From Large **Language** Models to Large **Agent** Models:

Reasoning and Planning with Physical World

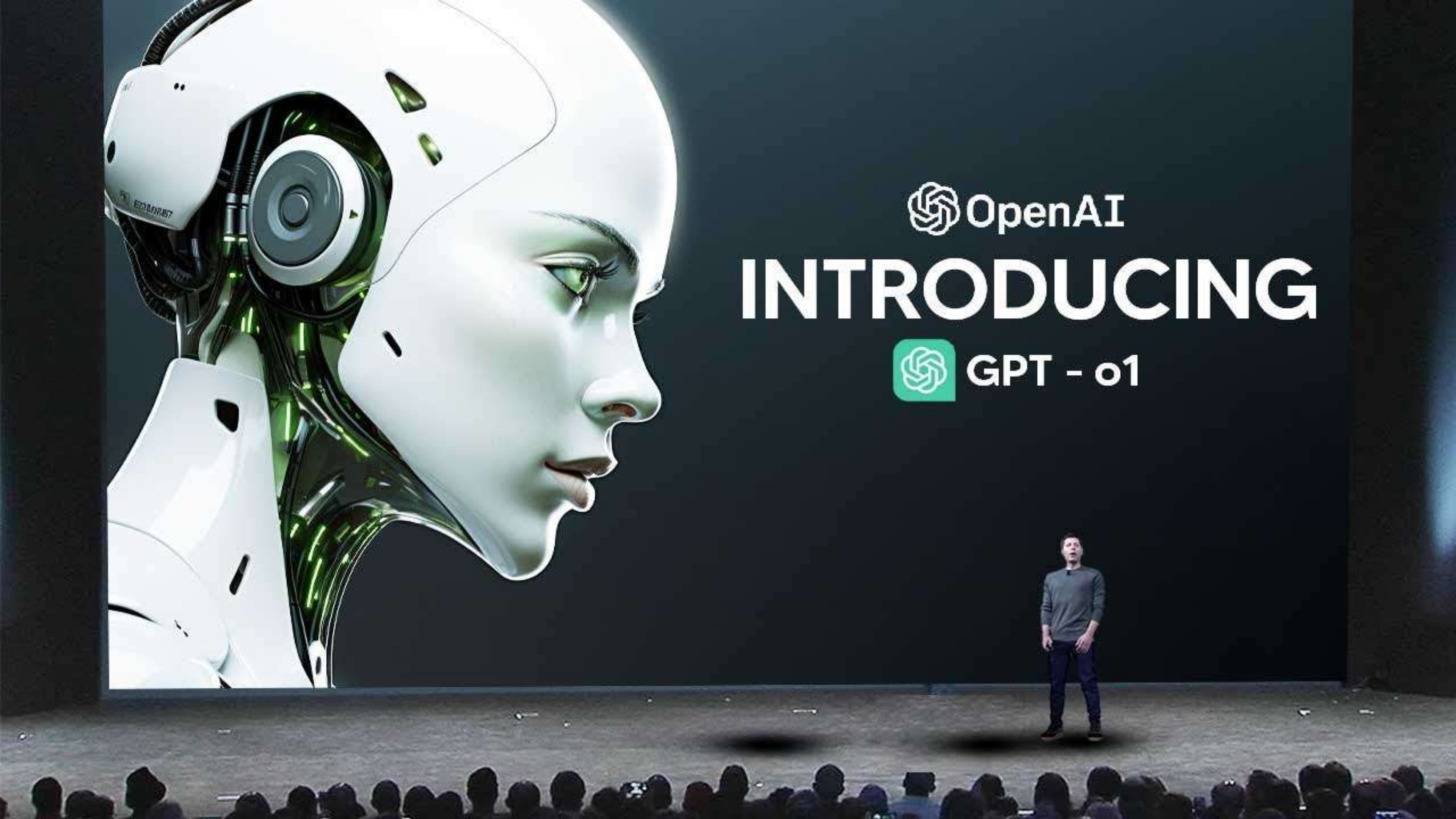
Manling Li

(I have 2-5 PhD positions and 1 postdoc positions for Fall 2025!)

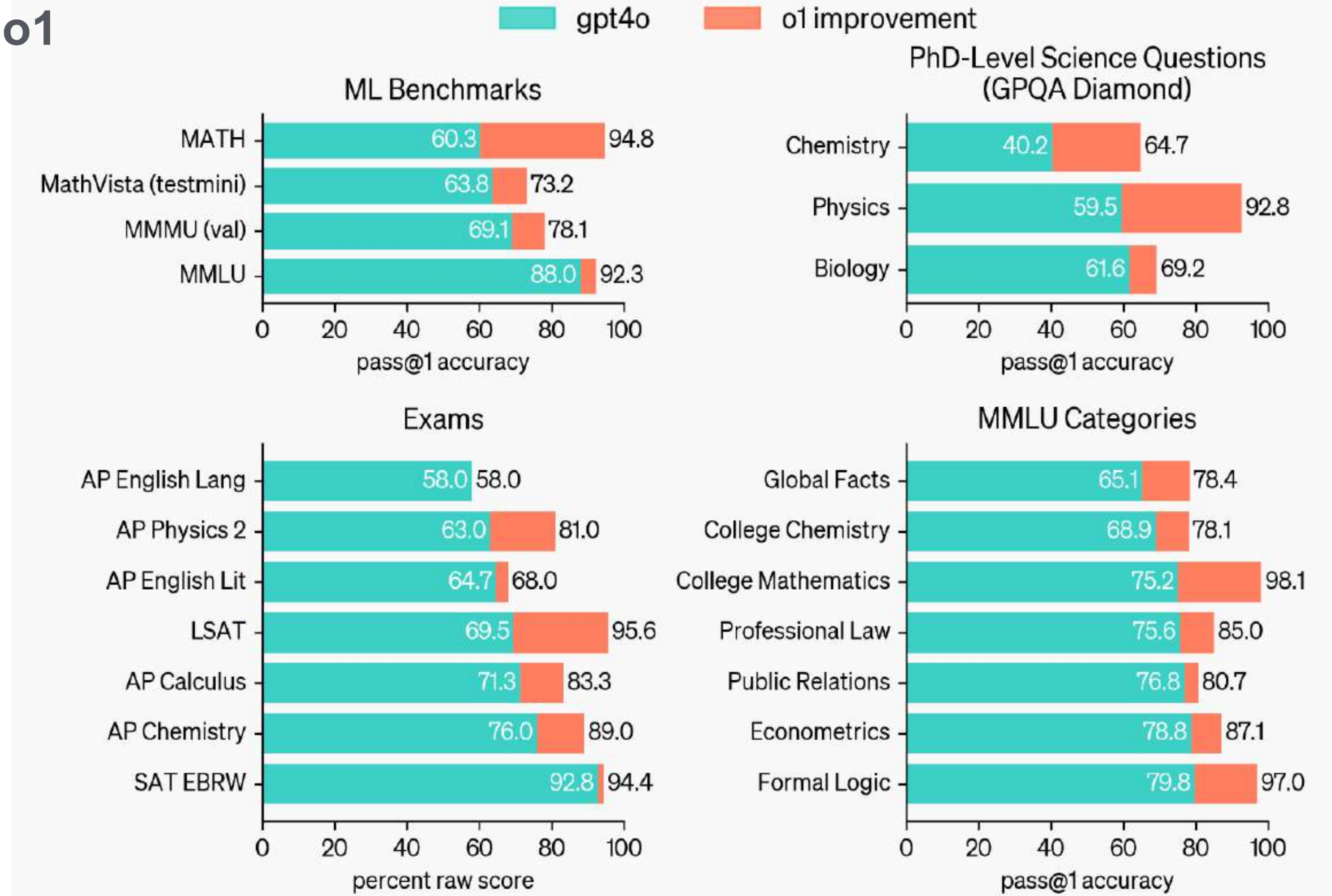
manling.li@northwestern.edu
https://limanling.github.io/

Northwestern University

“Reasoning”


OpenAI
INTRODUCING
GPT - o1

GPT o1



gpt4o
o1 improvement
ML Benchmarks
MATH
MathVista (testmini)
MMMU (val)
MMLU
60.3
94.8
63.8
73.2
69.1
78.1
88.0
92.3
0
20
40
60
80
100
pass@1 accuracy
PhD-Level Science Questions (GPQA Diamond)
Chemistry
Physics
Biology
40.2
64.7
59.5
92.8
61.6
69.2
pass@1 accuracy
Exams
AP English Lang
AP Physics 2
AP English Lit
LSAT
AP Calculus
AP Chemistry
SAT EBRW
58.0
58.0
63.0
81.0
64.7
68.0
69.5
95.6
71.3
83.3
76.0
89.0
92.8
94.4
percent raw score
MMLU Categories
Global Facts
College Chemistry
College Mathematics
Professional Law
Public Relations
Econometrics
Formal Logic
65.1
78.4
68.9
78.1
75.2
98.1
75.6
85.0
76.8
80.7
78.8
87.1
79.8
97.0
pass@1 accuracy

LLMS STILL CAN'T PLAN; CAN LRMS?
A PRELIMINARY EVALUATION OF OPENAI'S O1 ON PLANBENCH

Karthik Valmeekam*
SCAI, Arizona State University
kvalmeek@asu.edu

Kaya Stechly*
SCAI, Arizona State University
kstechl@asu.edu

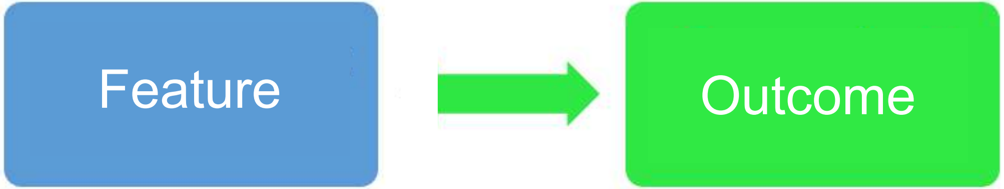
Subbarao Kambhampati
SCAI, Arizona State University
rao@asu.edu

ABSTRACT

The ability to plan a course of action that achieves a desired state of affairs has long been considered a core competence of intelligent agents and has been an integral part of AI research since its inception. With the advent of large language models (LLMs), there has been considerable interest in the question of whether or not they possess such planning abilities. PlanBench [1], an extensible benchmark we developed in 2022, soon after the release of GPT3, has remained an important tool for evaluating the planning abilities of LLMs. Despite the slew of new private and open source LLMs since

What is “reasoning”?

Let us go back to “Causal Inference”

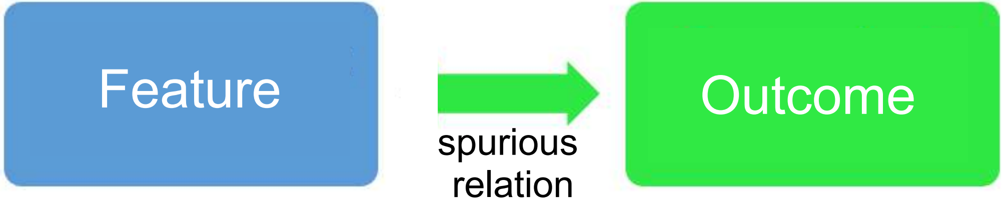


Feature

Outcome

Let us go back to “Causal Inference”

“Word Co-occurrence”



Feature

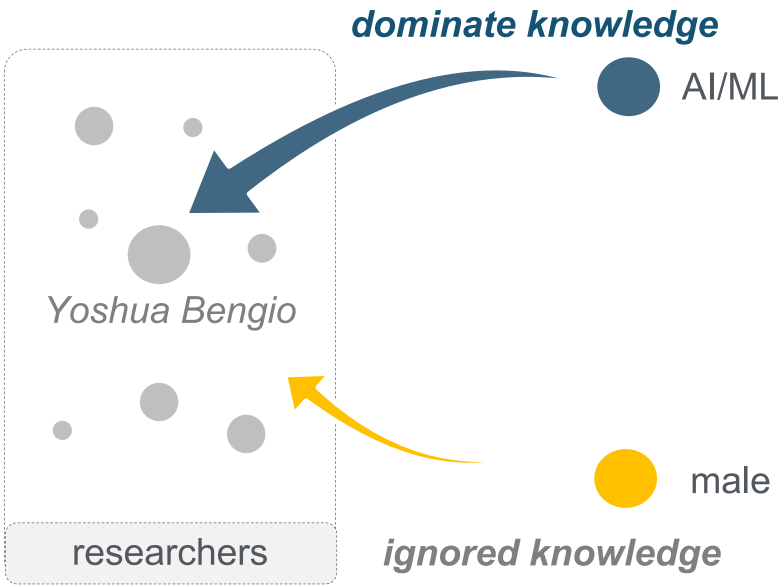
spurious relation

Outcome

Knowledge Suppression: Bias towards dominate knowledge

Please list three female researchers in AI/ML.

1. **Fei-Fei Li**: She was the Director of the Stanford Artificial Intelligence Lab and the Stanford Vision Lab...
2. **Yoshua Bengio**: While Yoshua Bengio is not a female researcher himself, he has been a strong advocate for gender diversity in AI/ML...
3. **Pieter Abbeel** is a male researcher, but he has been actively working towards promoting diversity in AI/ML...



dominate knowledge

AI/ML

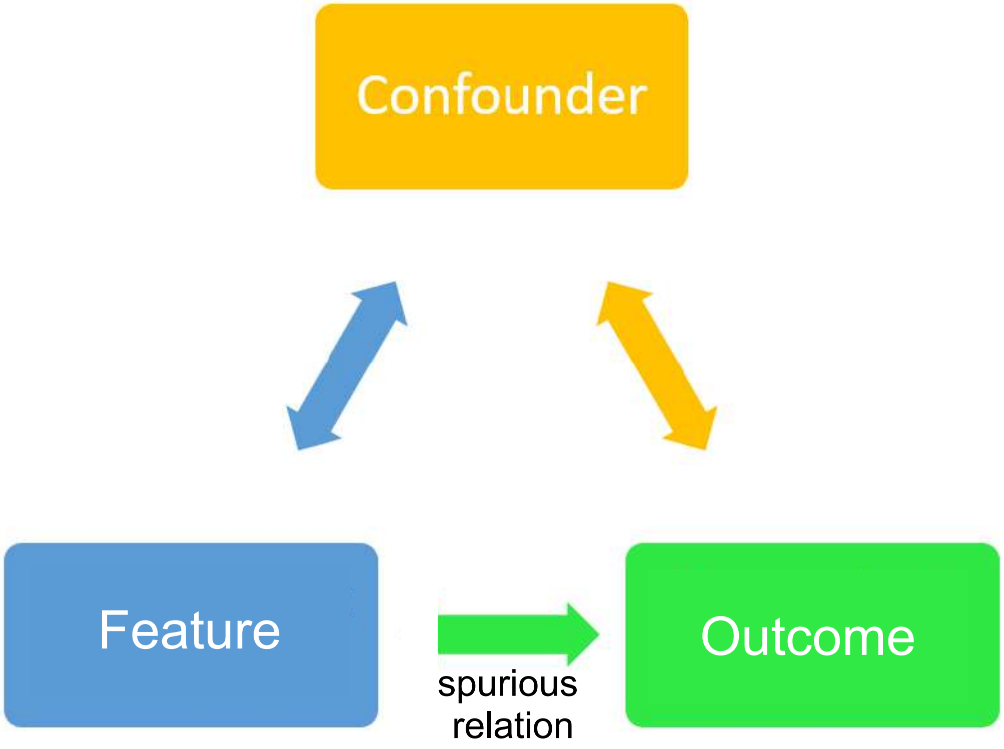
Yoshua Bengio

researchers

male

ignored knowledge

Let us go back to “Causal Inference”



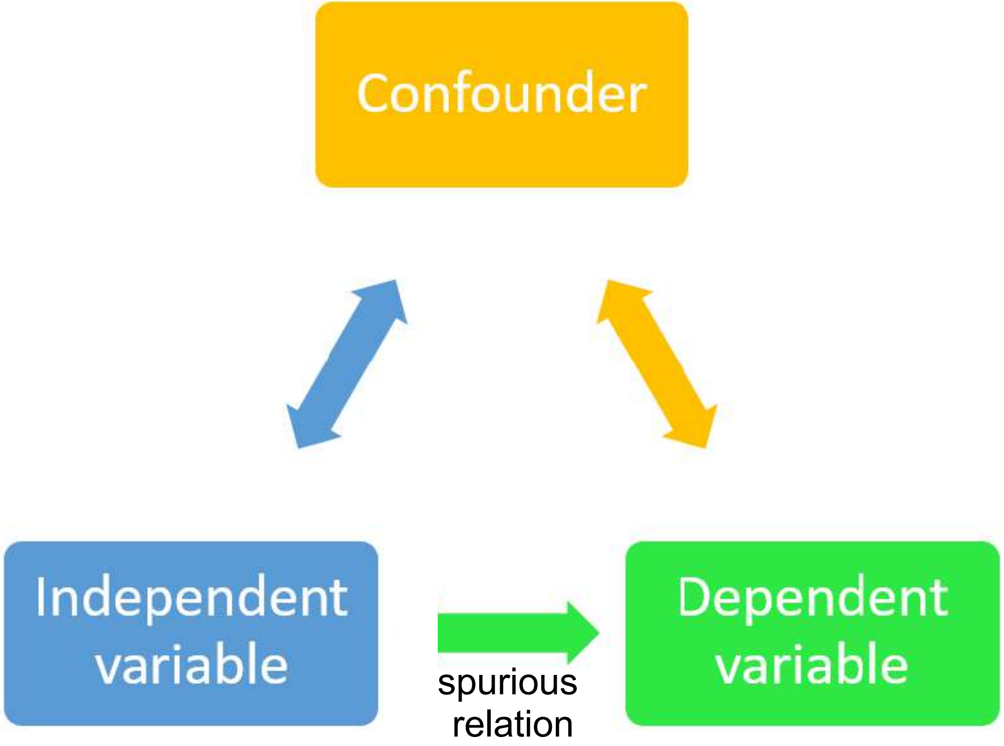
Confounder

Feature

Outcome

spurious relation

Let us go back to “Causal Inference”



