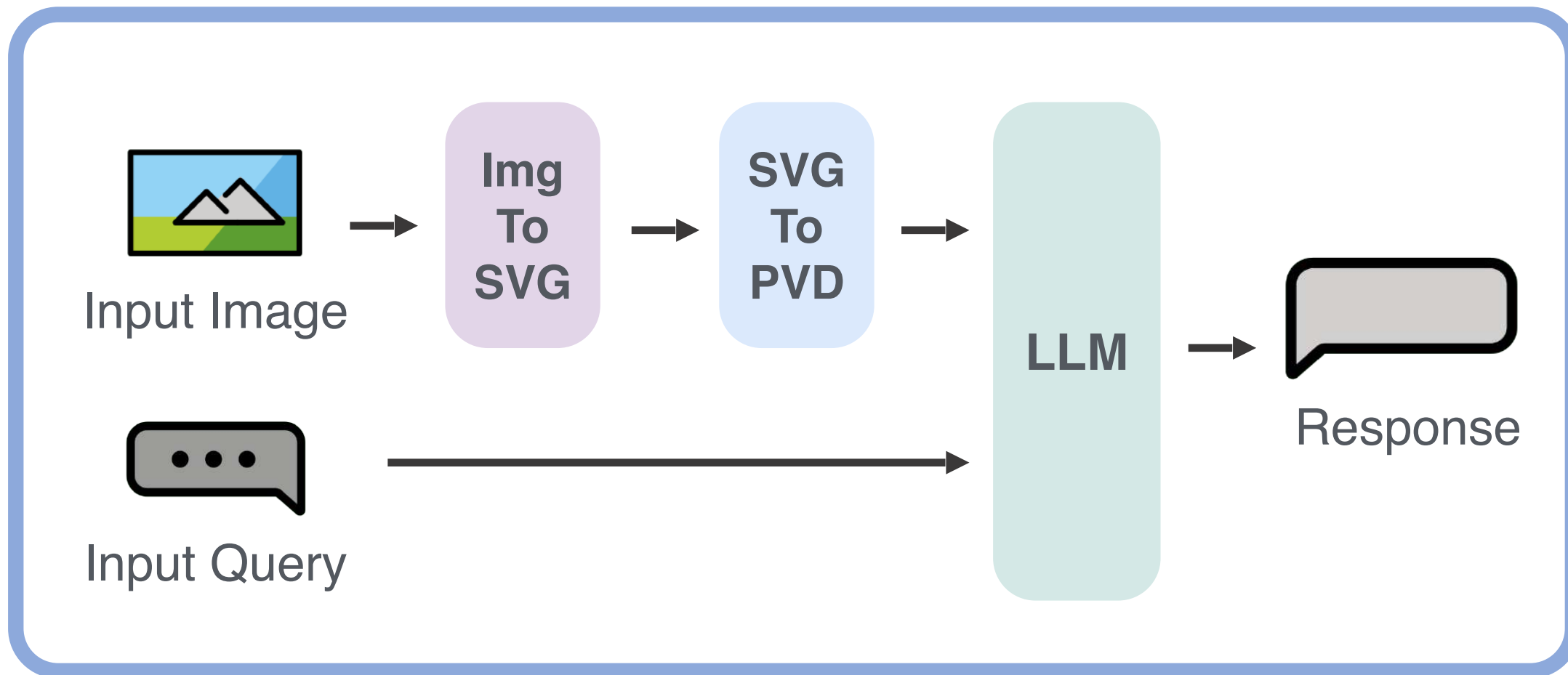
```
{
  "type": "path",
  "vertices": [[19, 255],[585,
    247], ...],
  "edges": [[[19, 255],[585,
    247]], ...],
  "style": "outlined shape",
  "color": [185, 81, 99],
  "line_width": 9
}
```

Target Output: PVD JSON

We develop a data generator leveraging PIL.ImageDraw and VTracer, which creates a large-scale ⟨SVG, PVD⟩ paired dataset.

Plug-In to any frozen foundation models

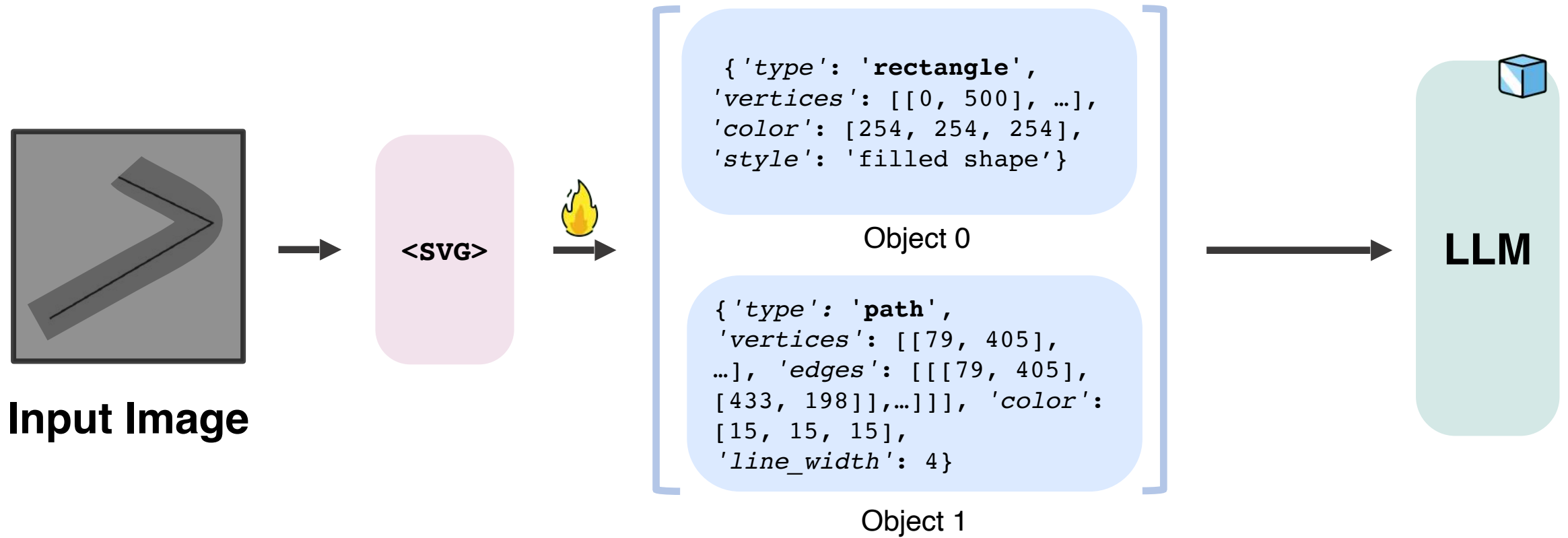
Overview



Overview

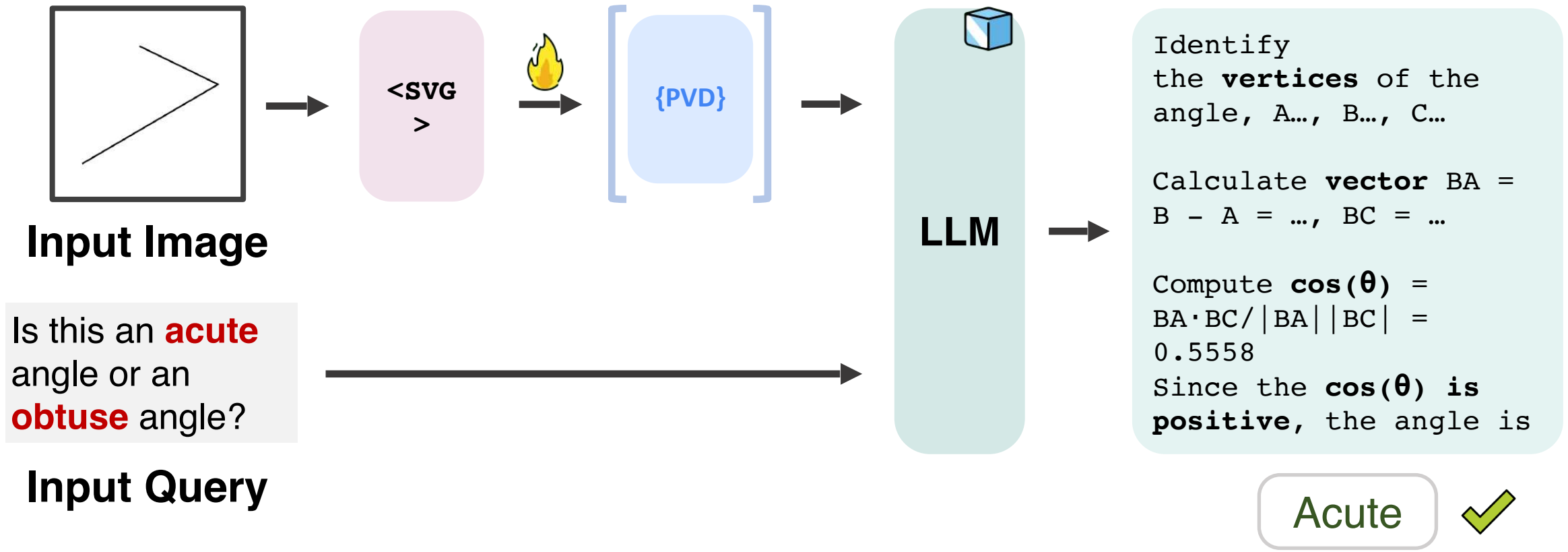


Thus, we **learn** an **intermediate symbolic representation**



Primal Visual Description (PVD)

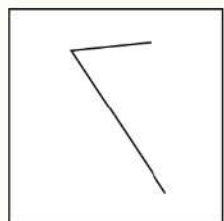
Primal Visual Description (PVD) enables **text-based reasoning with off-the-shelf LLMs**



Inference on Unseen Tasks

Zero-shot generalization to diverse tasks and domains

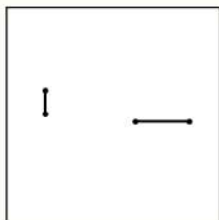
Zero-Shot Downstream Tasks



Is this an acute angle or an obtuse angle?

Acute

Angle Classification



Are the two lines of equal length?

No

Length Comparison



A yellow semicircle is to the left of a rectangle.

True

Shapeworld Spatial Reasoning (2Obj | MultiObj)



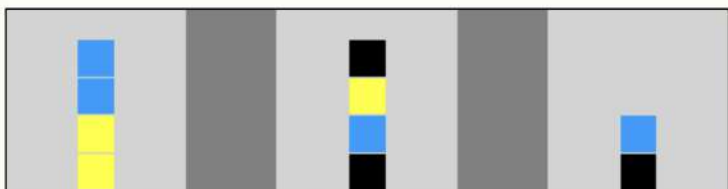
A cyan ellipse is to the right of a semicircle.