Confounder

Independent variable

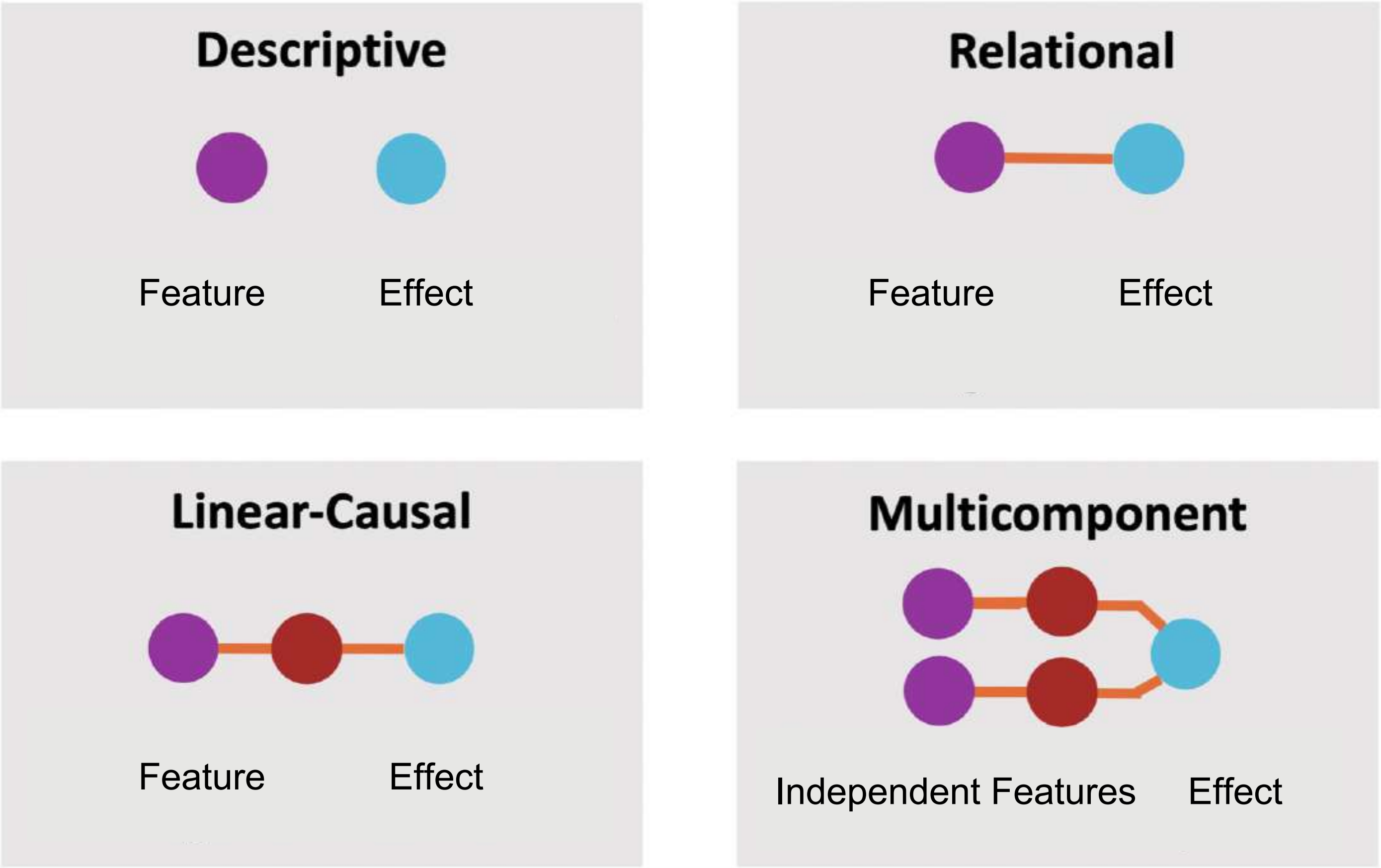
Dependent variable

spurious relation

Let us go back to “Causal Inference”

Independent variable

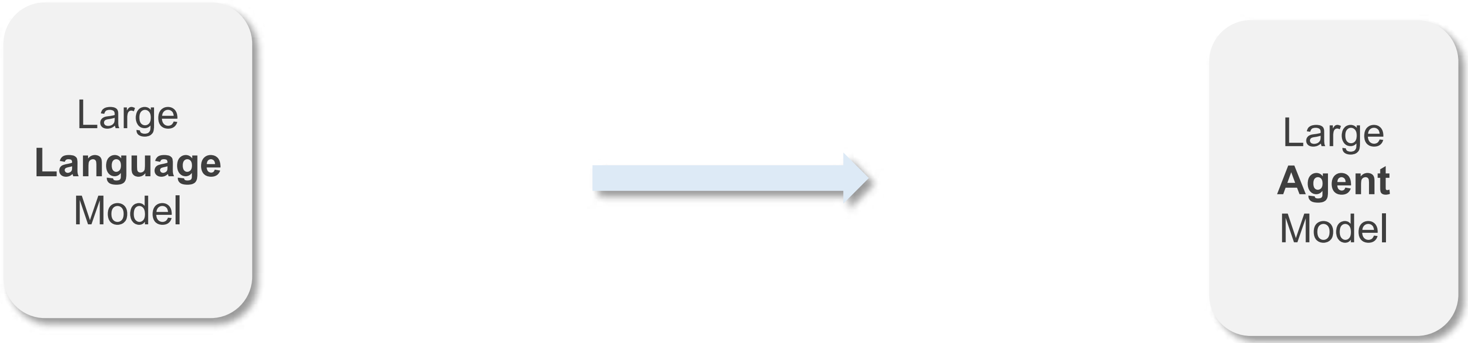
“Disentanglement” is the key of casual reasoning!



https://www.flynnresearchgroup.com/scientific-argumentation

“Disentanglement”

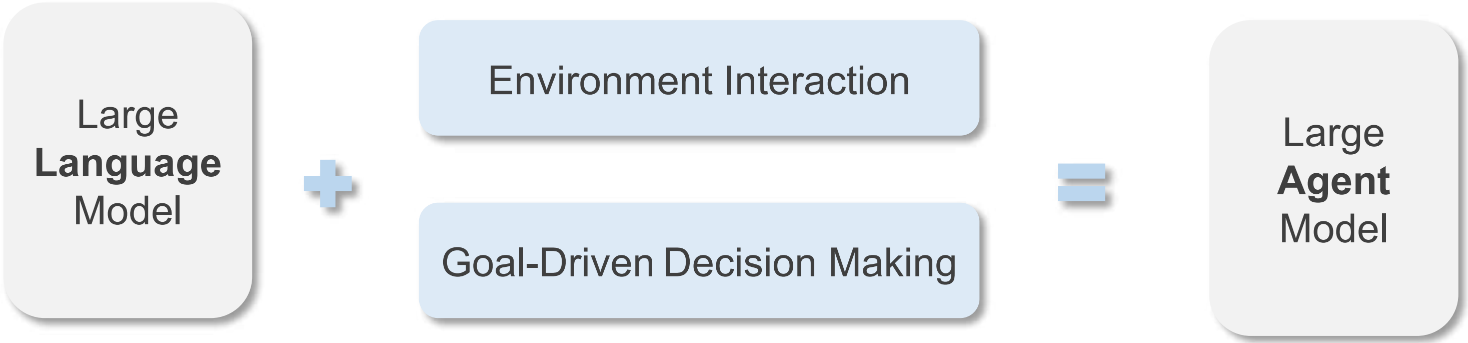
Large **Language** Models → Large **Agent** Models



Large
Language
Model

Large
Agent
Model

Large **Language** Models → Large **Agent** Models



Large **Language** Model

+

Environment Interaction

Goal-Driven Decision Making

=

Large **Agent** Model

Such disentanglement enables a generalist agent

What is a generalist agent?

Having a robot that can do many tasks, across many environments.

Figure credits: Jiayuan Mao

What is a generalist agent?

Having a robot that can do many tasks, across many environments.

Interact with environment

What is a generalist agent?

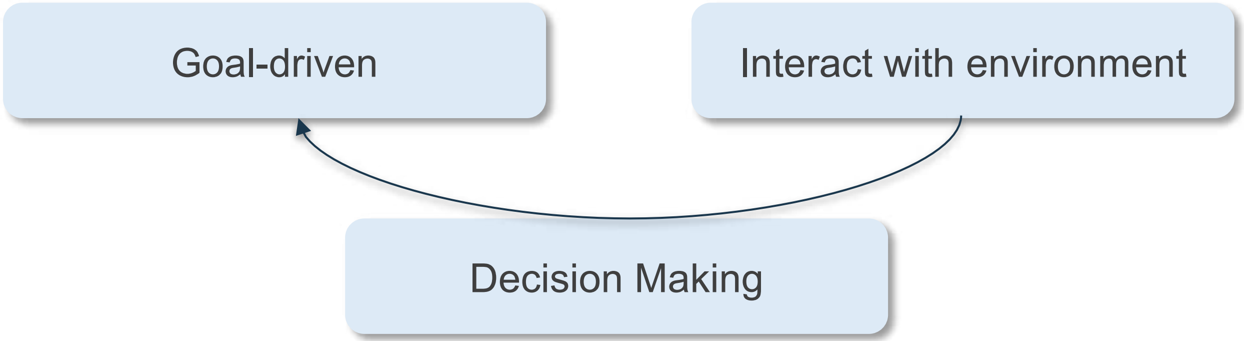
Having a robot that can do many tasks, across many environments.

Goal-driven

Interact with environment

What is a generalist embodied agent?

Having a robot that can do many tasks, across many environments.

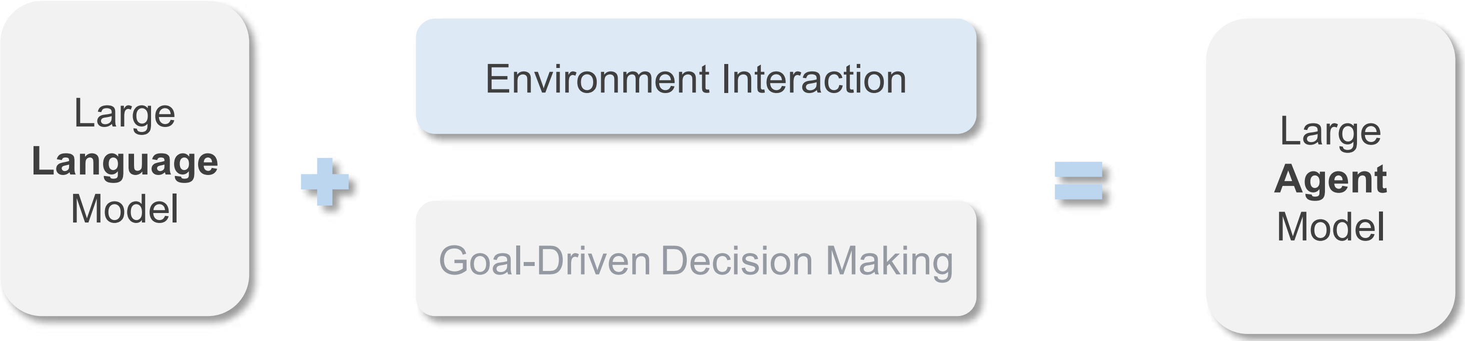


Goal-driven

Interact with environment

Decision Making

Large **Language** Models → Large **Agent** Models



Large
Language
Model

+

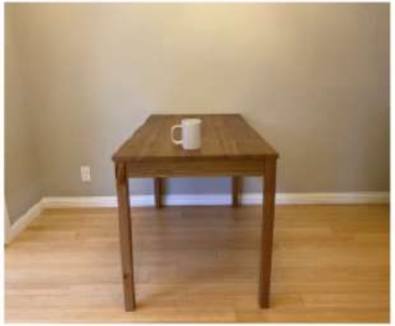
Environment Interaction

Goal-Driven Decision Making

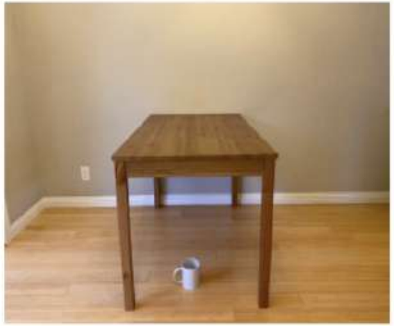
=

Large
Agent
Model

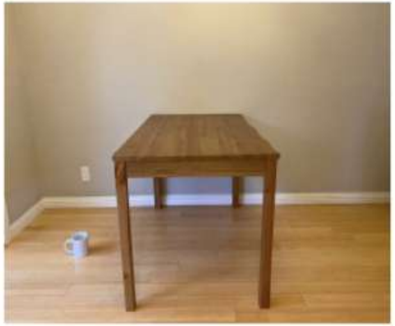
Environment Interaction



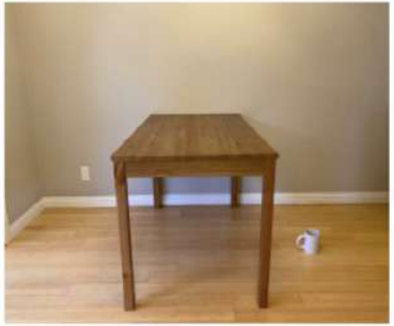
Cup is on top of the table.



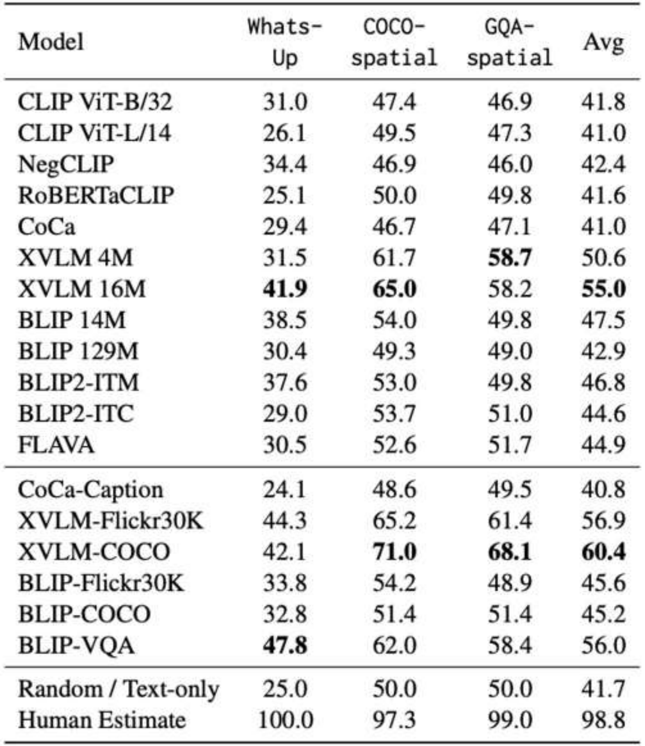
Cup is on top of the table. ✗



Cup is on top of the table. ✗



Cup is on top of the table. ✗

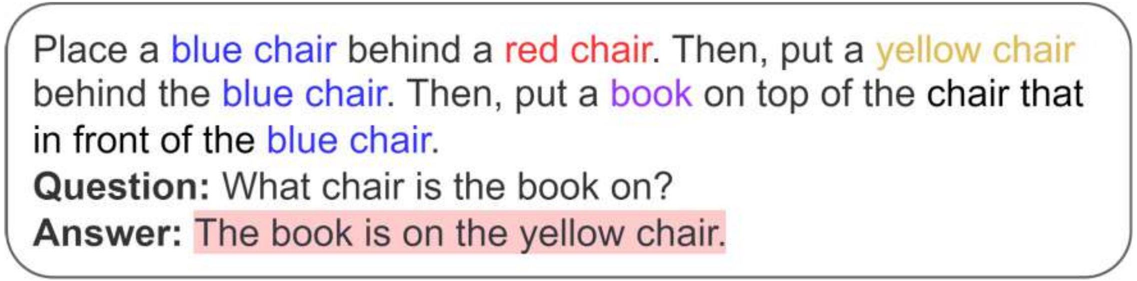


Model	Whats-Up	COCO-spatial	GQA-spatial	Avg
CLIP ViT-B/32	31.0	47.4	46.9	41.8
CLIP ViT-L/14	26.1	49.5	47.3	41.0
NegCLIP	34.4	46.9	46.0	42.4
RoBERTaCLIP	25.1	50.0	49.8	41.6
CoCa	29.4	46.7	47.1	41.0
XVLM 4M	31.5	61.7	**58.7**	50.6
XVLM 16M	**41.9**	**65.0**	58.2	**55.0**
BLIP 14M	38.5	54.0	49.8	47.5
BLIP 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

Amita et al, What's "up" with vision-language models? Investigating their struggle with spatial reasoning, EMNLP 2023

What is missing: Physical World State Changes

Current **LLMs** still **fall short** on understanding concepts involving complex **physical interactions**.



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 in front of the blue chair.

Question: What chair is the book on?

Answer: The book is on the yellow chair.

GPT4 failed example

Mental models, which can be viewed as **internal representations of the physical world**, enable **humans** to understand such concepts with ease.

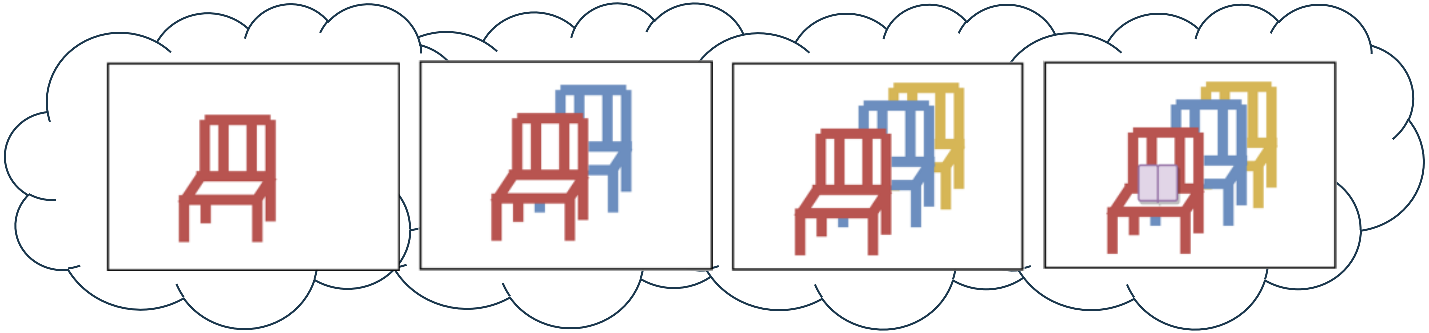
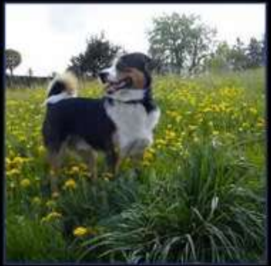
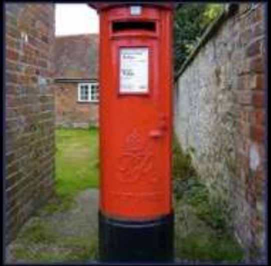


Illustration Of Mental Simulation



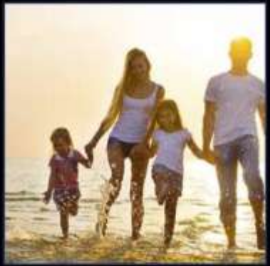
Q: Is the dog facing left or right from the camera's perspective?

The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.



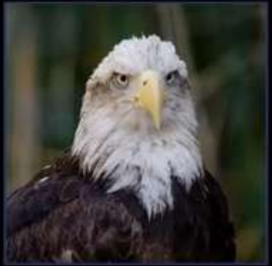
Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.



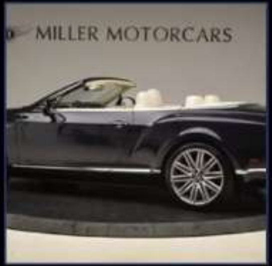
Q: Do the individuals in the picture face the front or the back?

The individuals in the picture are **facing away from the camera**, so we see them from the back.



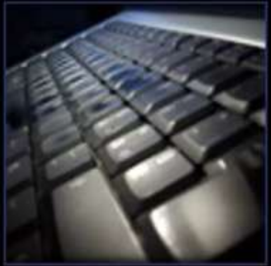
Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.



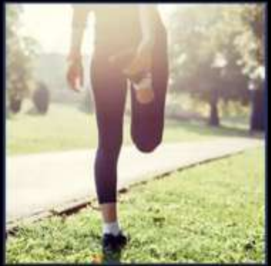
Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.



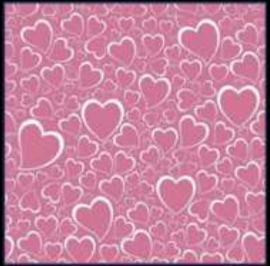
Q: Does the keyboard have a backlight?

Based on the image, it appears that the keyboard **does have a backlight**. This is indicated by the illumination seen around the keys, which suggests that there is a light source beneath them to enhance visibility in low-light conditions.



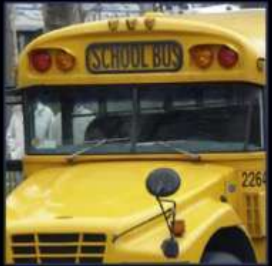
Q: Is the person in the picture on the grass or on the gravel path?

The person in the image appears to be running **on a gravel path**. You can discern this from the texture and color contrast between the light-colored path and the green grass surrounding it.



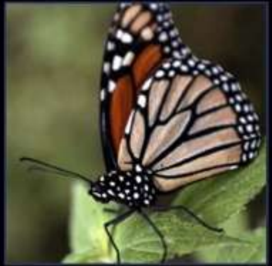
Q: Is the edge of the heart in the picture white or dark-colored?

The image you've provided is small and of low resolution, but it appears that the hearts have a **dark-colored edge** or outline. This outline varies in visibility, likely due to the quality of the image.



Q: Is the school bus parked facing the camera or away from the camera?

The school bus is parked **facing away from the camera**. You can tell because the rear lights and the back of the exterior rearview mirror are visible. The text "SCHOOL BUS" is also reversed, which is typically seen on the front of a bus, designed so that it reads correctly in a rearview mirror.



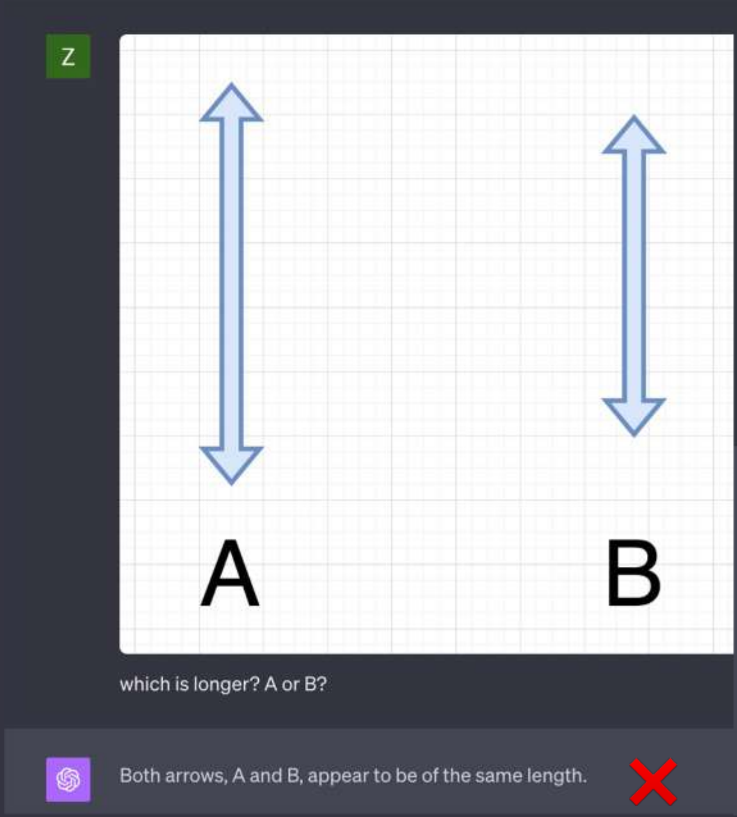
Q: Are the butterfly's feet visible?

The butterfly's feet are **not visible** in this image. The wings and part of the body of the butterfly are in view, but the feet are either hidden beneath the body or out of the frame of the picture. Butterflies have delicate, slender feet which are often not the most noticeable feature when they are perched, especially from the angle shown in this image.

Tong, Shengbang, et al. "Eyes wide shut? exploring the visual shortcomings of multimodal llms." CVPR2024

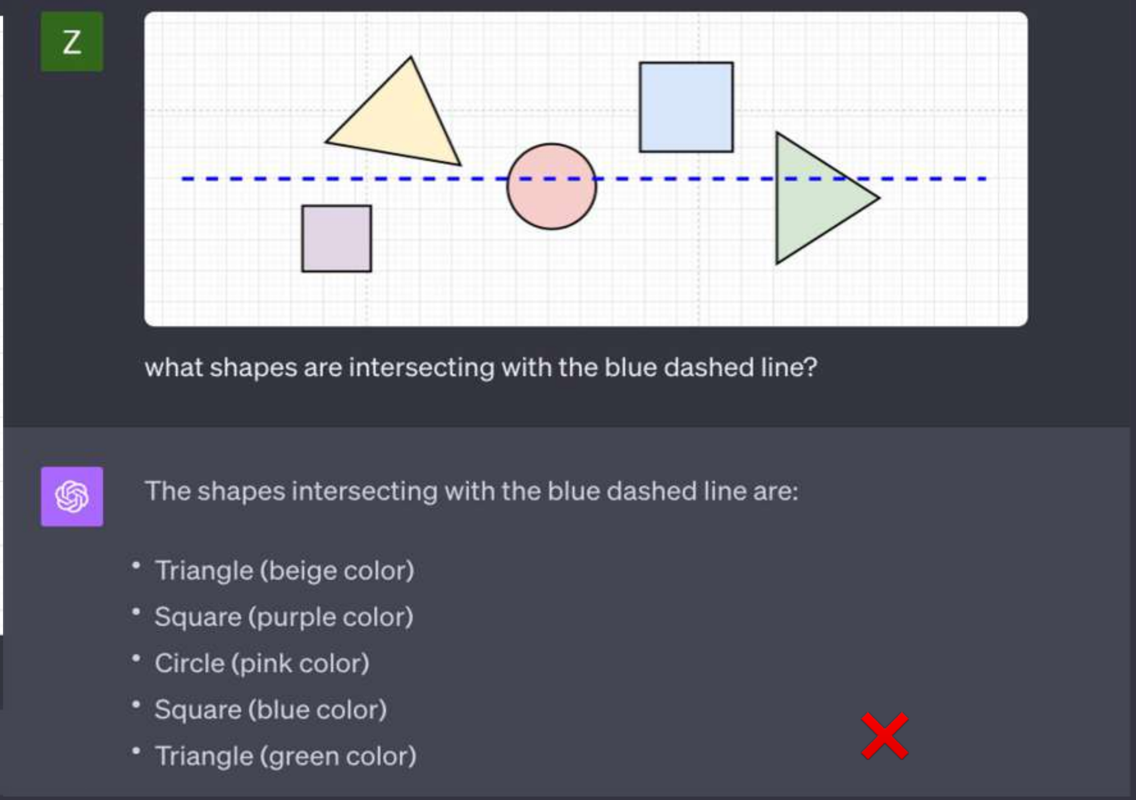
Current VLMs have Poor Geometric Knowledge [Wang et al., 2024]

- Shape/Length



which is longer? A or B?

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

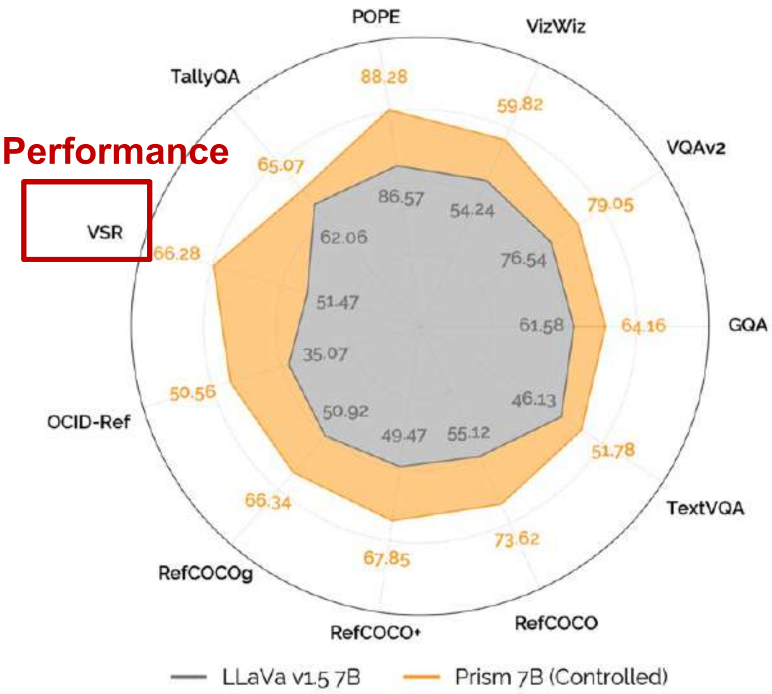


what shapes are intersecting with the blue dashed line?

The shapes intersecting with the blue dashed line are:

- Triangle (beige color)
- Square (purple color)
- Circle (pink color)
- Square (blue color)
- Triangle (green color)

Low Performance on Visual Spatial Reasoning

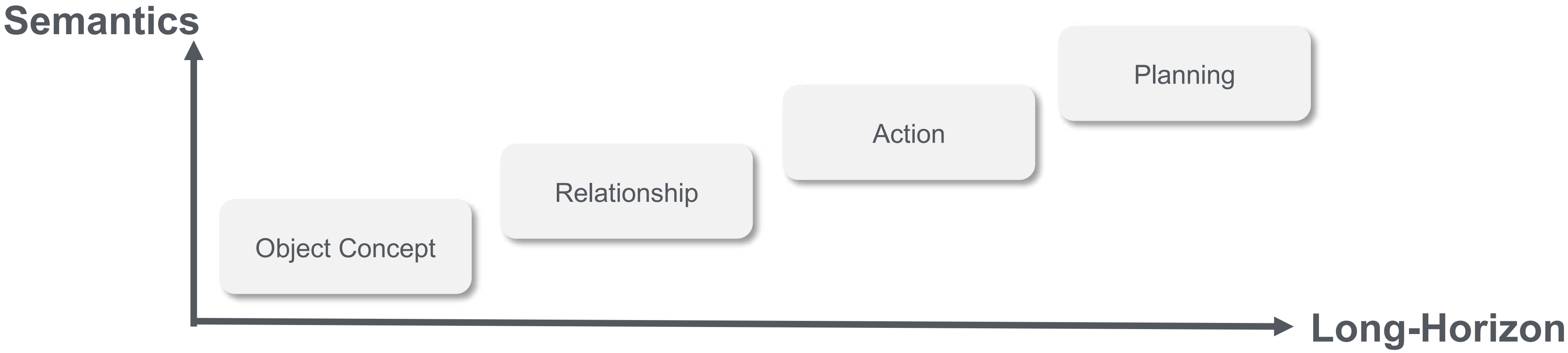


Low Performance

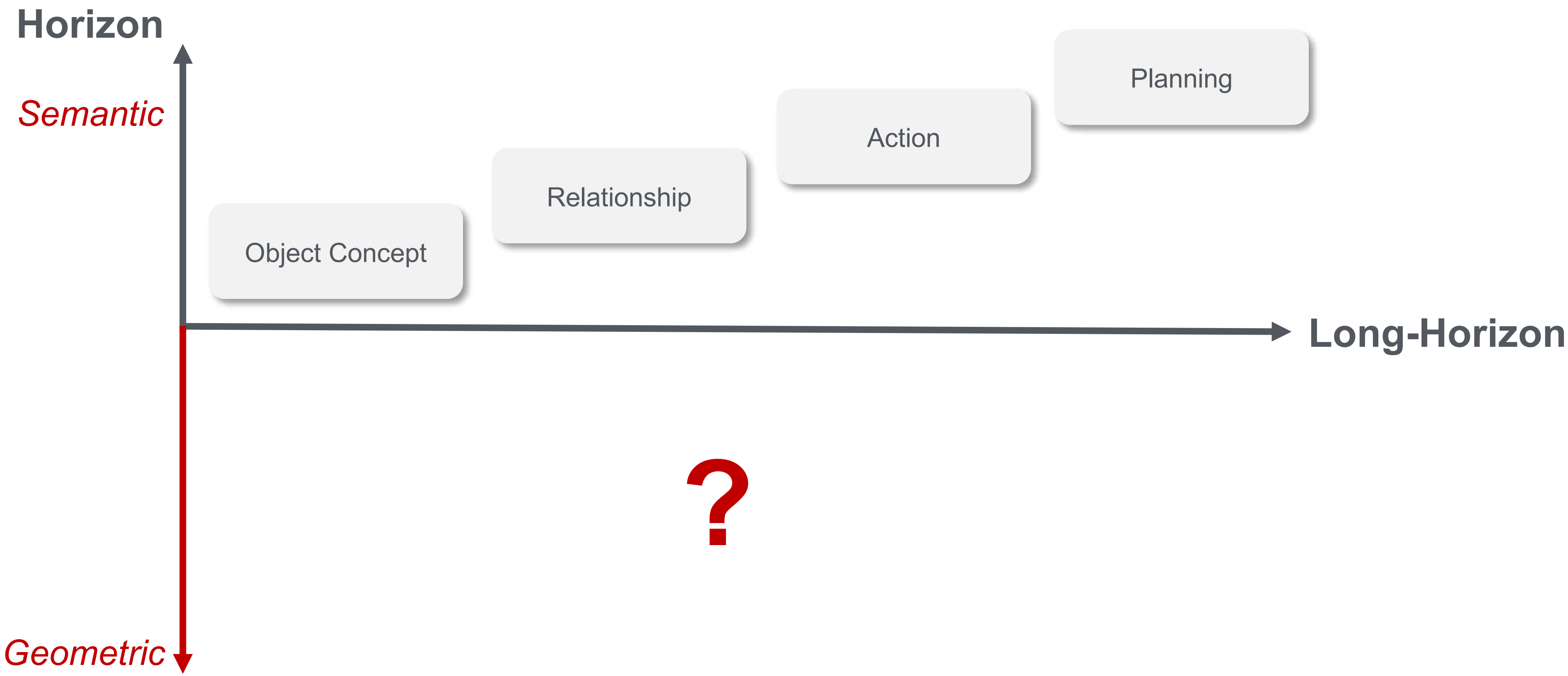
Video-Language Foundation Models

Prismatic VLMs (2024)