False

The rightmost shape is a magenta cross

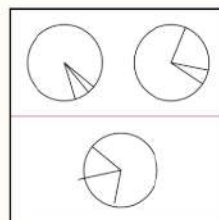
False

Shapeworld Superlative



There is 1 tower with 2 yellow blocks. **True**

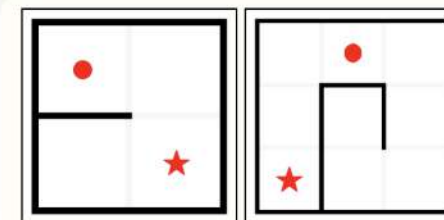
NLVR



Task: Determine if the test example (bottom) depicts the same concept as the two reference examples (top)

Target Output: **False**

Geoclidean 2-Shot Learning

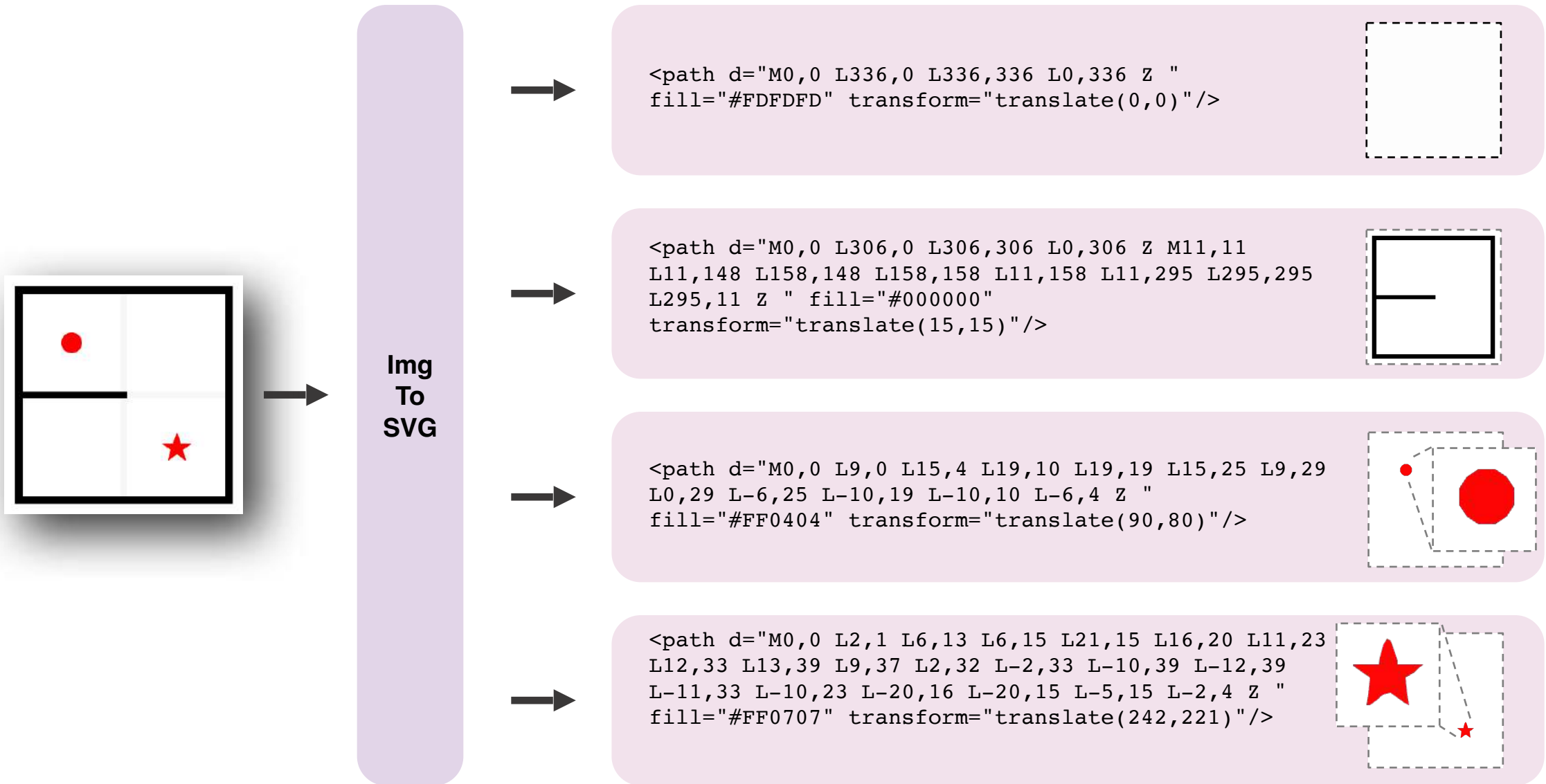


Task: Solve the maze

Target Output (2x2):
[(0,0), (0,1), (1,1)]

Maze Solving (2x2 | 3x3)

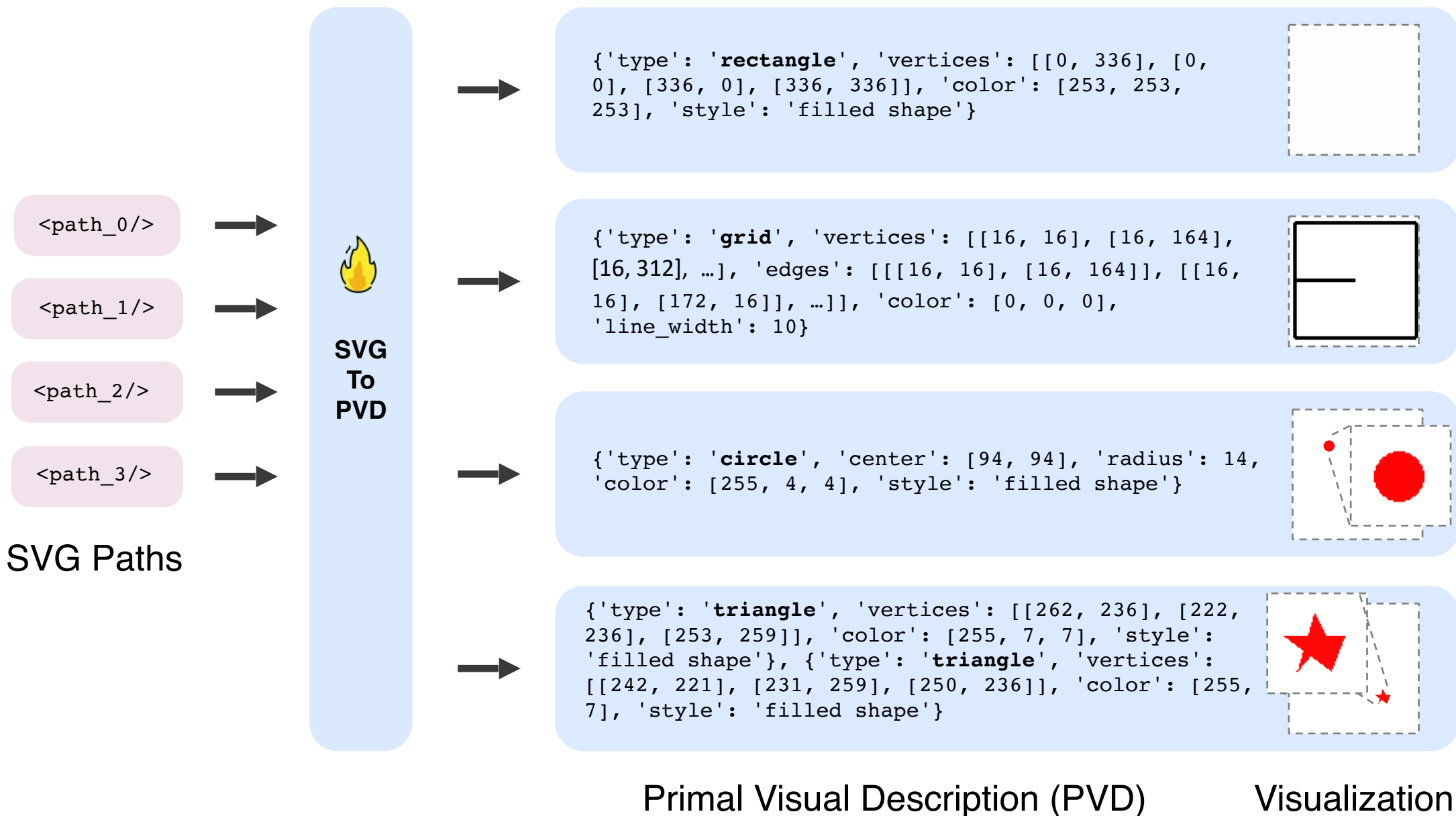
Step 1: Transform the image into SVG with a rule-based converter



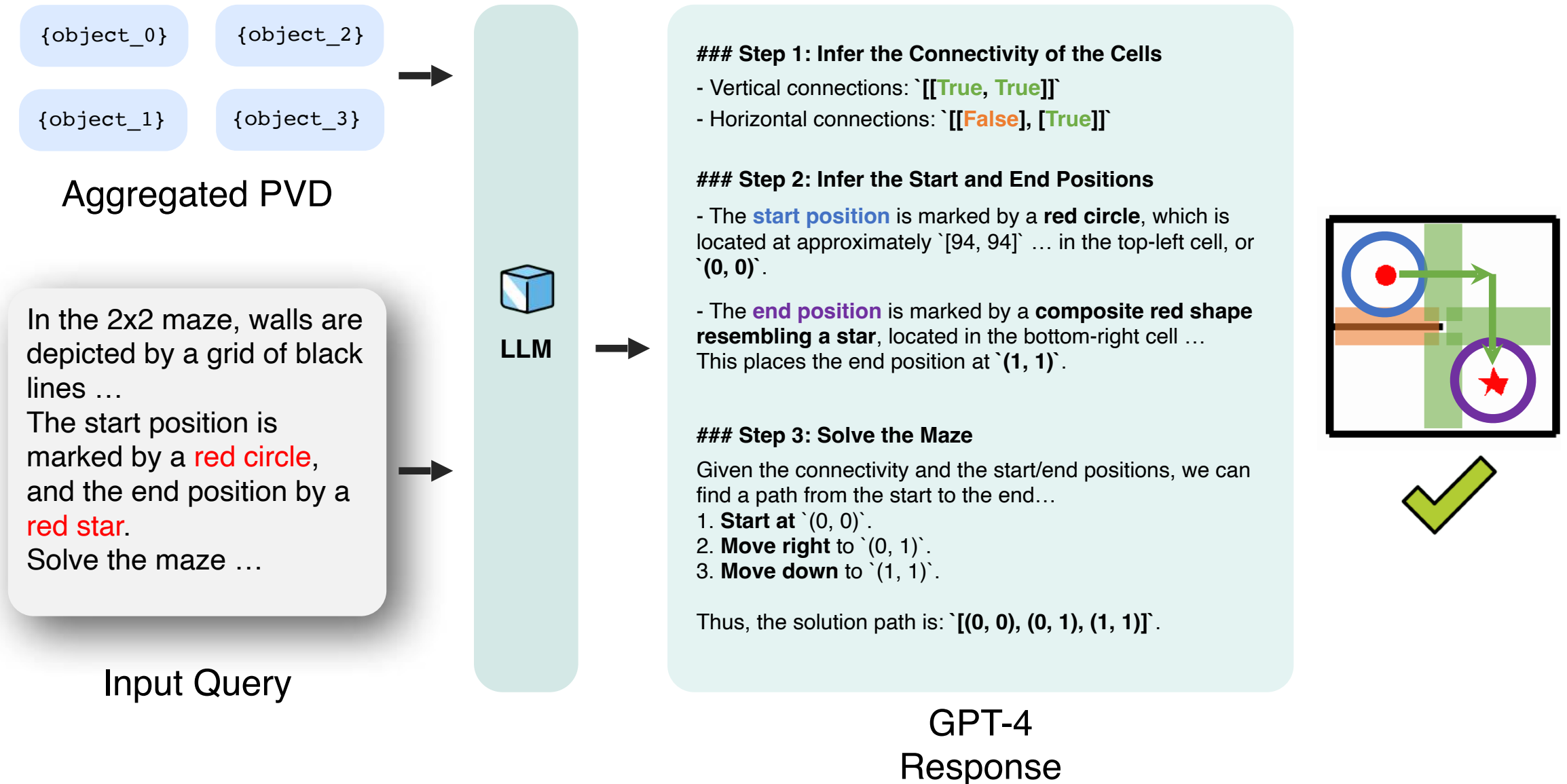
Decomposed Single SVG Paths

Visualization

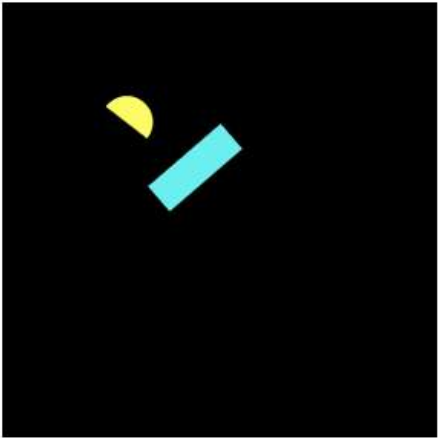
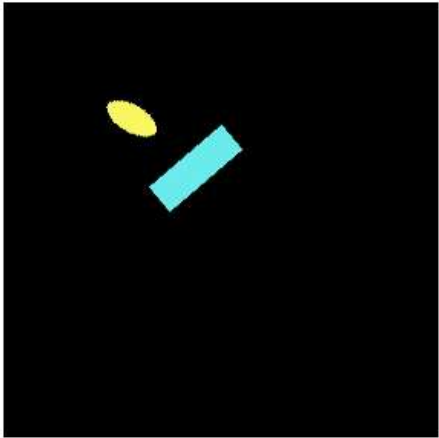
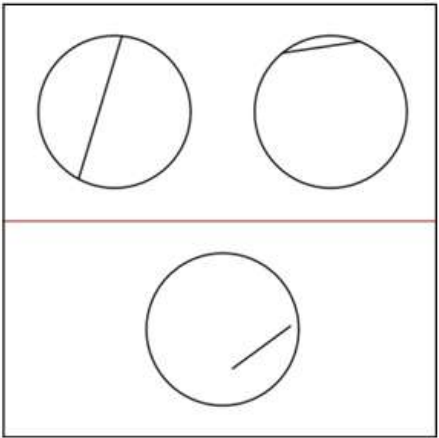
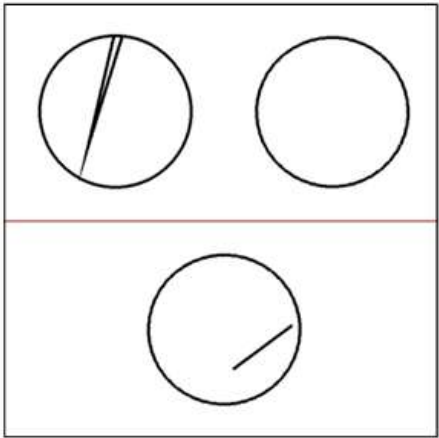
Step 2: Generate intermediate visual descriptions from SVG with a learned language model



Step 3: Reasoning about the task with an inference-only LLM



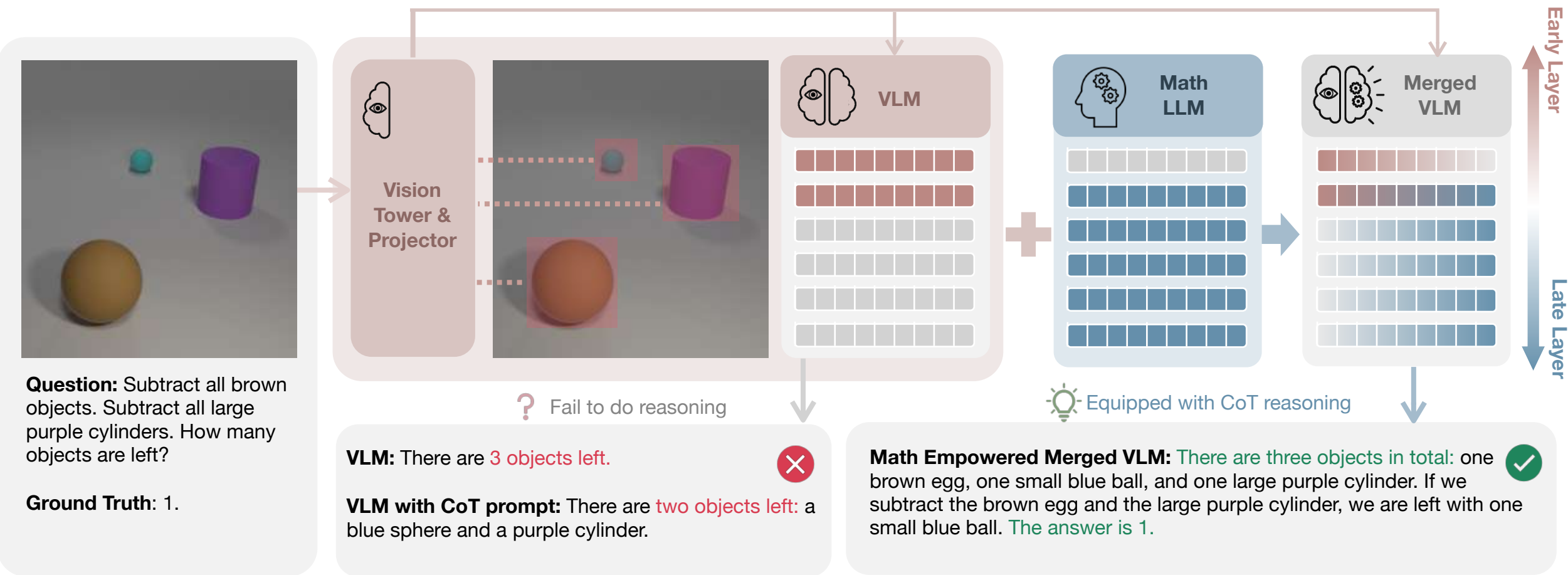
Limitations: Perception Errors

Error Type	Input Image	PVD Perception	PVD Perception Visualization
Novel shape (semicircle)		<pre>{... 'object_2': [{ 'type': 'ellipse', 'center': [99, 90], 'major_axis_length': 21, 'minor_axis_length': 10, 'rotation': 150, 'color': [249, 249, 62], 'style': 'filled shape' }]}</pre>	
Accurate constraints (circle segment)		<pre>{ 'object_0': [{ 'type': 'circle', ... }, { 'type': 'triangle', ... }], ... 'object_1': [{ 'type': 'ellipse', ... }, { <missing line_segment in the circle on the right> }] }</pre>	

Bring Reason to Vision: Understanding Perception and Reasoning through Model Merging

ICML 2025

Know where is perception layers via Model Merging

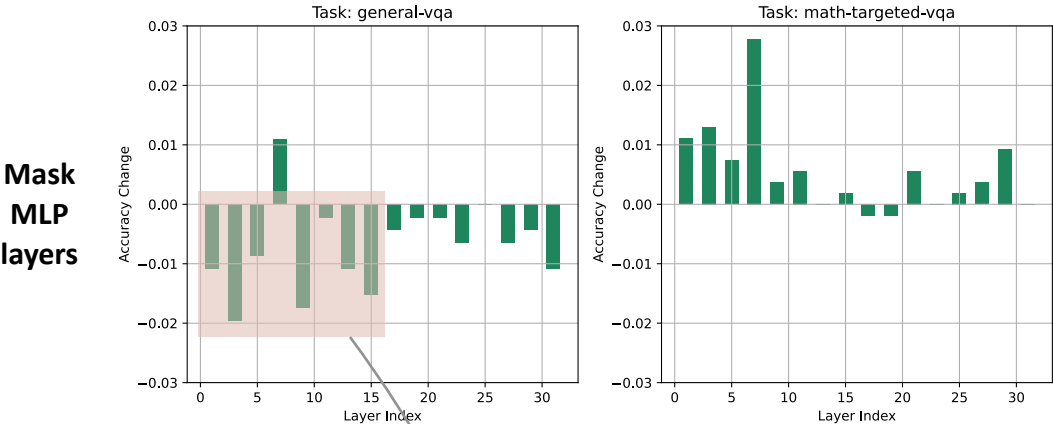


Perception: Early Layers, Reasoning: Later Layers

Where is the perception ability located?

①

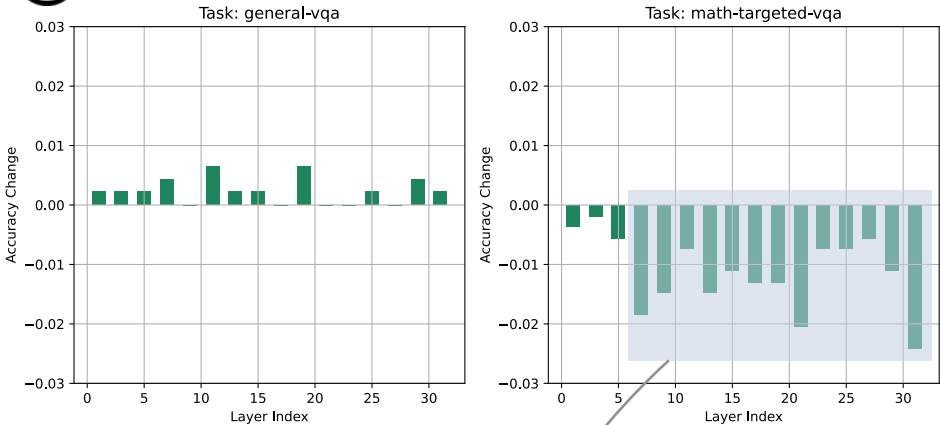
Accuracy Change by Blocking-Out certain MLP Layers



Where is the reasoning ability located?