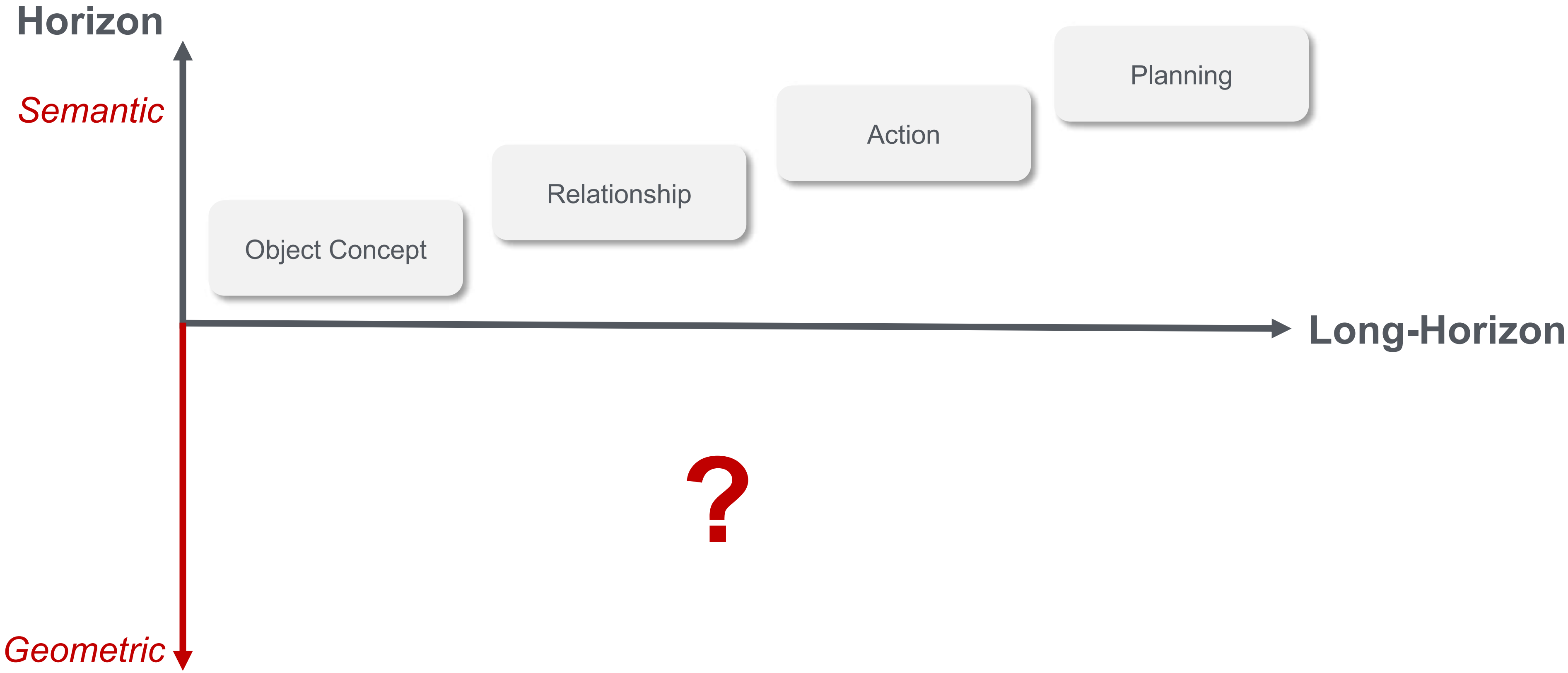
What we know about physical world



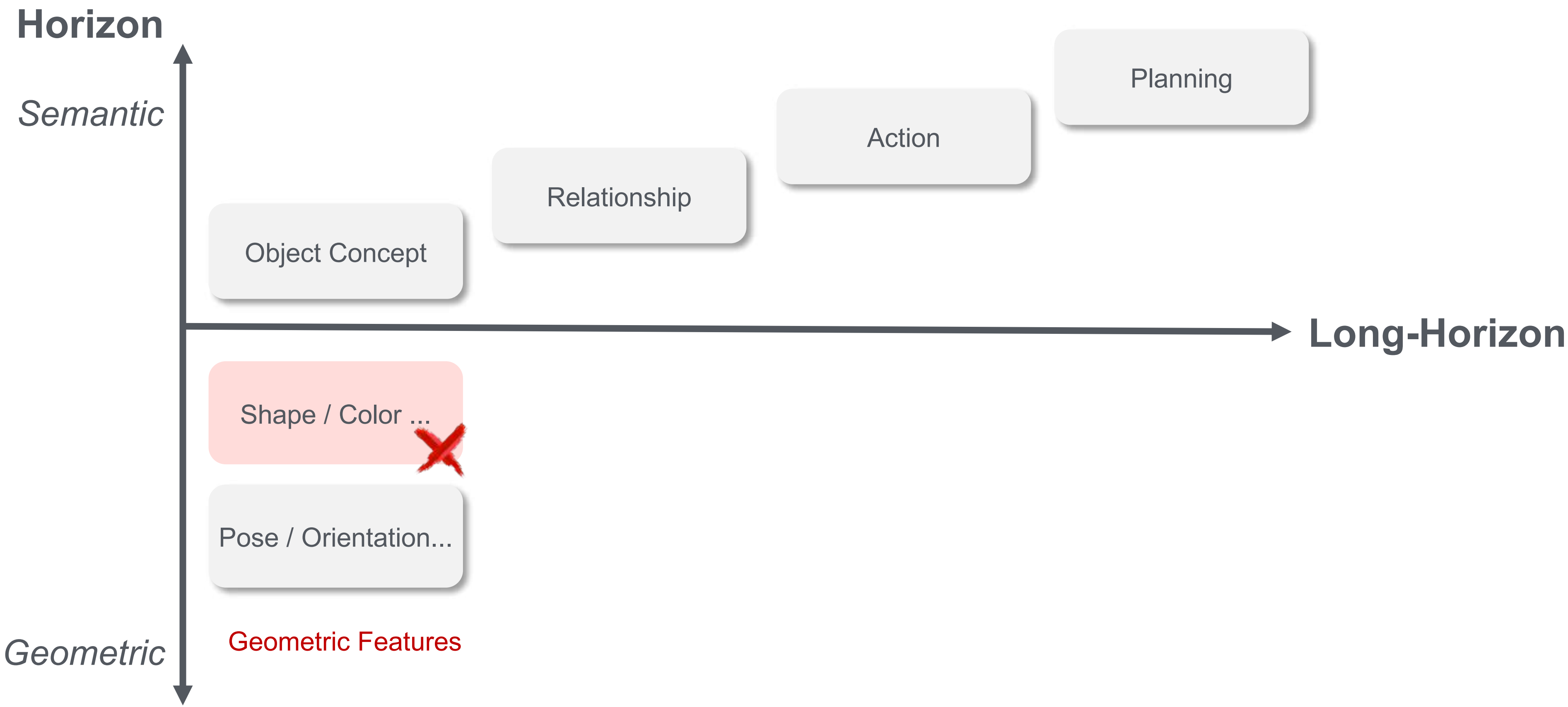
Missing knowledge about physical world



Missing knowledge about physical world



Missing knowledge about physical world



Horizon

Semantic

Planning

Action

Relationship

Object Concept

Long-Horizon

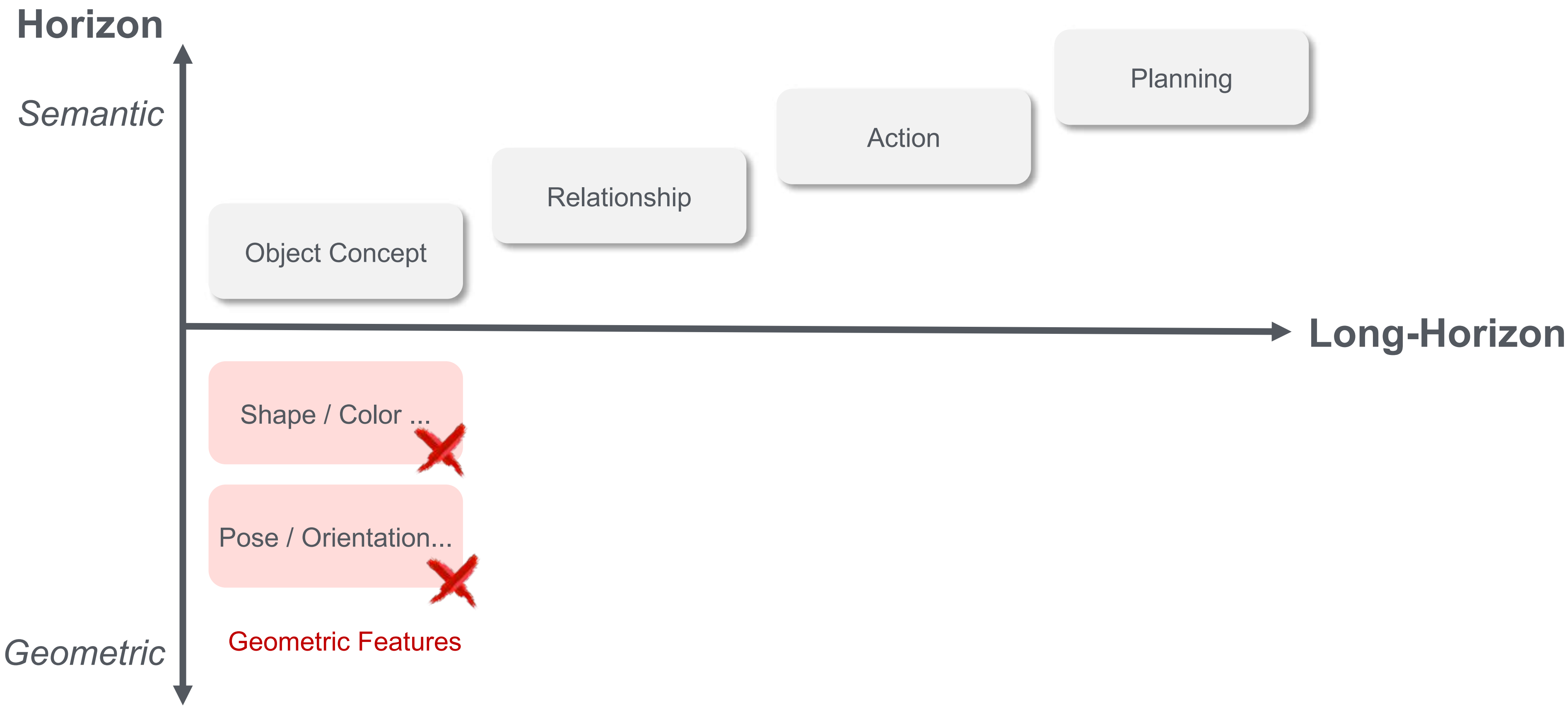
Shape / Color ...

Pose / Orientation...

Geometric Features

Geometric

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

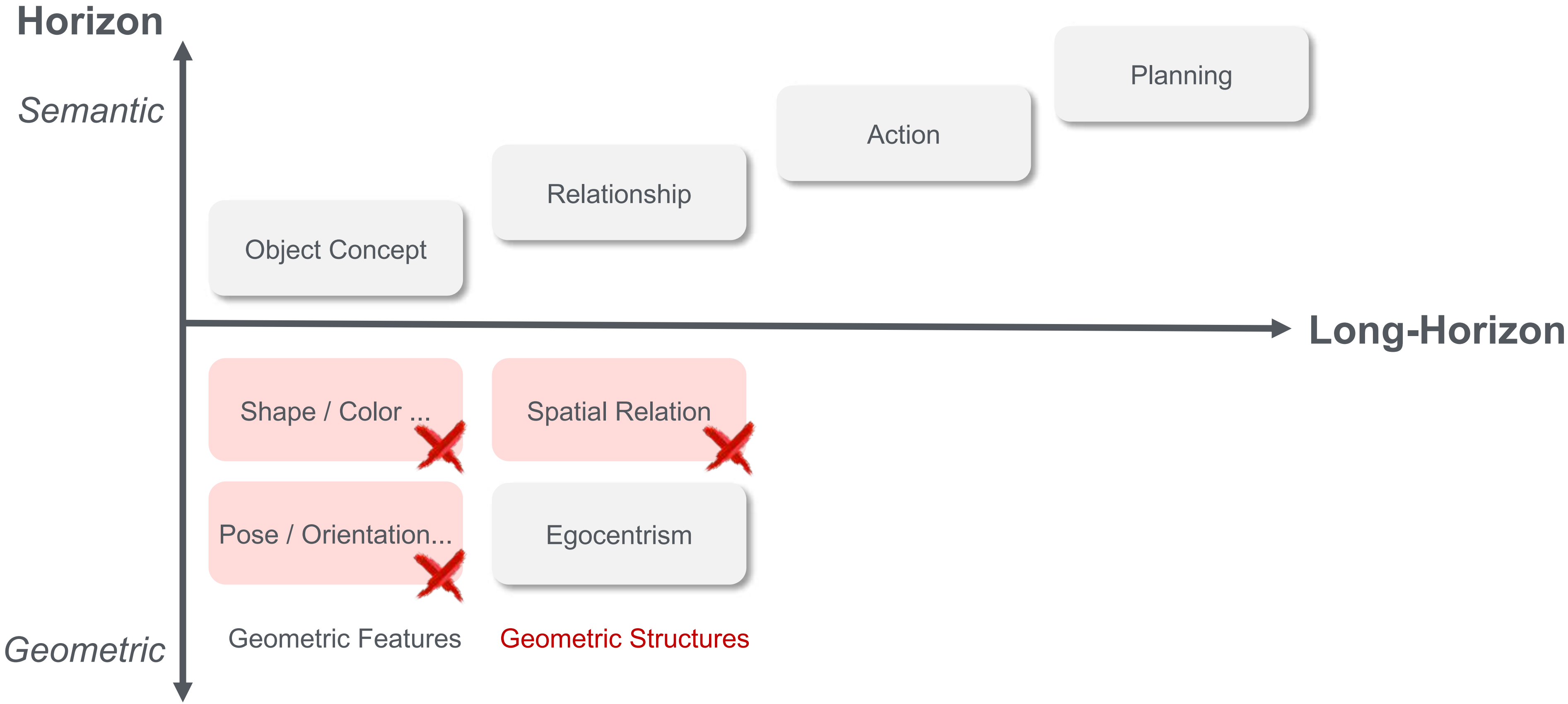
Shape / Color ...

Pose / Orientation...