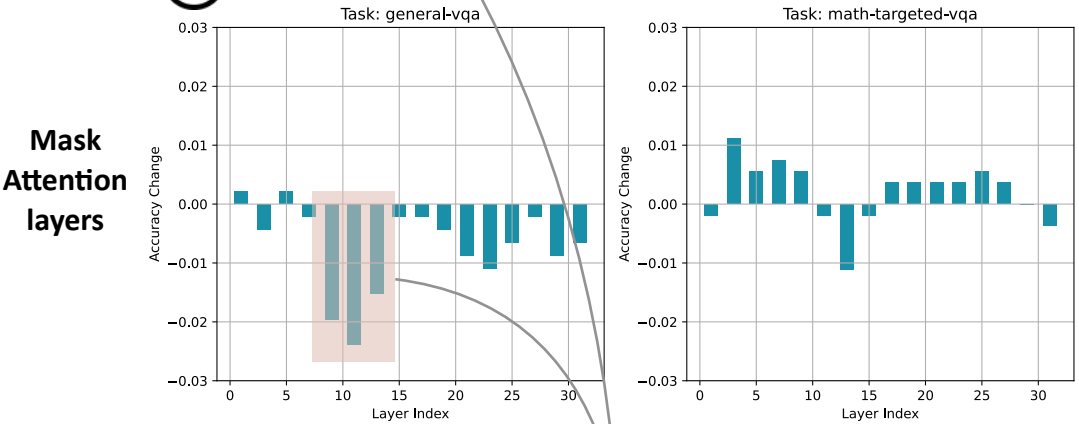
②

Accuracy Change by Blocking-Out certain MLP Layers



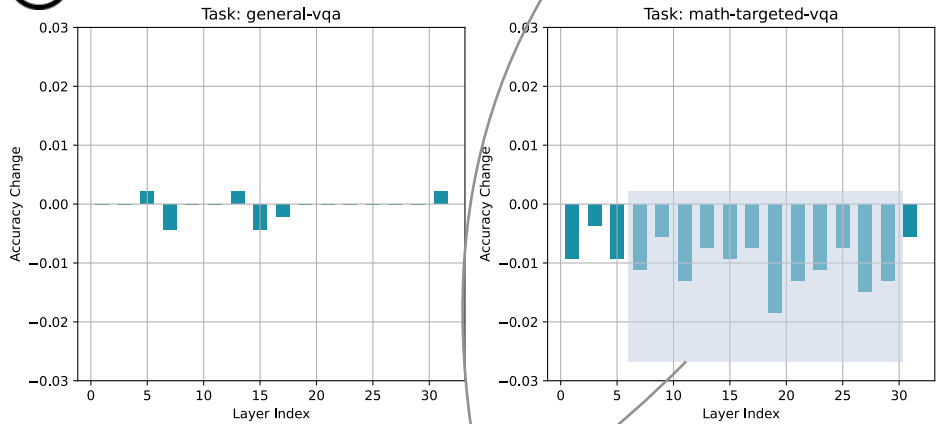
③

Accuracy Change by Blocking-Out certain Self Attention Layers



④

Accuracy Change by Blocking-Out certain Self Attention Layers



Perception: New Paradigm?

Theory of Space

How LLMs/VLMs develop Spatial Intelligence?

When a child enters an unfamiliar room...



Concept:

Q: Are there any **televisions**?
A: Yes

Q: Is there a **sofa** in the room with a **printer**?
A: Yes

Counting:

Q: How many **chairs** are **close** to the **table** in the room with **plants** on the **cabinet**? A: 6

Q: How many rooms have **sofas**? A: 1

Relation:

Q: **Facing** the **computer** from the **curtain**, is there a **lamp** on the **right**? A: Yes

Q: What's on the **cabinet** in the **smaller** room? A: Plant

Comparison:

Q: Are there **fewer** **pictures** in the **larger** room than the other room? A: No

Q: Is the **computer** **closer** to a **printer** or a **lamp**?
A: Printer



They are not executing a plan like finding a specific toy

Curiosity-driven approach

When a child enters an unfamiliar room...

Passive Exploration







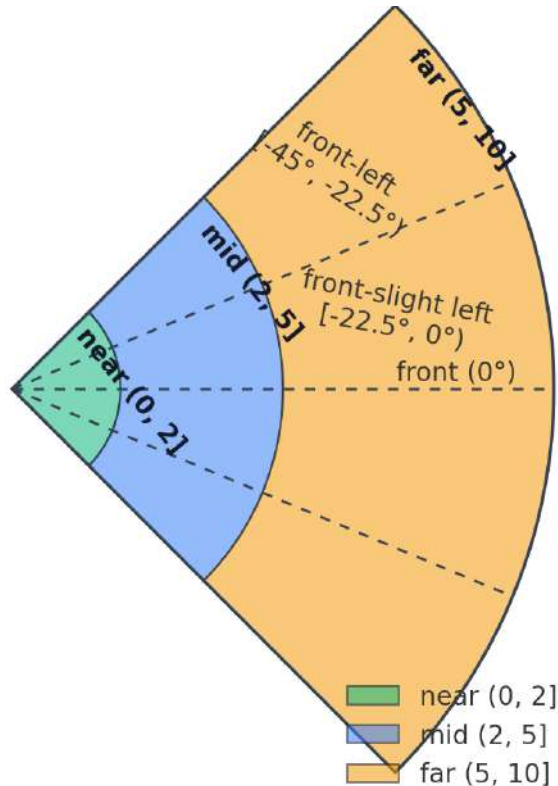
Active Exploration

They are not executing a plan like finding a specific toy

Curiosity-driven approach

Exploration — Actions

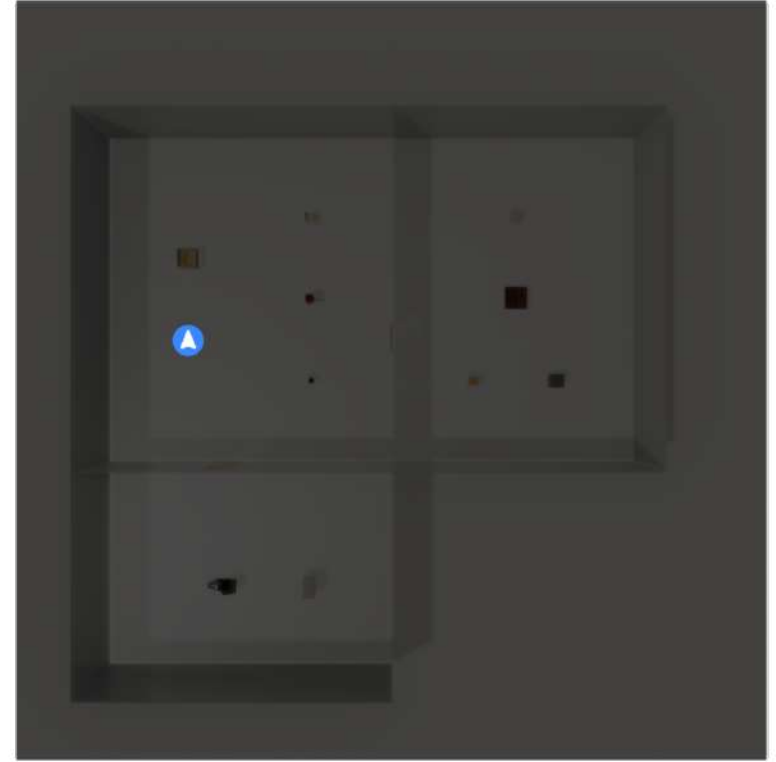
1. Move(A): move to object A 
2. Rotate(D): rotate D degree, D in [0, 90, 180, 270] 
3. Observe(): observe the objects in field of view (fov) 
 - a. Direction: [-45, 22.5) front left, [22.5, 0) front-slight left, 0 front
 - b. Distance: (0, 2] near, (2, 5] mid, (5, 10] far, ...
4. Term(): Terminate the exploration stage 



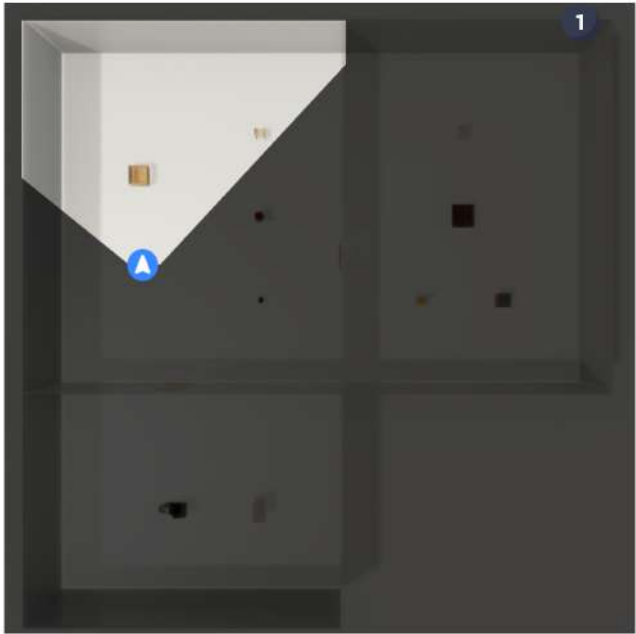
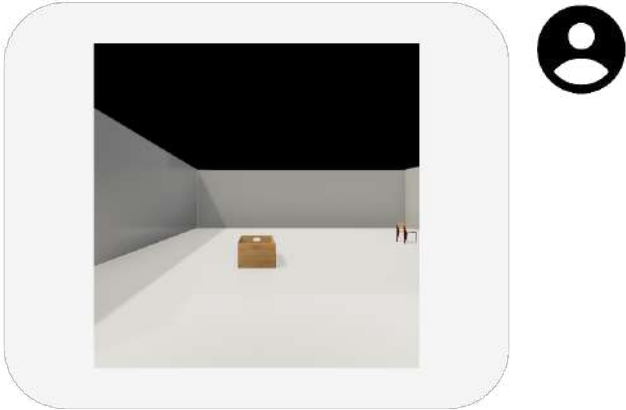
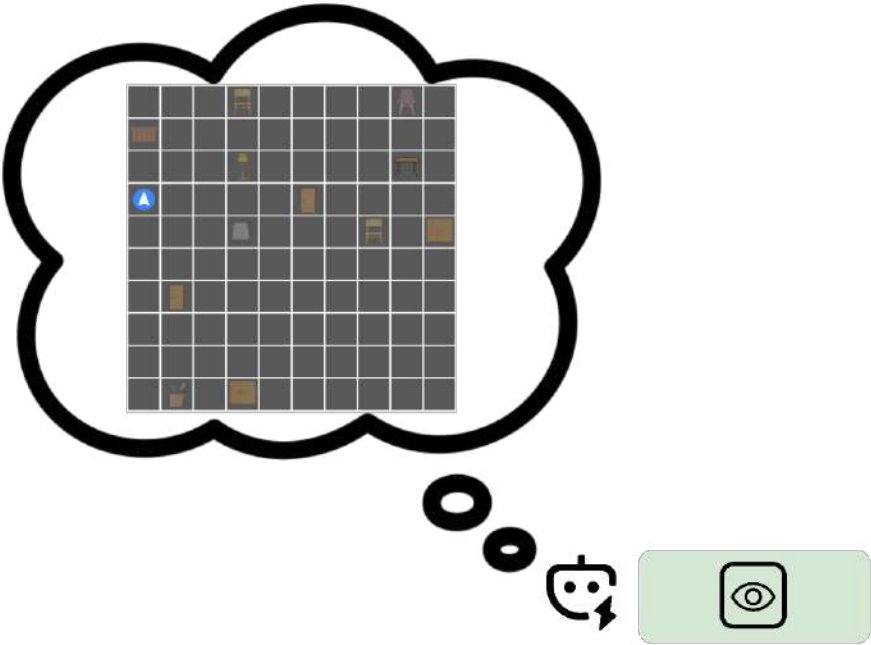
Text: You observe a basket at your front, mid distance

Vision:

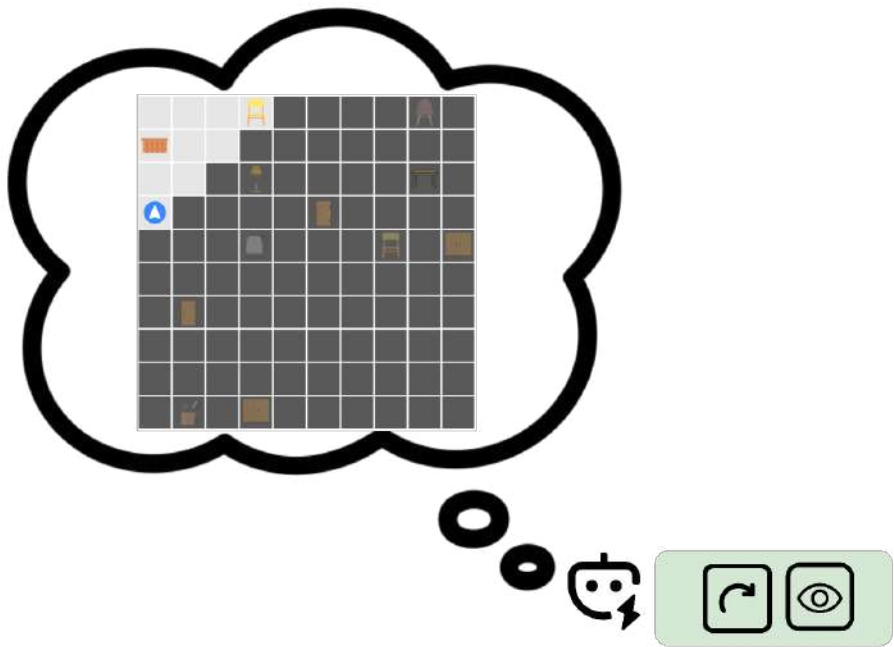




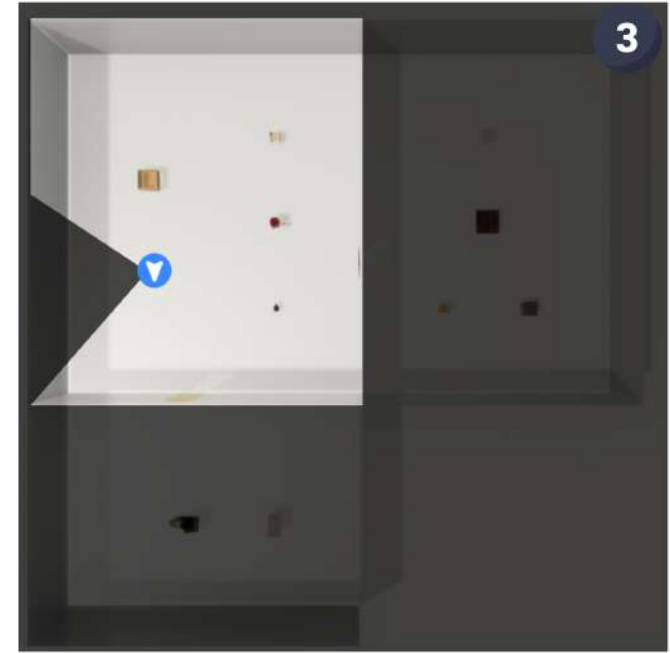
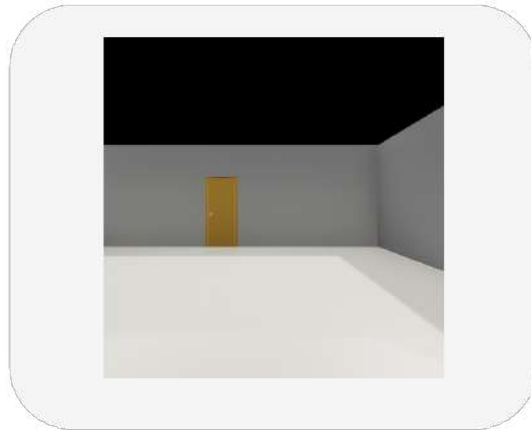
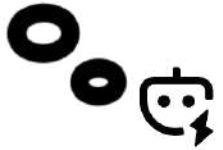
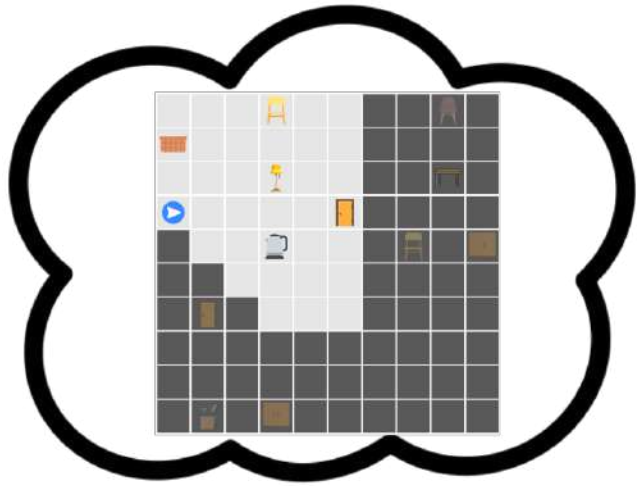
Exploration



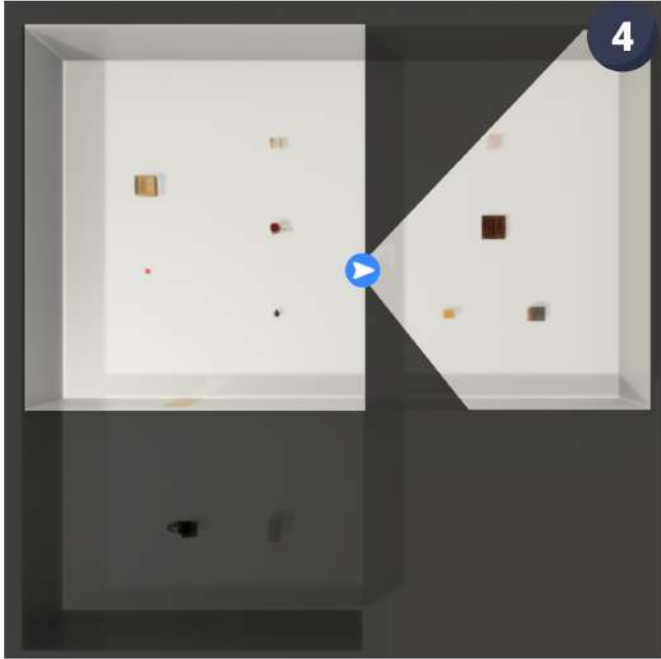
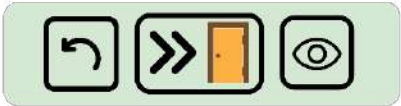
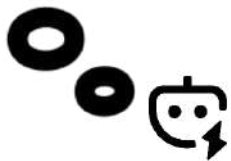
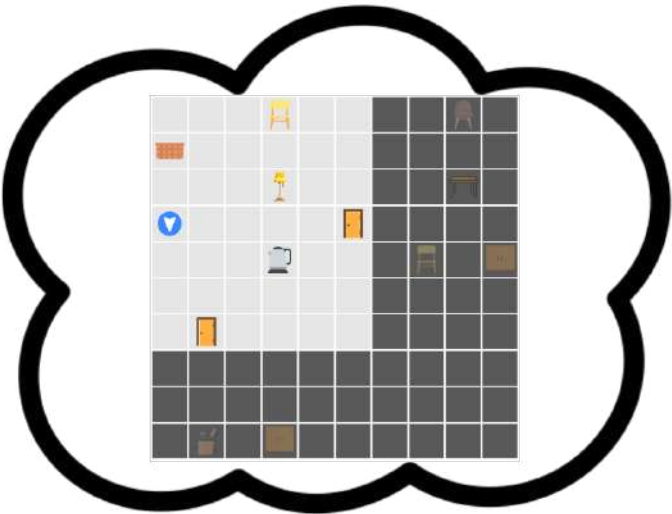
Exploration



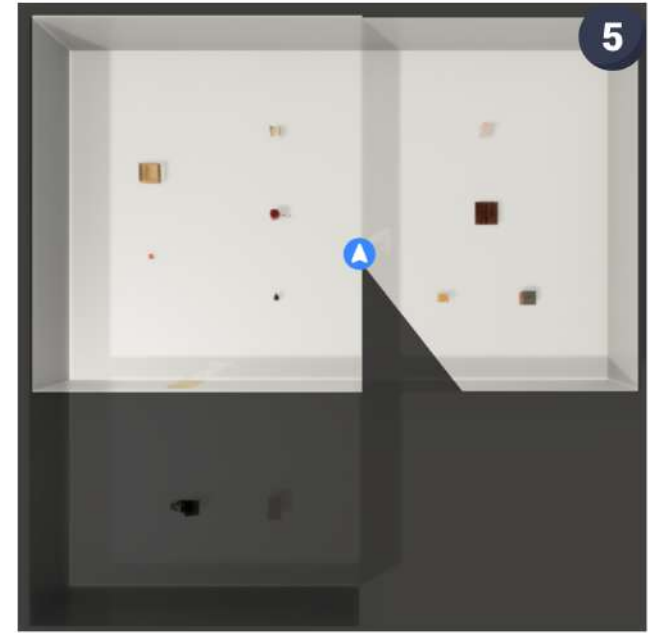
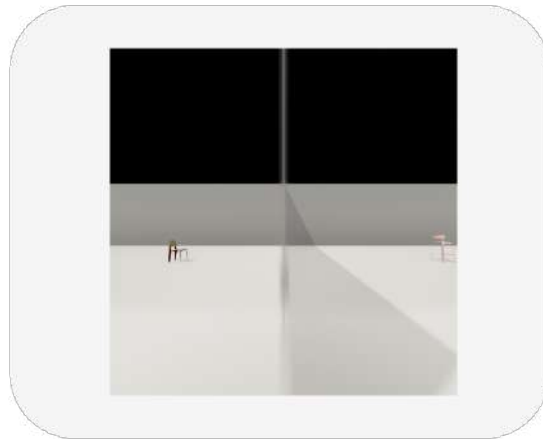
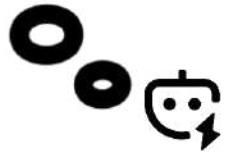
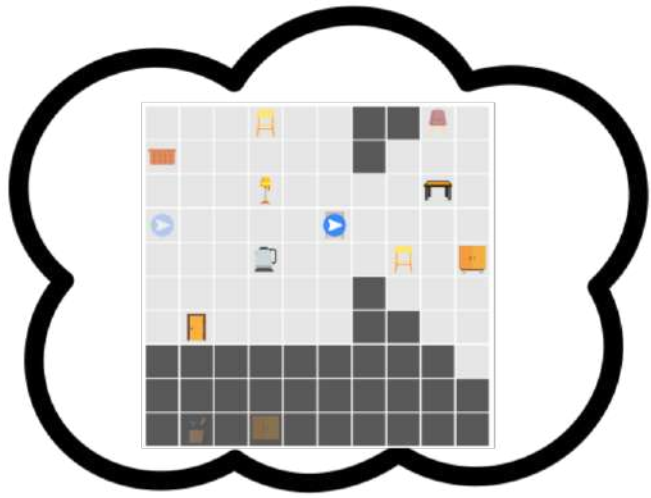
Exploration