Geometric Features

Geometric

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ... ✗

Spatial Relation ✗

Pose / Orientation... ✗

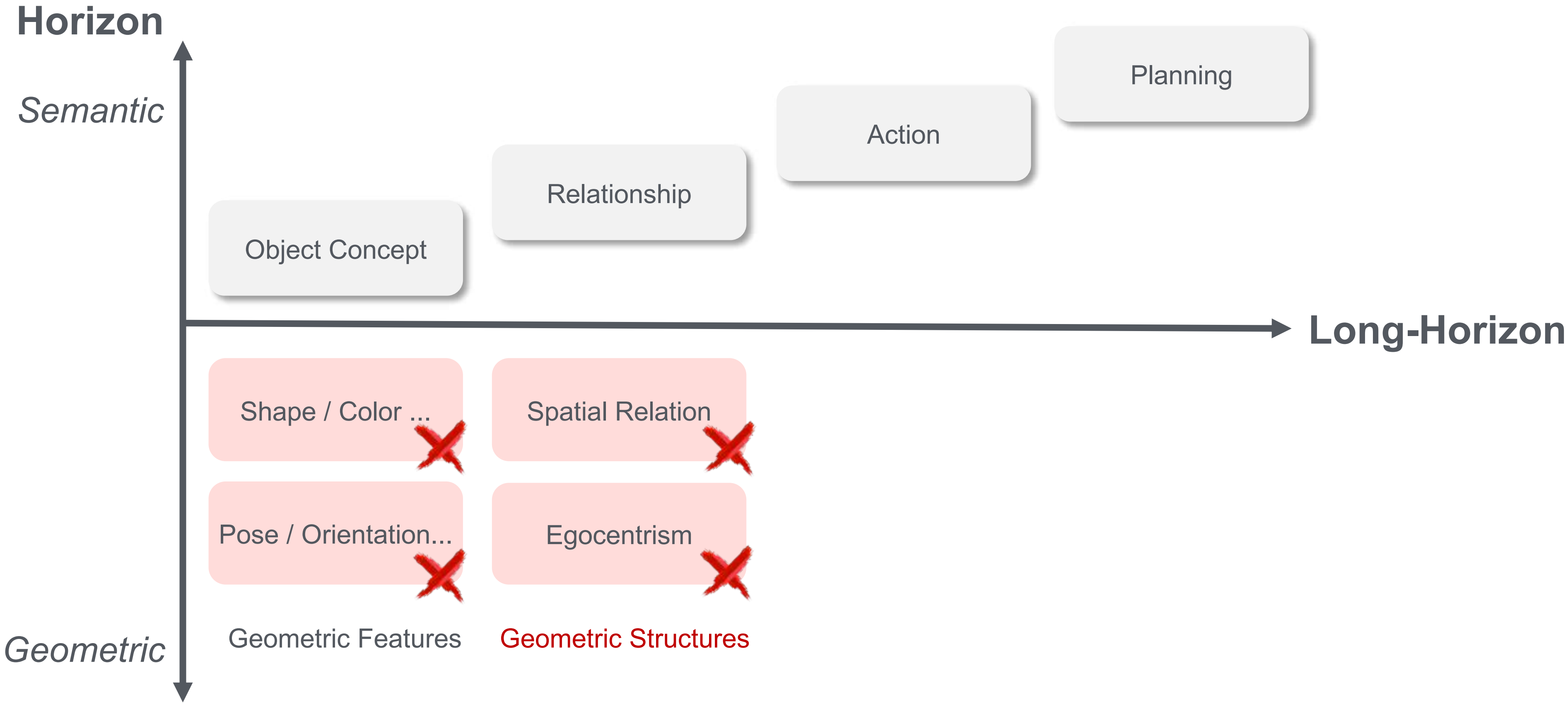
Egocentrism

Geometric

Geometric Features

Geometric Structures

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ...

Spatial Relation

Pose / Orientation...

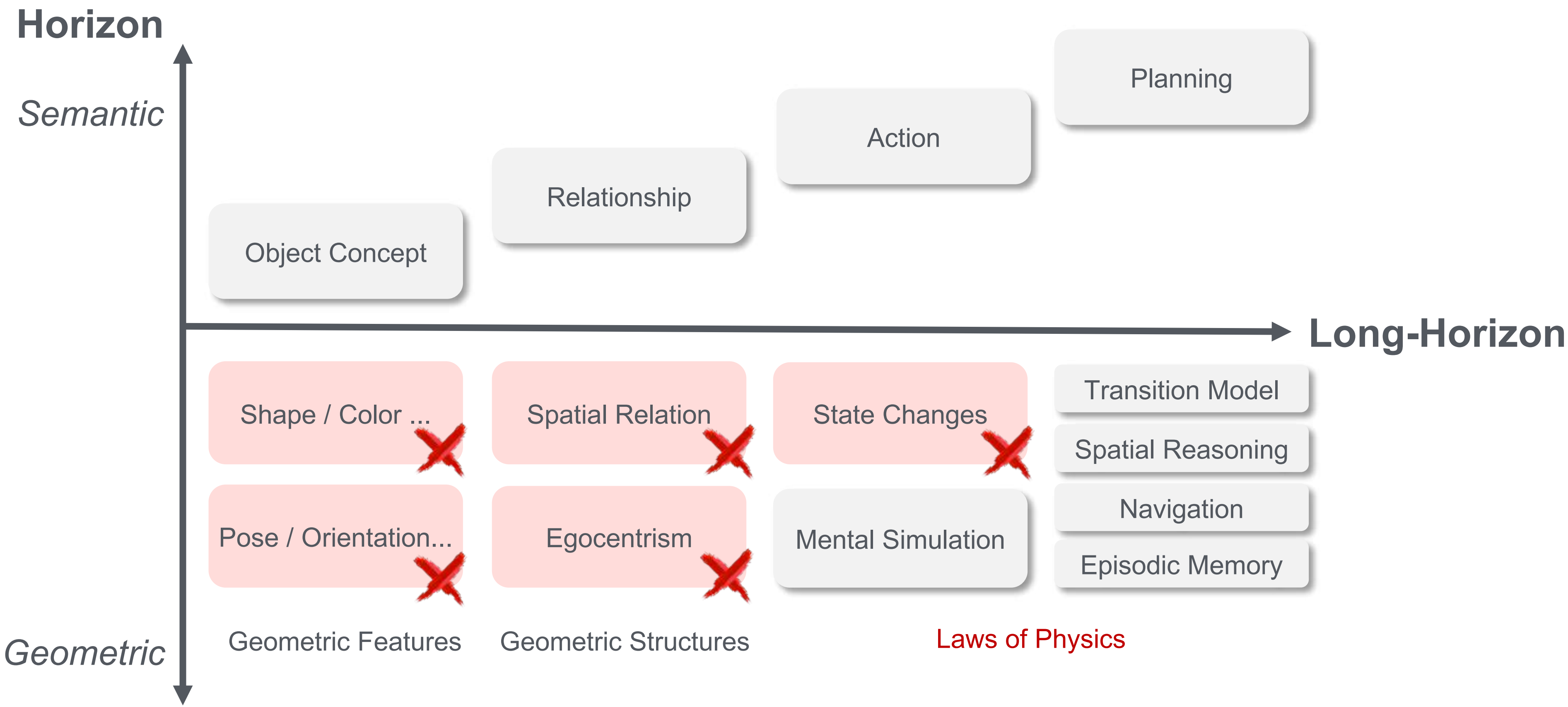
Egocentrism

Geometric

Geometric Features

Geometric Structures

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ... ✗

Pose / Orientation... ✗

Spatial Relation ✗

Egocentrism ✗

State Changes ✗

Mental Simulation

Transition Model

Spatial Reasoning

Navigation

Episodic Memory

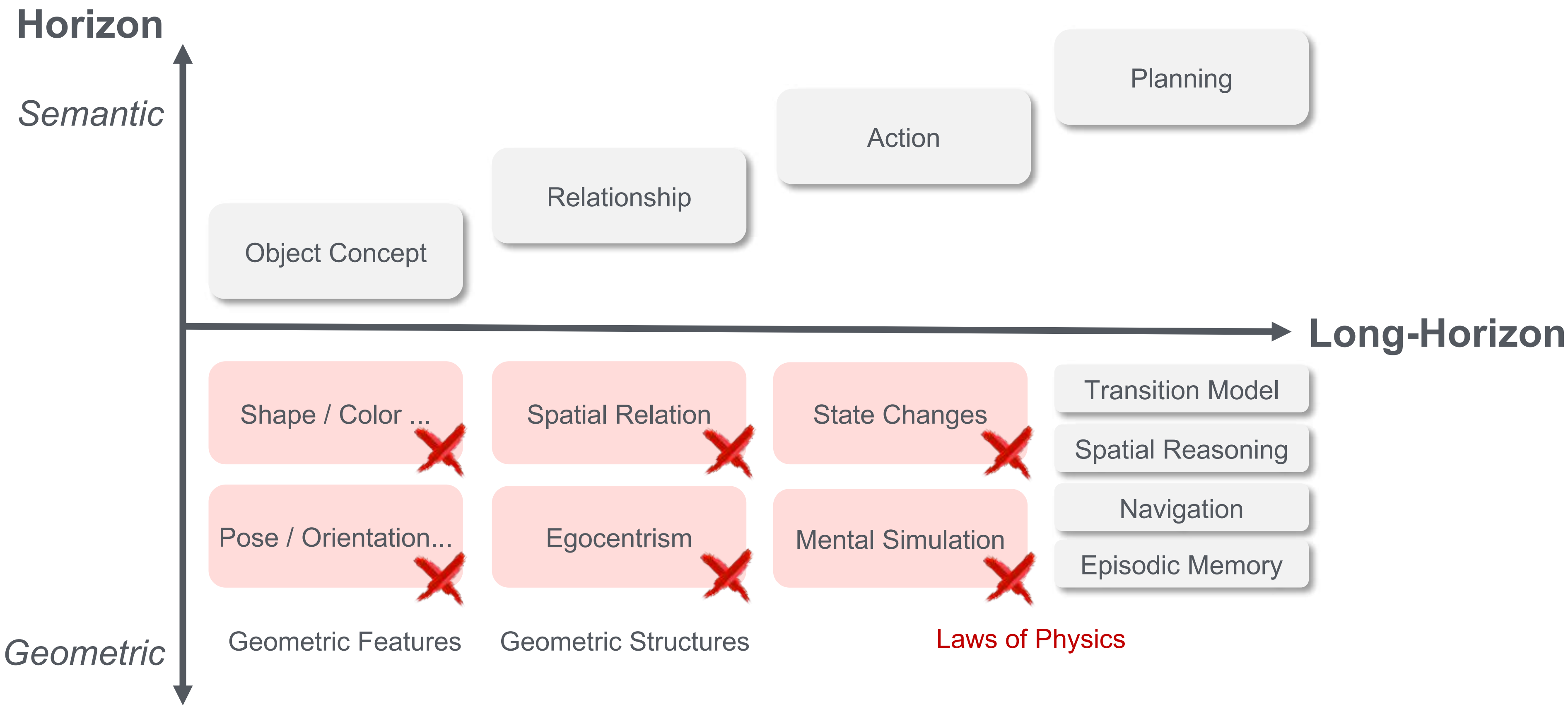
Geometric

Geometric Features

Geometric Structures

Laws of Physics

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ...

Spatial Relation

State Changes

Transition Model

Spatial Reasoning

Pose / Orientation...