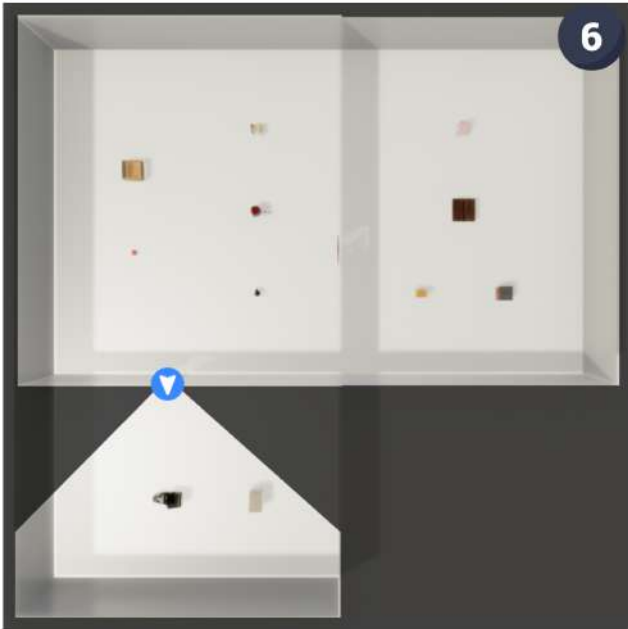
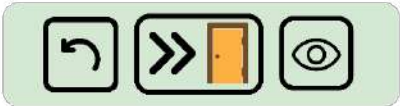
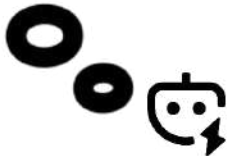
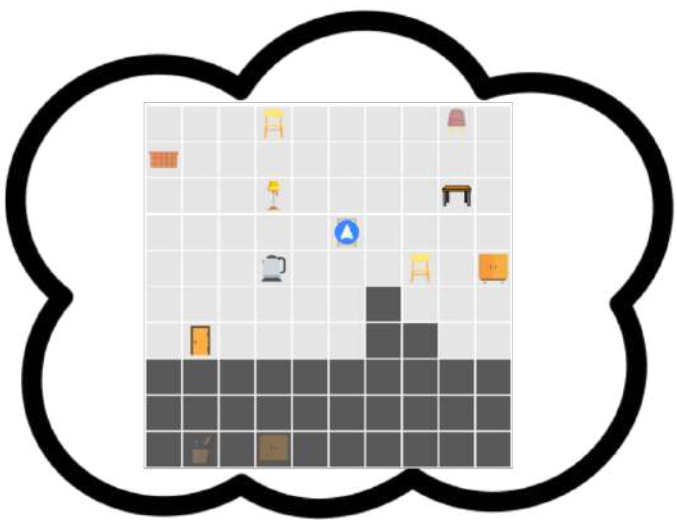
Exploration