Egocentrism

Mental Simulation

Navigation

Episodic Memory

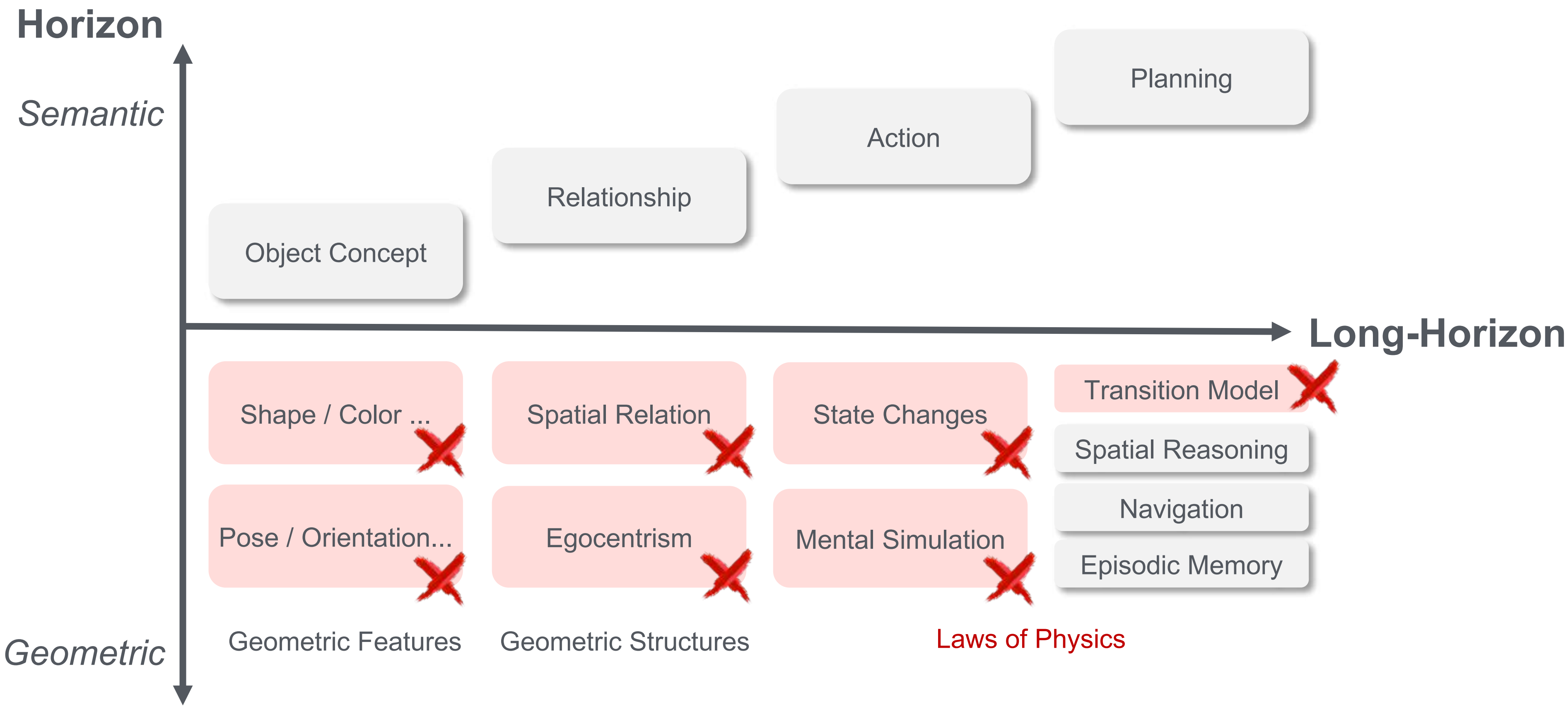
Geometric

Geometric Features

Geometric Structures

Laws of Physics

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ... ✗

Pose / Orientation... ✗

Spatial Relation ✗

Egocentrism ✗

State Changes ✗

Mental Simulation ✗

Transition Model ✗

Spatial Reasoning

Navigation

Episodic Memory

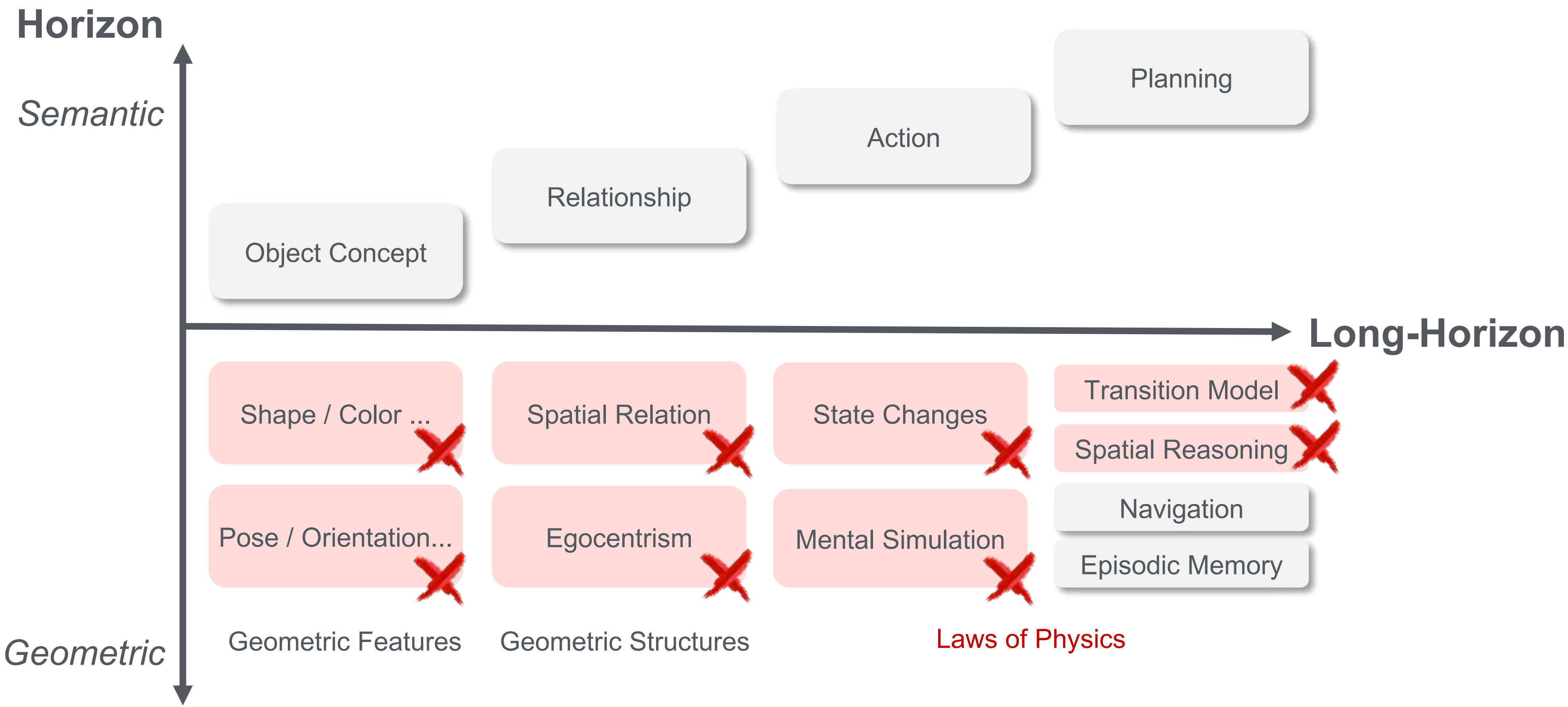
Geometric

Geometric Features

Geometric Structures

Laws of Physics

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ... ✗

Spatial Relation ✗

State Changes ✗

Transition Model ✗

Spatial Reasoning ✗

Pose / Orientation... ✗

Egocentrism ✗

Mental Simulation ✗

Navigation

Episodic Memory

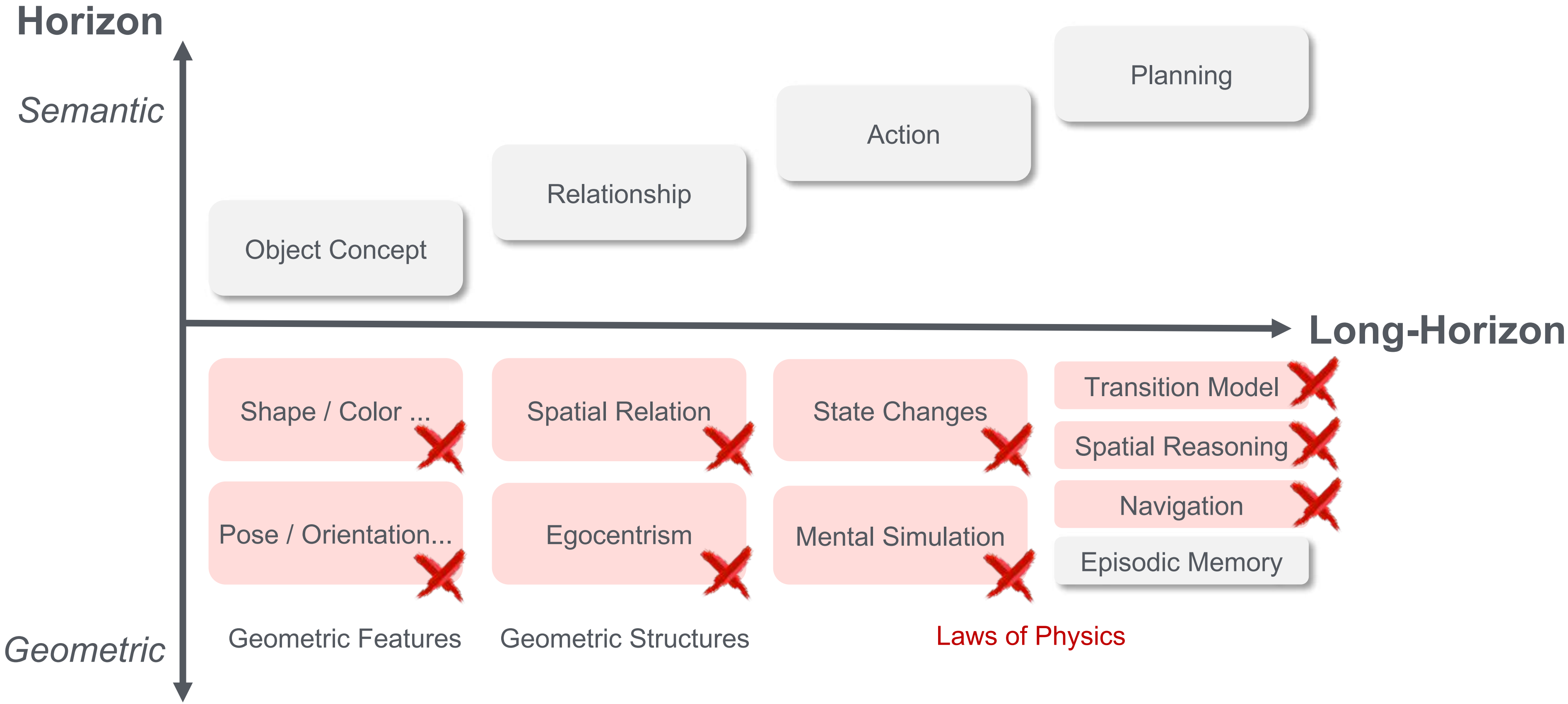
Geometric

Geometric Features

Geometric Structures

Laws of Physics

Missing knowledge about physical world



Horizon

Semantic

Planning

Action

Relationship

Object Concept

Long-Horizon

Shape / Color ... ✗

Pose / Orientation... ✗

Spatial Relation ✗

Egocentrism ✗

State Changes ✗

Mental Simulation ✗

Transition Model ✗

Spatial Reasoning ✗

Navigation ✗

Episodic Memory

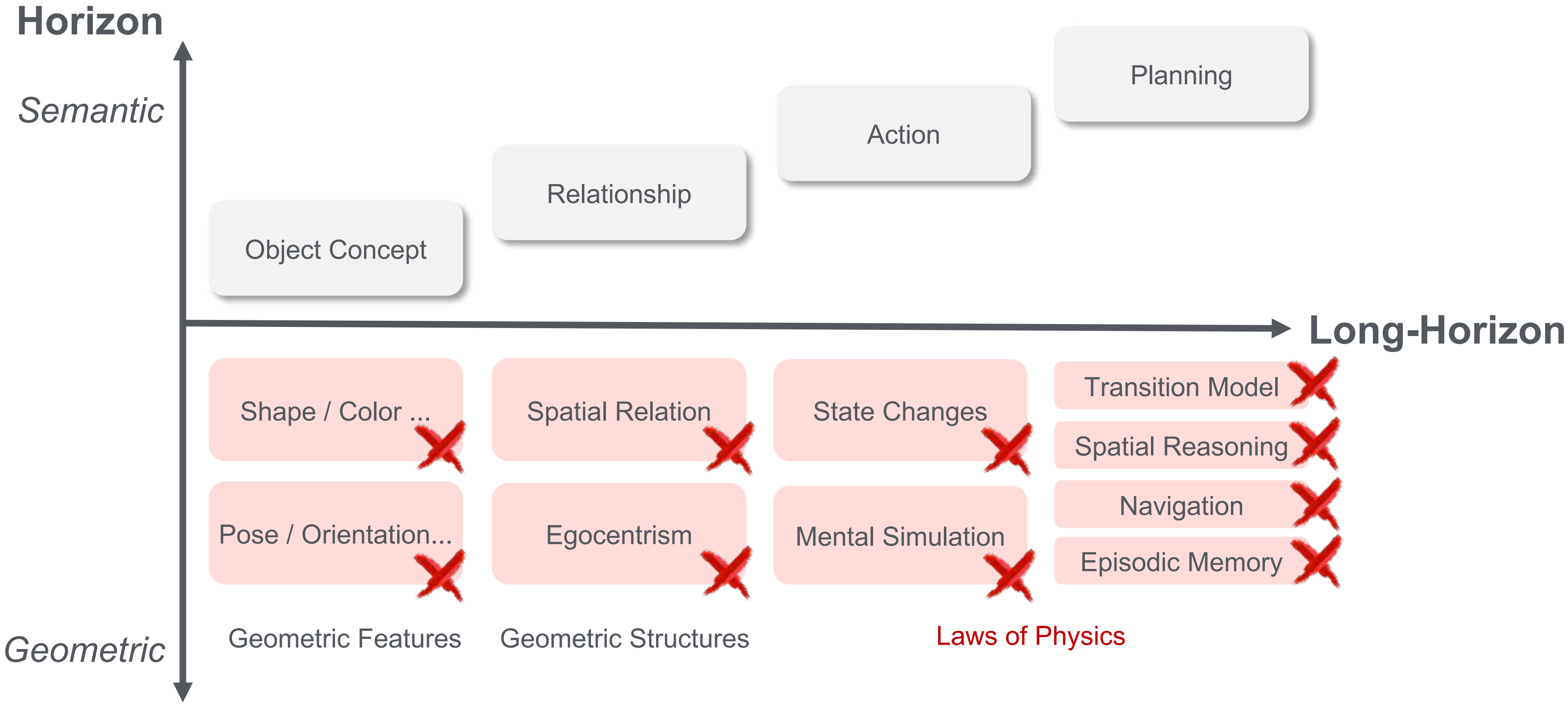
Geometric Features

Geometric Structures

Laws of Physics

Geometric

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ... ✗

Pose / Orientation... ✗

Spatial Relation ✗

Egocentrism ✗

State Changes ✗

Mental Simulation ✗

Transition Model ✗

Spatial Reasoning ✗

Navigation ✗

Episodic Memory ✗

Geometric

Geometric Features

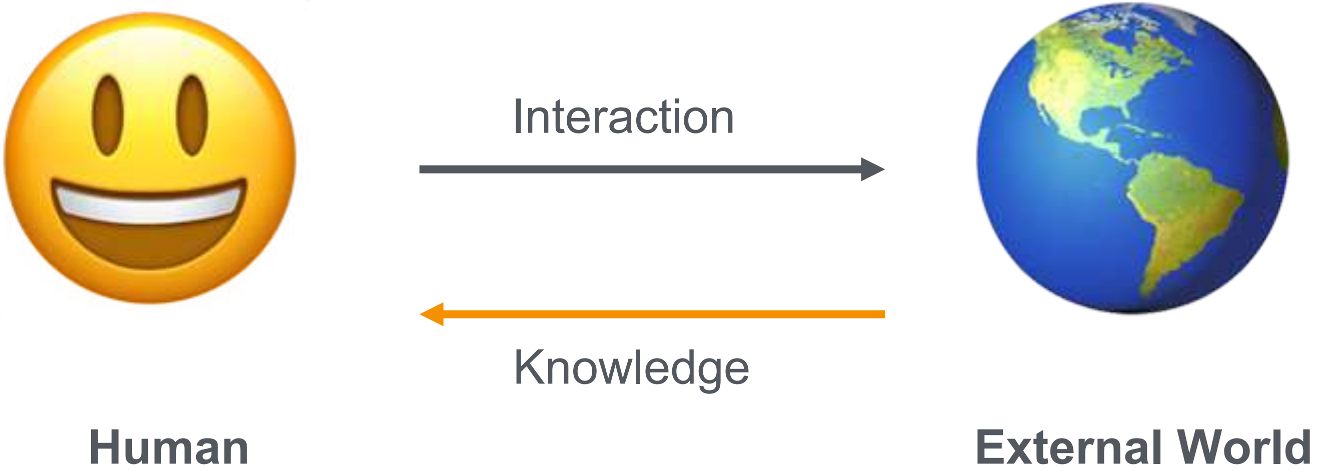
Geometric Structures

Laws of Physics

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

Why?