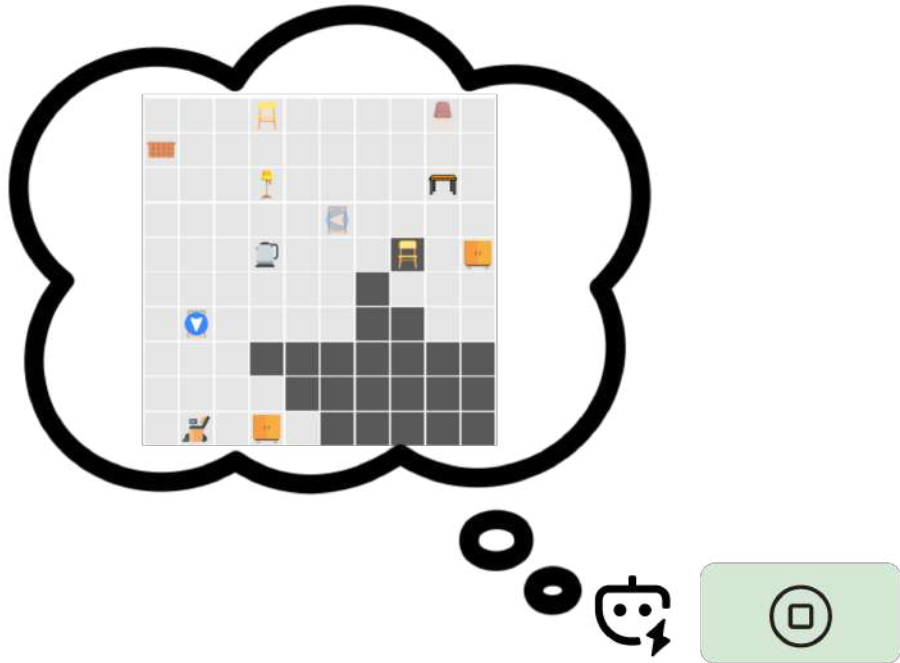
Exploration



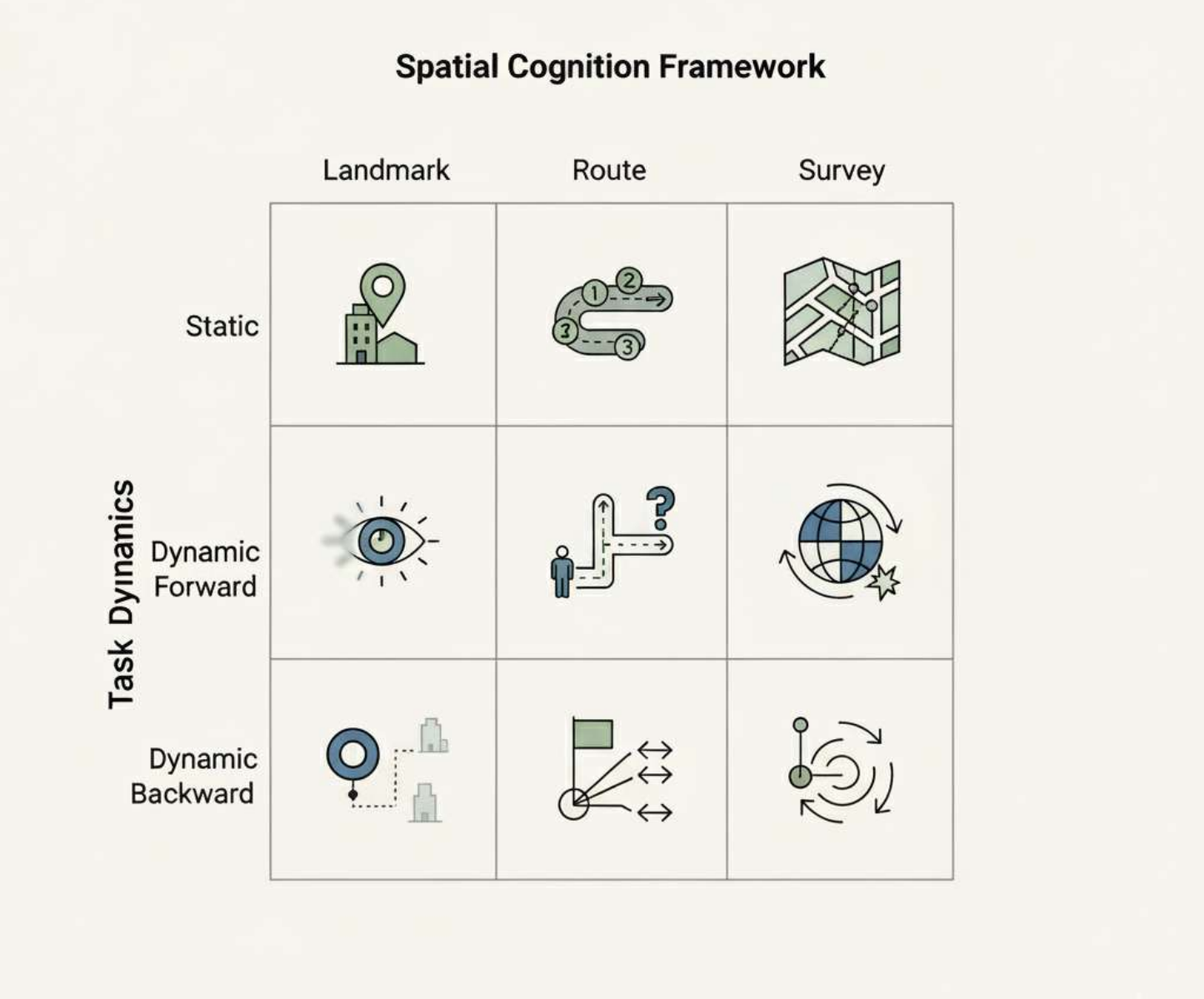
Exploration



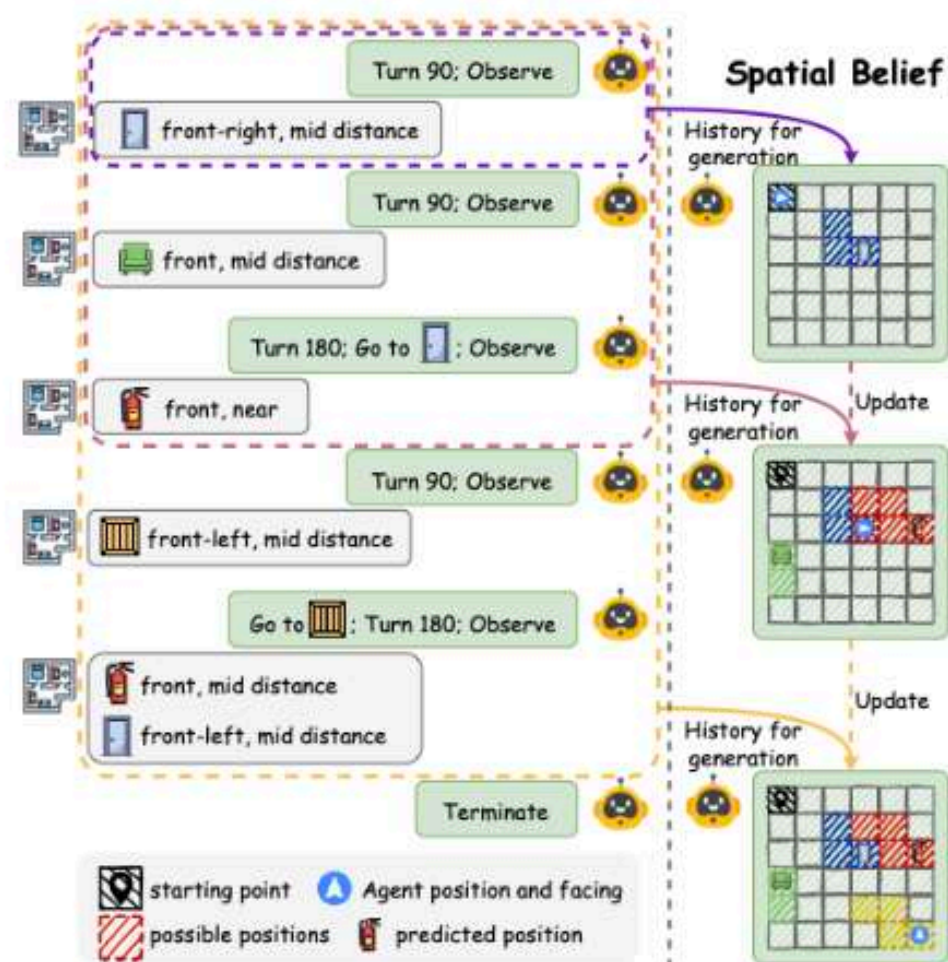
Exploration



Evaluating at each interaction step:

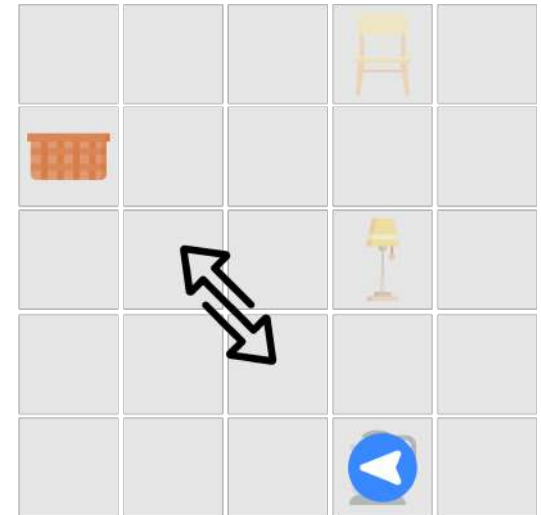
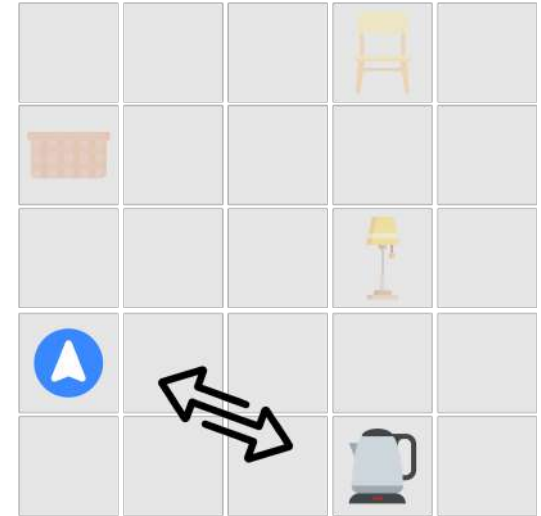
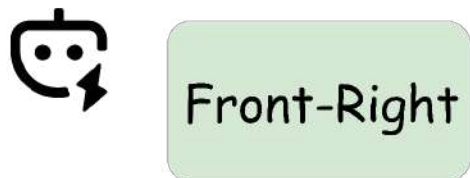
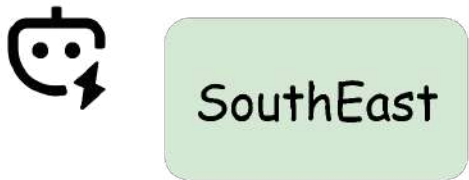


Evaluating at each interaction step:



Evaluation — Route, Static

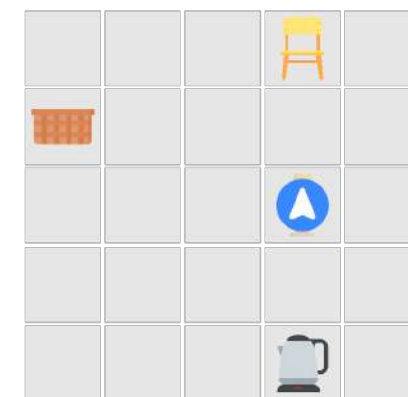
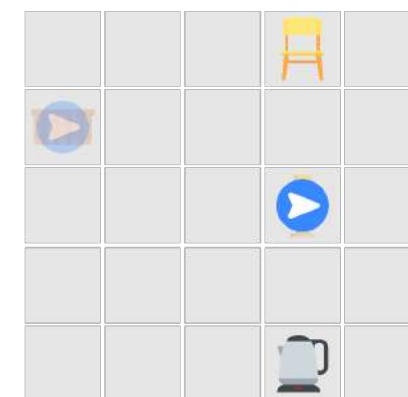
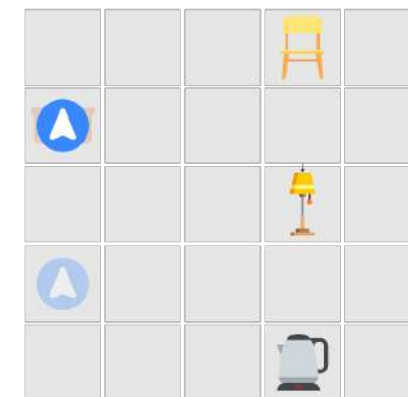
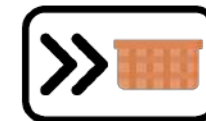
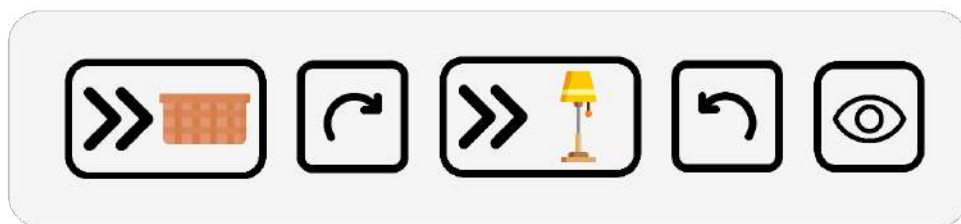
1. Directional Relationship (Dir): Allocentric pairwise relation reasoning anchored to a fixed global frame.
2. Perspective Taking (PT): Egocentric pairwise relation reasoning anchored to a viewpoint-local frame defined by a given facing rule.



Evaluation — Route, Dynamic

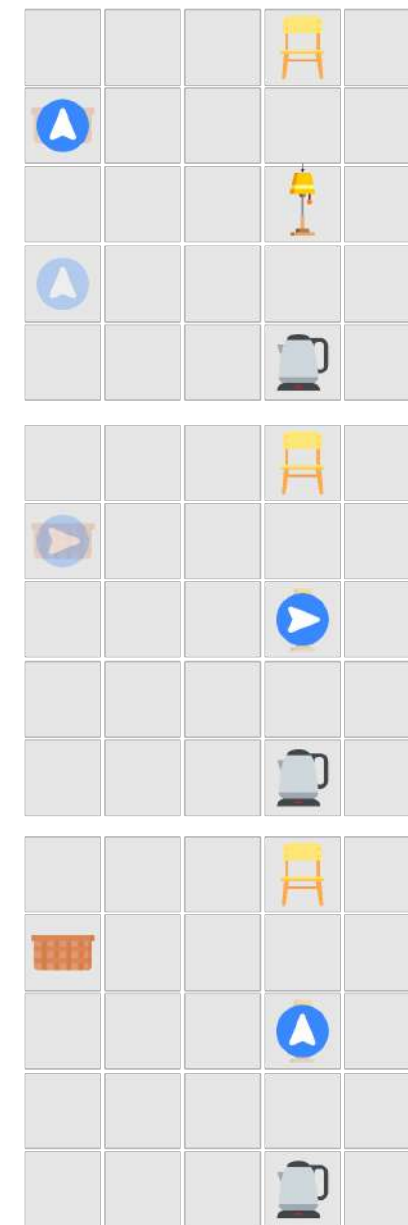
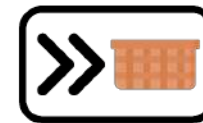
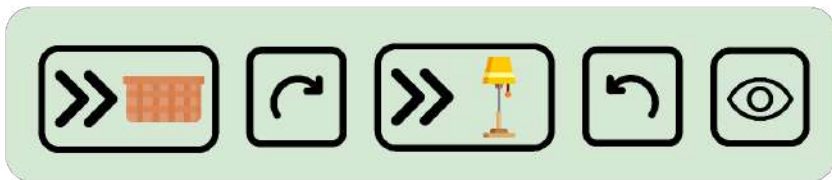
1. Forward (Dynamic)

- Tests: Egocentric simulation (predict next observation under discrete actions).
- Setup: Given an action sequence composed of moves (between objects) and rotations, predict the resulting observation (objects in FOV with egocentric bins).



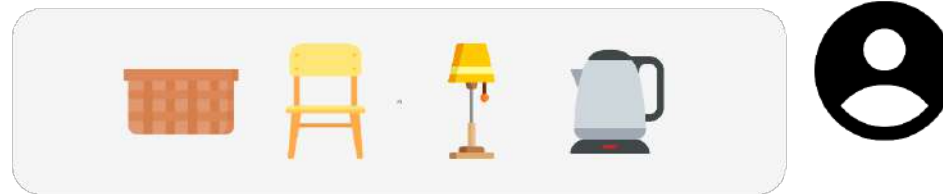
Evaluation — Route, Dynamic

1. Backward (Dynamic)
 - a. Tests: Inverse egocentric inference (recover discrete action sequence from a final view).
 - b. Setup: Given a final observation, choose the correct object-to-object action sequence and discrete rotations that reach the goal view.

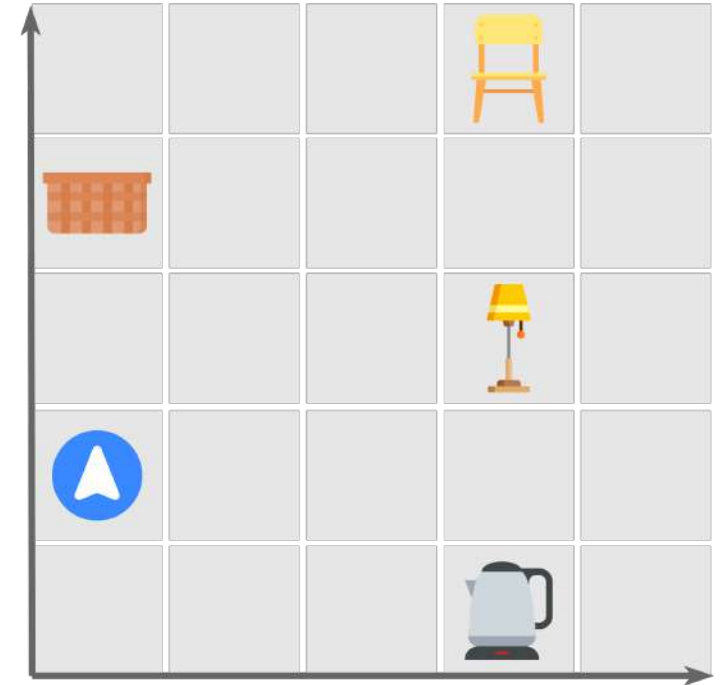


Evaluation — Survey, Static

1. Allocentric Mapping (AM):
 - a. Tests: Global, self-independent mapping (cognitive map construction).
 - b. Setup: Predict object coordinates and headings in the global frame.

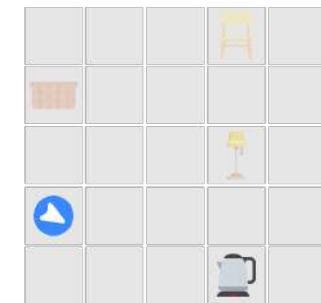
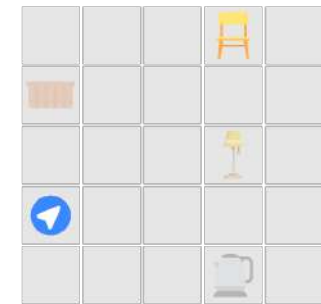


$(0,3), (3,4), (2,3), (3,0)$



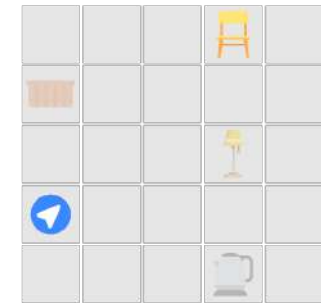
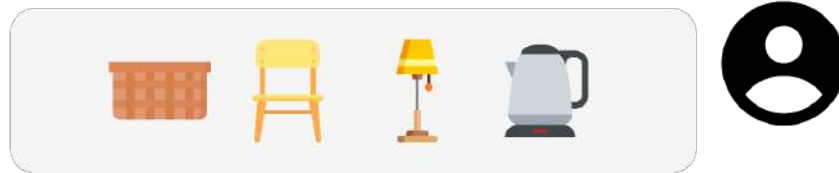
Evaluation — Survey, Dynamic

1. Mental Rotation (MR, Forward):
 - a. Tests: Egocentric mental transformation in a global map.
 - b. Setup. Given the current pose, imagine rotating the agent around and report which object would be straight ahead.



Evaluation — Survey, Dynamic

1. SpinDecide (Backward): Determine rotation direction given an object sequence

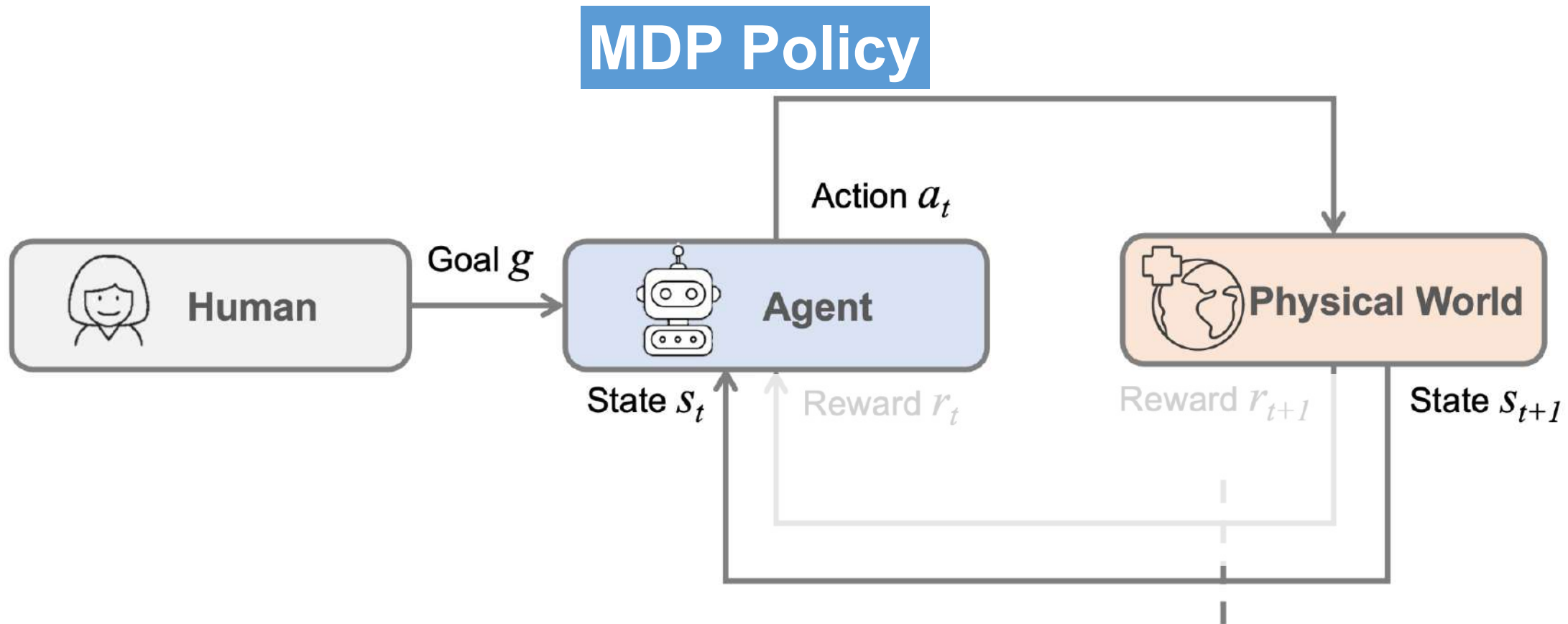


What we discovered?

Current models do not know "when to stop"

Current models do not know "when to go to a new room"

Let us go back to MDPs (Markov Decision Processes)



Tutorial on Foundation Models Meet Embodied Agents

<https://foundation-models-meet-embodied-agents.github.io/>



Manling Li
Northwestern



Yunzhu Li
Columbia



Jiayuan Mao
MIT



Wenlong Huang
Stanford



Northwestern
University



COLUMBIA



Stanford
University

Challenge on Foundation Models Meet Embodied Agents

<https://foundation-models-meet-embodied-agents.github.io/>

Foundation Models Meet Embodied Agents

Tutorials Workshops Challenges

🏆 EAI x BEHAVIOR: Co-Hosted at NeurIPS!

September 1 2025

We're thrilled to announce that the BEHAVIOR Challenge is joining forces with the Embodied Agent Interface Challenge at this year's NeurIPS Competition Track. Two challenges, one stage — bringing richer benchmarks, diverse tasks, and a united embodied AI community. [Learn more about the BEHAVIOR Challenge here!](#)

🚀 EAI Challenge Launch!

August 15 2025

The EAI Challenge officially kicks off at 12:00 PM (CDT)! We are thrilled to welcome all participants and can't wait to see your innovative solutions. Check out our challenge on [EvalAI](#) and the [Participate](#) section for all the details and resources you need to get started. Good luck to everyone!

🔧 Beta Testing Phase

June 30 2025

The beta test for our competition platform, Eval AI, is now underway! Get ready for the official launch and public registration in late July or early August. Please stay tuned for more updates!



Fei-Fei Li @drfeifei · Sep 2

(1/N) How close are we to enabling robots to solve the long-horizon, complex tasks that matter in everyday life?

🌟 We are thrilled to invite you to join the 1st BEHAVIOR Challenge @NeurIPS 2025, submission deadline: 11/15.

🏆 Prizes:
🥇 \$1,000
🥈 \$500
🥉 \$300

How far are we from robots
that can perform everyday tasks?



Pantheon-CLI

Your own analyst, why not try now?



```
Aristotle • 2025
PANTHEON

We're not just building another CLI tool.
We're redefining how scientists interact with data in the AI era.
Pantheon-CLI is a research project, use with caution.

-- MODEL -----
• gpt-5

-- HELP -----
• /exit    to quit
• /help    for commands
• /model   for available models
• /api-key for API keys

-- CONTROL -----
Use ↑/↓ arrows for command history
Enter your message (press Enter twice to finish)
> 
```

pantheonos.stanford.edu



Xiaojie Qiu



MLL Lab

Machine Learning and Language

We develop intelligent language + X (vision, robotics, etc) models that reason, plan, and interact with the physical world.

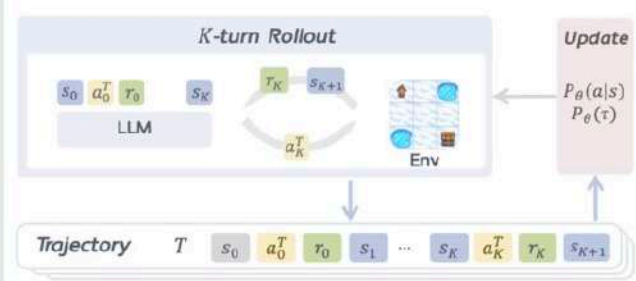
[Join Us](#)

Announcing **RAGEN**: Training RL Agents - Github 1.6k >



RAGEN

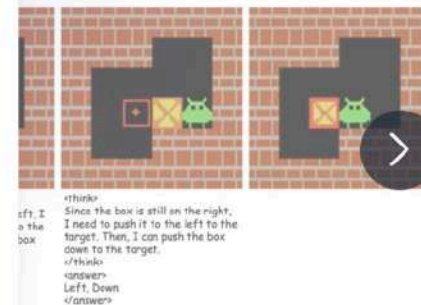
StarPO (State-Thinking-Actions-Reward Policy Optimization)



Dynamic Tasks

- Single-turn Stochastic: Bandit
- Multi-turn Non-Stochastic: Sokoban
- Multi-turn Stochastic: Frozen Lake

GEN



LLM Agent + multi-turn RL

Thank You