Humans learn knowledge through **interactions**



Interaction

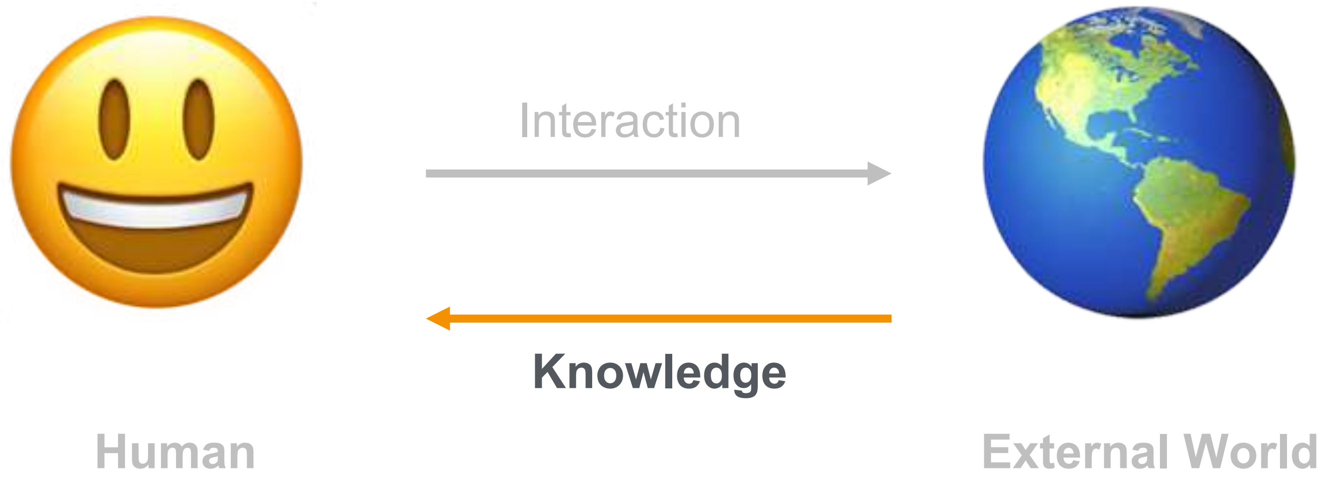
Knowledge

Human

External World

Credits: Zhenhailong

Humans learn knowledge through **interactions**



Interaction

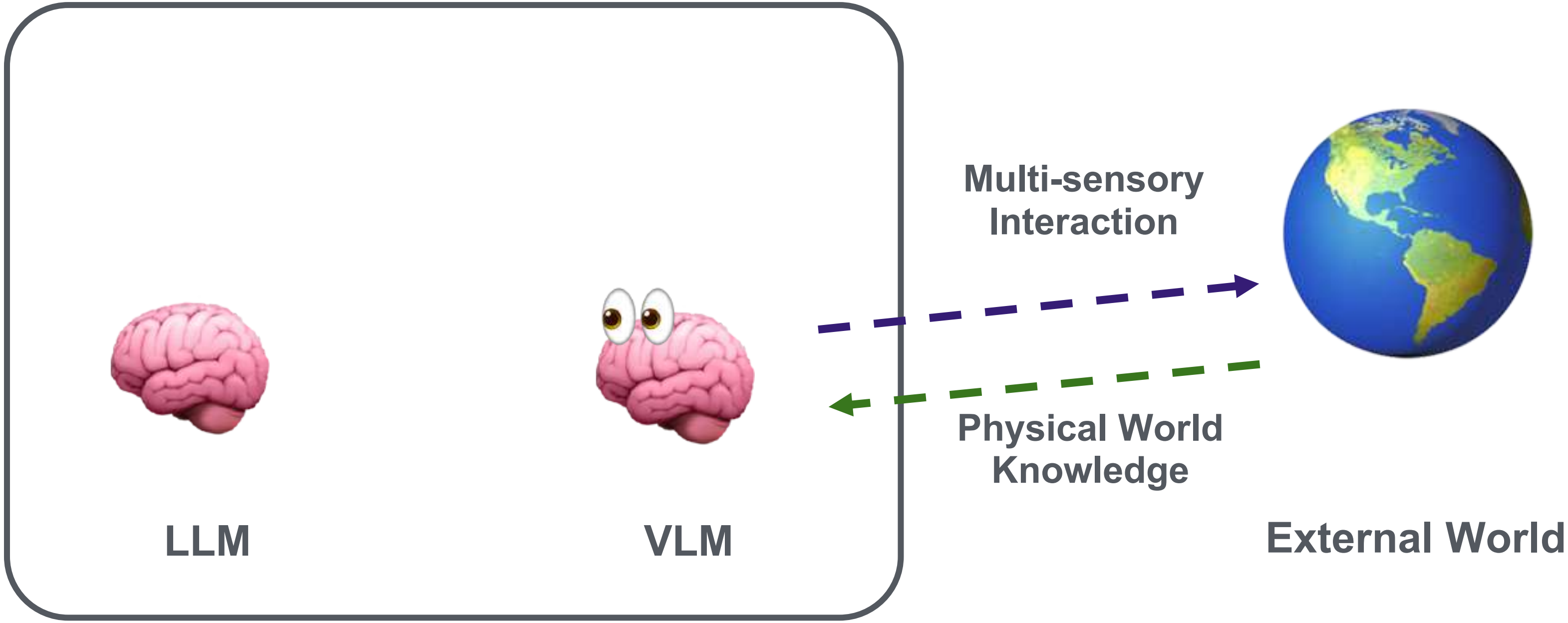
Knowledge

Human

External World

Credits: Zhenhailong

Machines learn knowledge **w/o interactions**



Multi-sensory Interaction

Physical World Knowledge

LLM

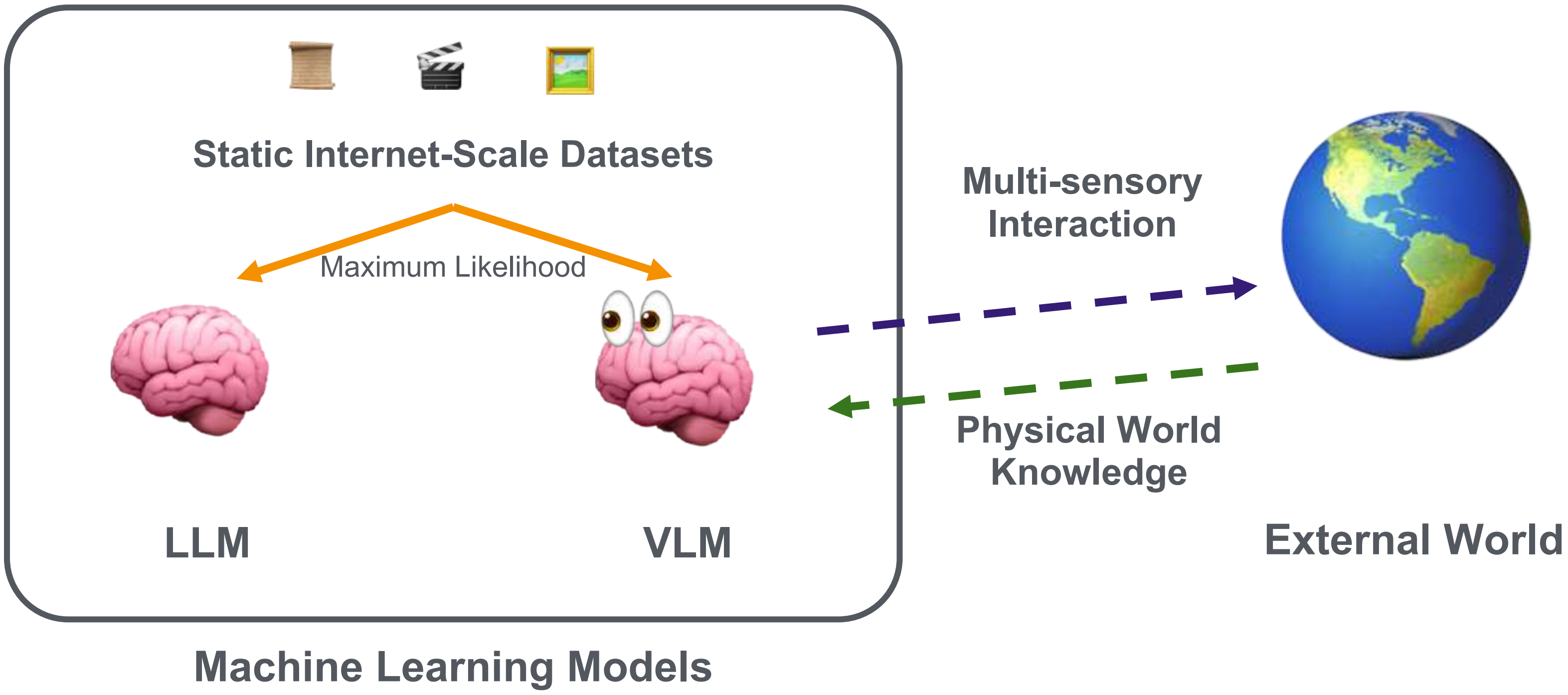
VLM

External World

Machine Learning Models

Credits: Zhenhailong

Machines learn knowledge **w/o interactions**



Static Internet-Scale Datasets

Maximum Likelihood

LLM

VLM

Machine Learning Models

Multi-sensory Interaction

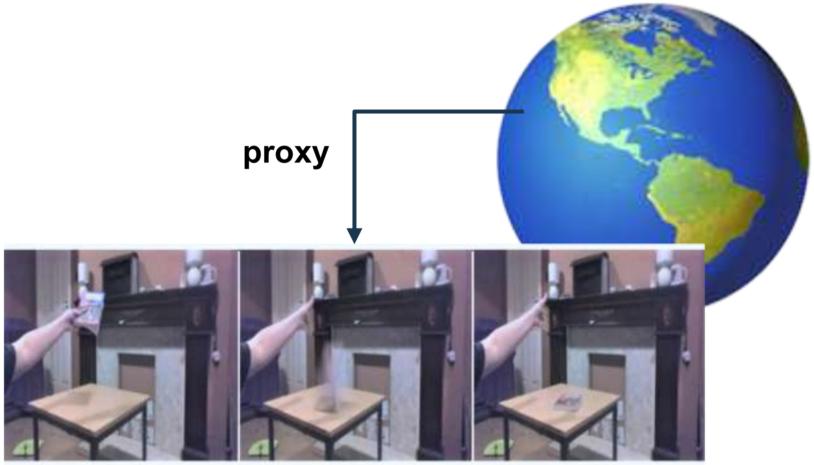
Physical World Knowledge

External World

Credits: Zhenhailong

Static Internet-Scale Datasets

Video: A “Visual Recording” of World State Changes

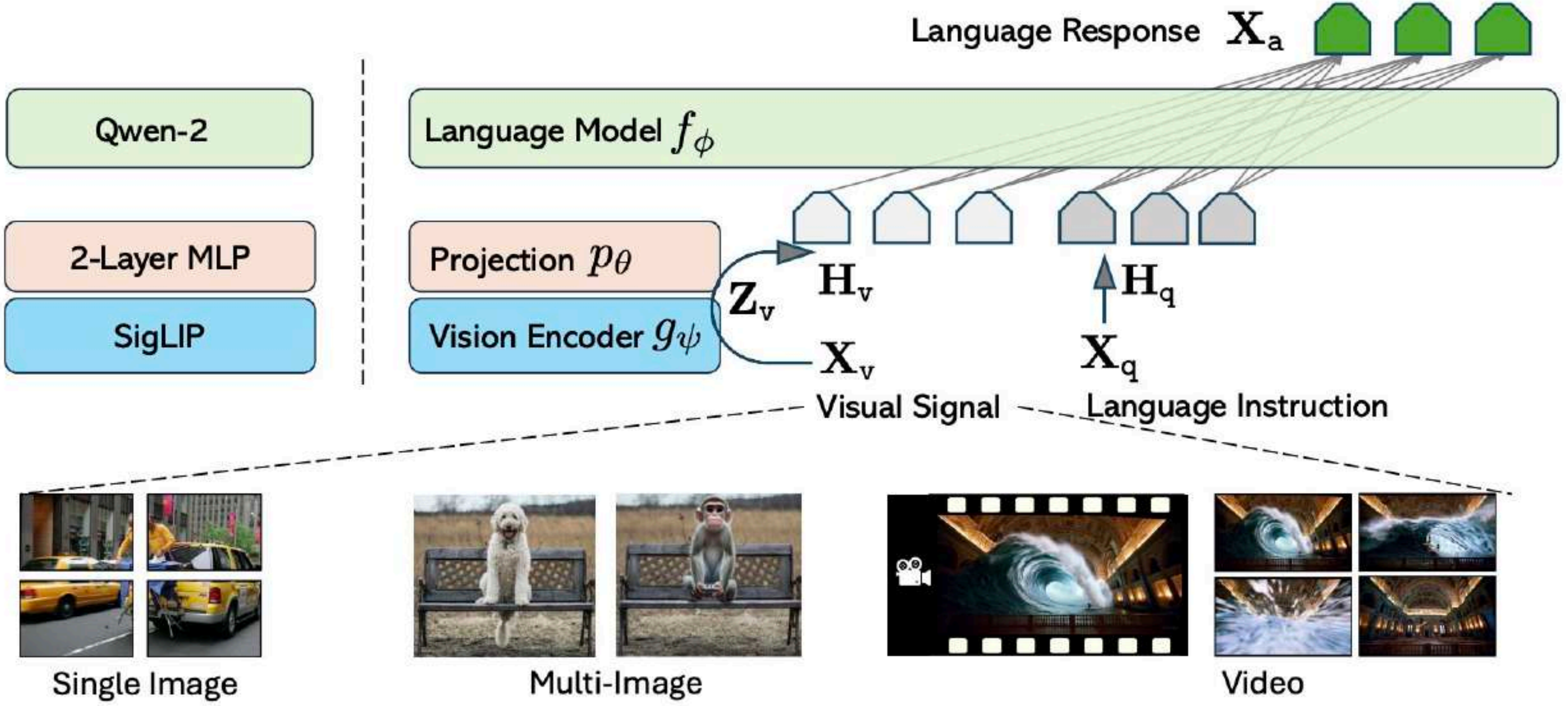


“Book falling like a rock”

Video-Language Datasets

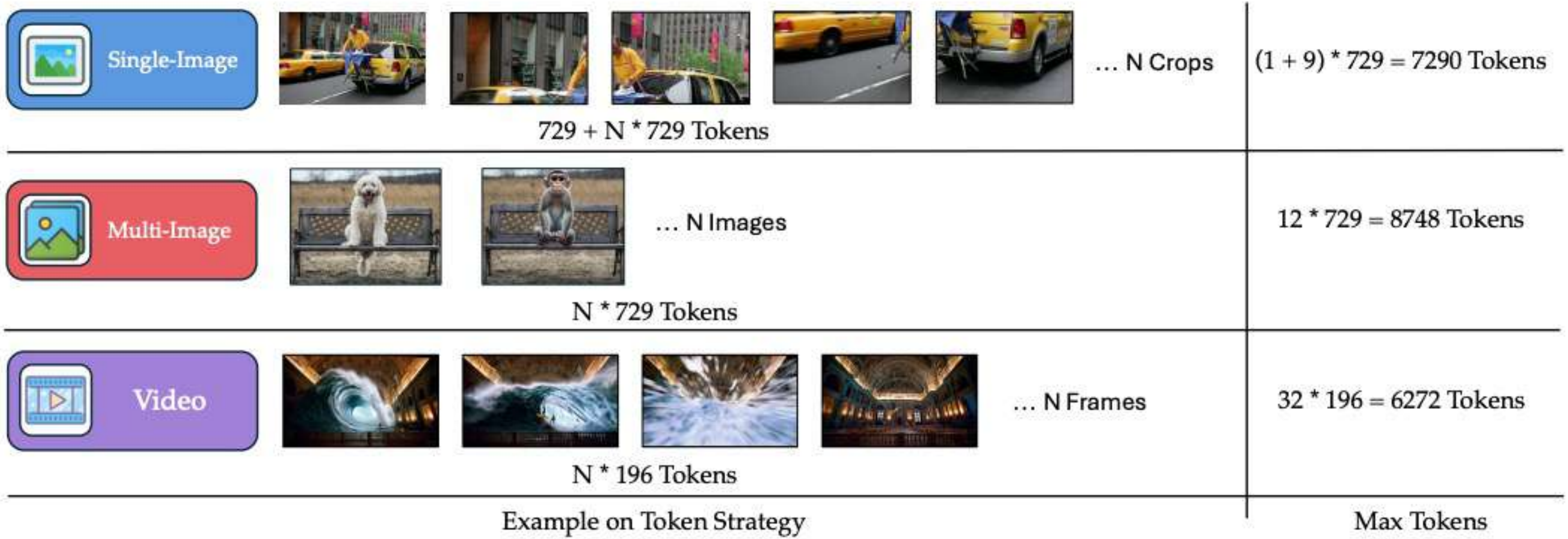
Credits: Zhenhailong

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



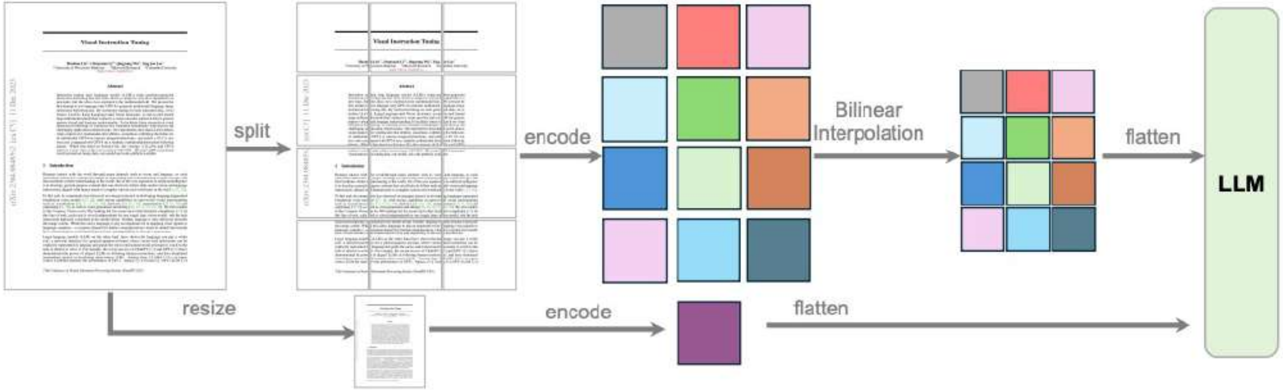
Credits: LLaVA-OneVision

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

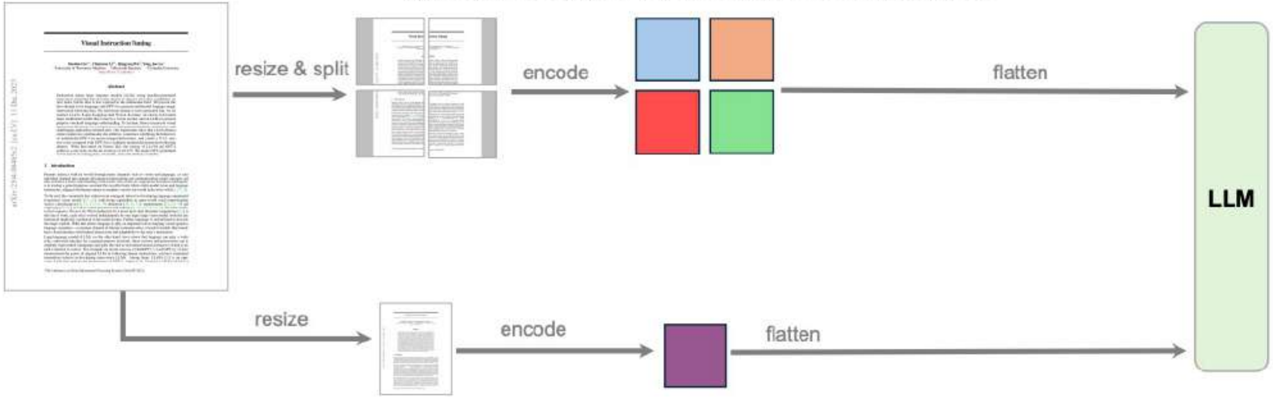


Credits: LLaVA-OneVision

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

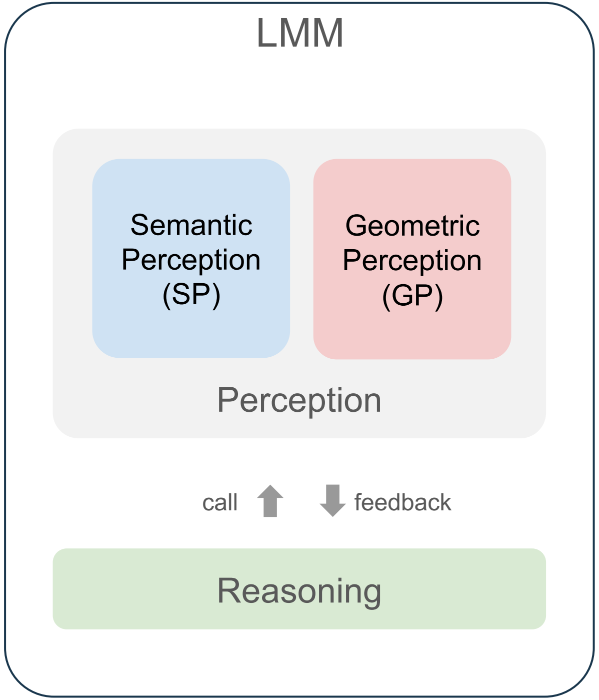


(a) Higher AnyRes with Bilinear Interpolation



Credits: LLaVA-OneVision

Semantic vs Geometric



LMM

Semantic Perception (SP)

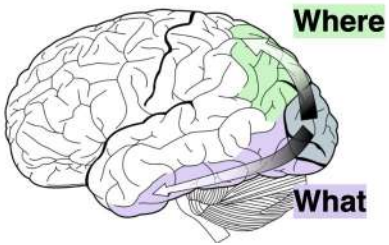
Geometric Perception (GP)

Perception

call

feedback

Reasoning

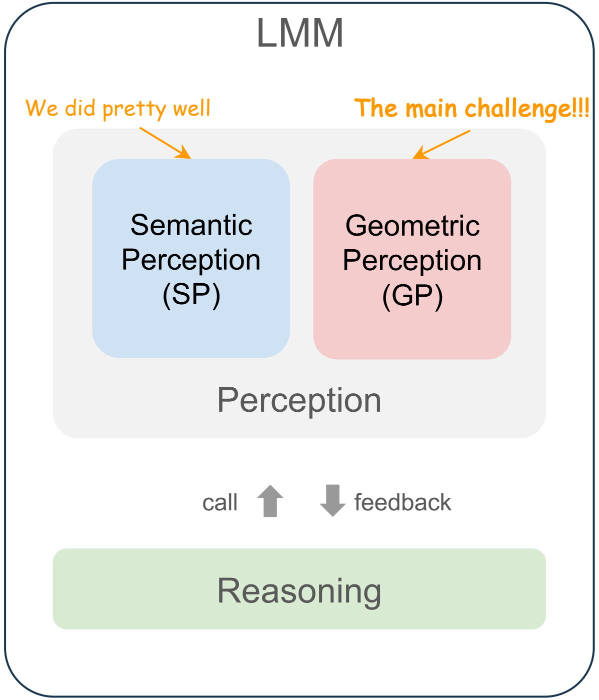


Two-streams Hypothesis an interesting human cognition analogy:

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 vs Geometric


LMM

We did pretty well

The main challenge!!!

Semantic Perception (SP)

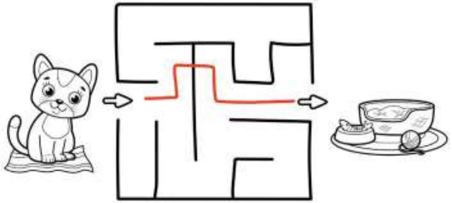
Geometric Perception (GP)

Perception

call

feedback

Reasoning



Semantic-centric Question:	**Expected response:**
Is there a dog or a cat in the image?	**[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:	**Expected response:**
Given that the black lines are walls that cannot be crossed, is the red line a valid path through the maze?	**[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 understanding is bad:

VL Encoders vs V-only Encoders

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

VL encoder (CLIP...):

LLaVA (= CLIP+LLM layer):

Why geometric understanding is bad:

VL Encoders vs V-only Encoders

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

semantic geometric

VL encoder (CLIP...):

semantic geometric

LLaVA (= CLIP+LLM layer):

semantic geometric

Why geometric understanding is bad:

VL Encoders vs V-only Encoders

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

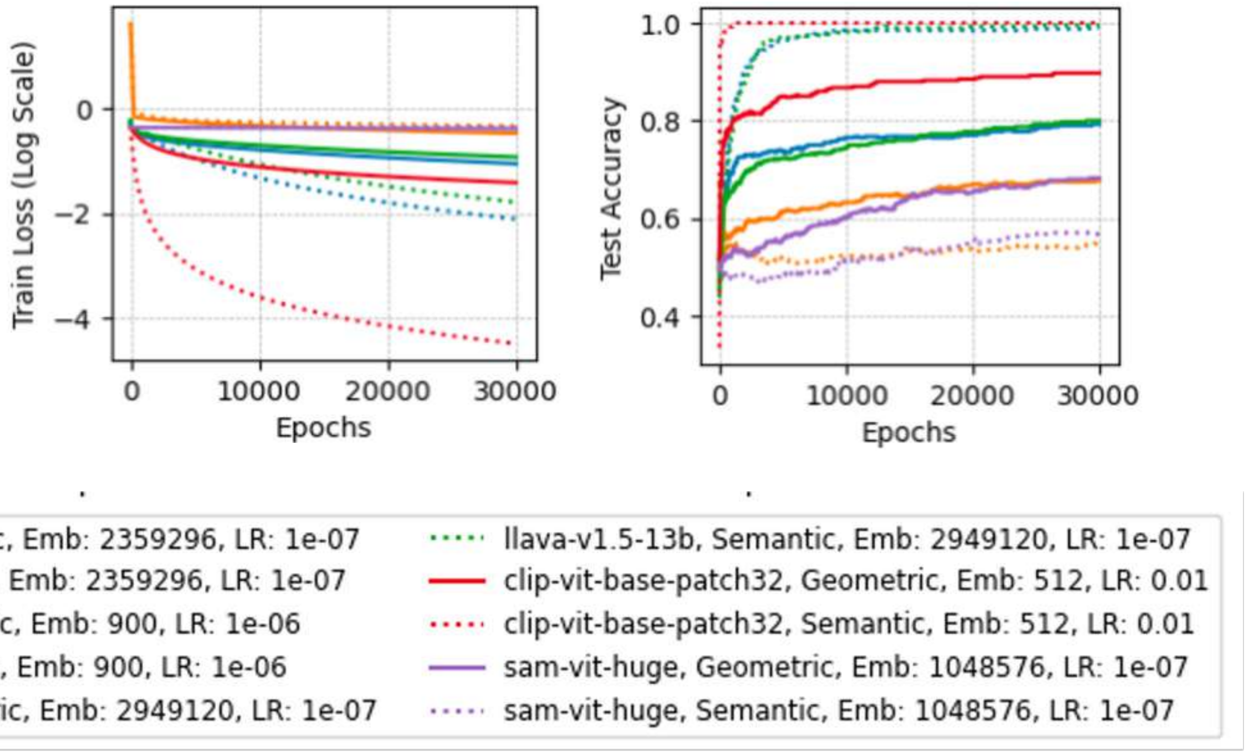
semantic << geometric

VL encoder (CLIP...):

semantic >> geometric

LLaVA (= CLIP+LLM layer):

semantic geometric



Why geometric understanding is bad:

VL Encoders vs V-only Encoders

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

semantic << geometric

VL encoder (CLIP...):

semantic >> geometric

LLaVA (= CLIP+LLM layer):

semantic geometric

Why geometric understanding is bad:

LLaVA < CLIP: LLM layer swallows the geometric features.

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

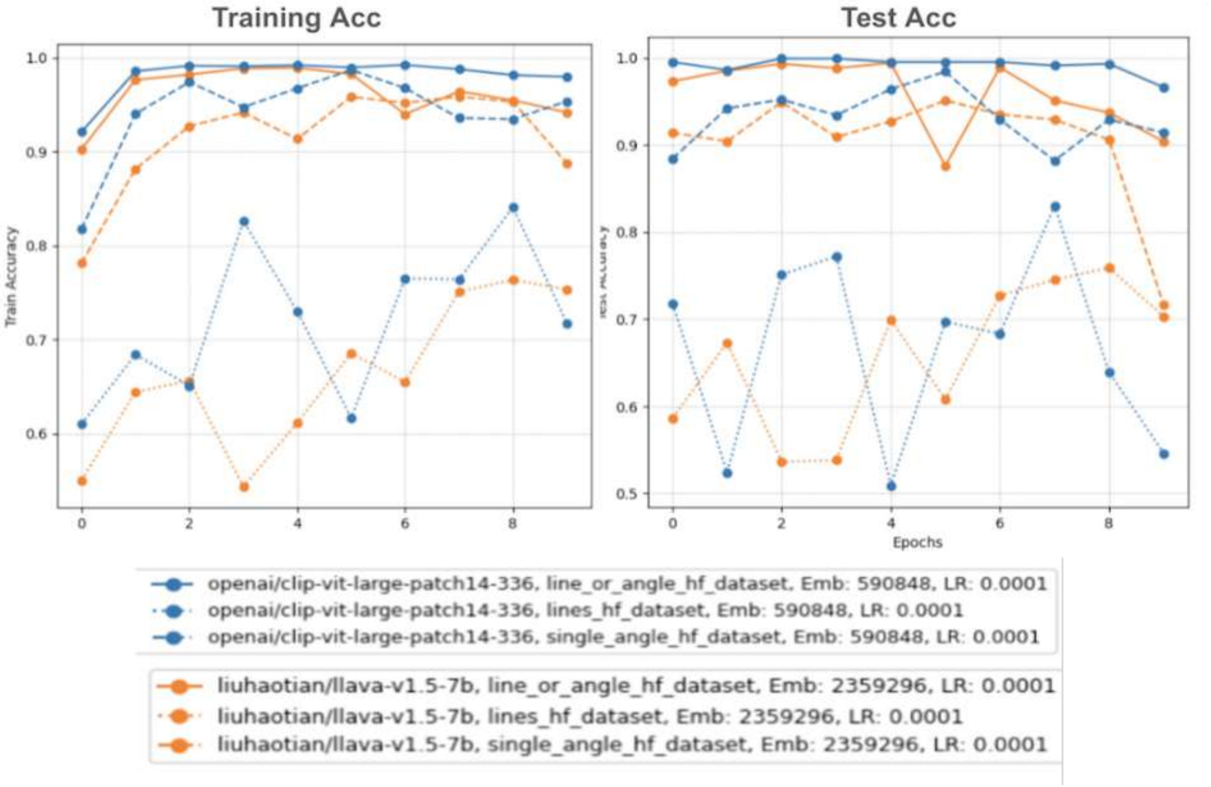
semantic << geometric

VL encoder (CLIP...):

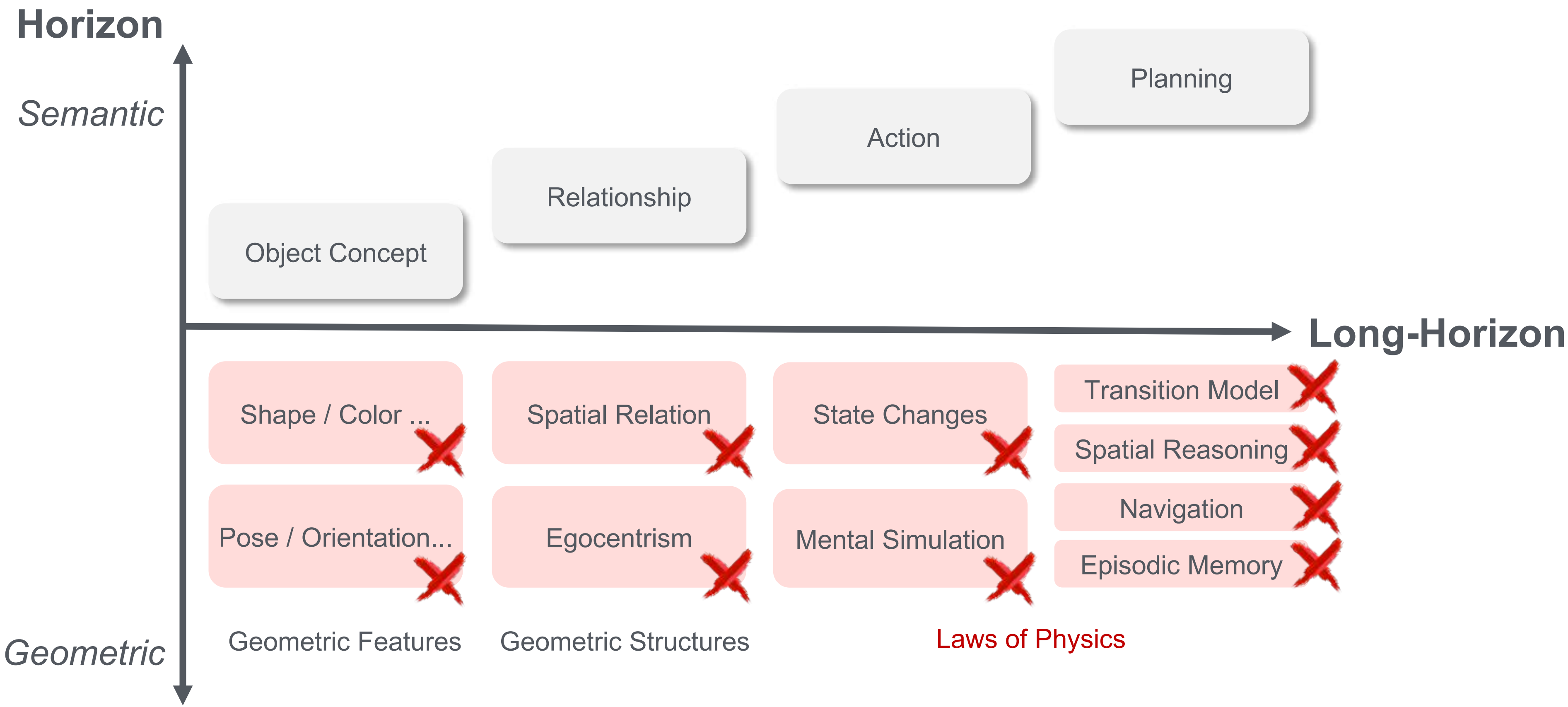
semantic >> geometric

LLaVA (= CLIP+LLM layer):

semantic >> geometric



Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

Shape / Color ... ✗

Pose / Orientation... ✗

Spatial Relation ✗

Egocentrism ✗

State Changes ✗

Mental Simulation ✗

Transition Model ✗

Spatial Reasoning ✗

Navigation ✗

Episodic Memory ✗

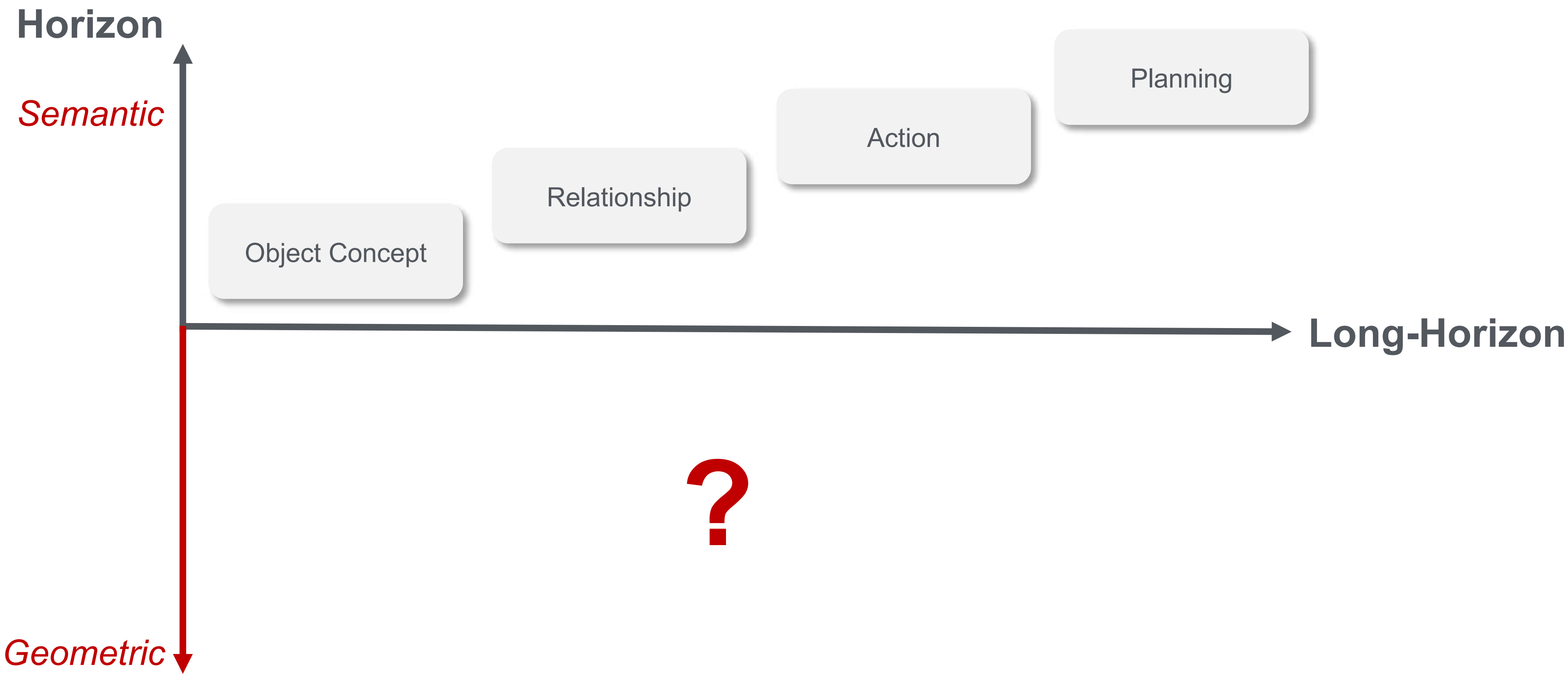
Geometric

Geometric Features

Geometric Structures

Laws of Physics

Missing knowledge about physical world



Horizon

Semantic

Object Concept

Relationship

Action

Planning

Long-Horizon

?

Geometric

My Goal: Physical World Knowledge

Horizon

Semantic

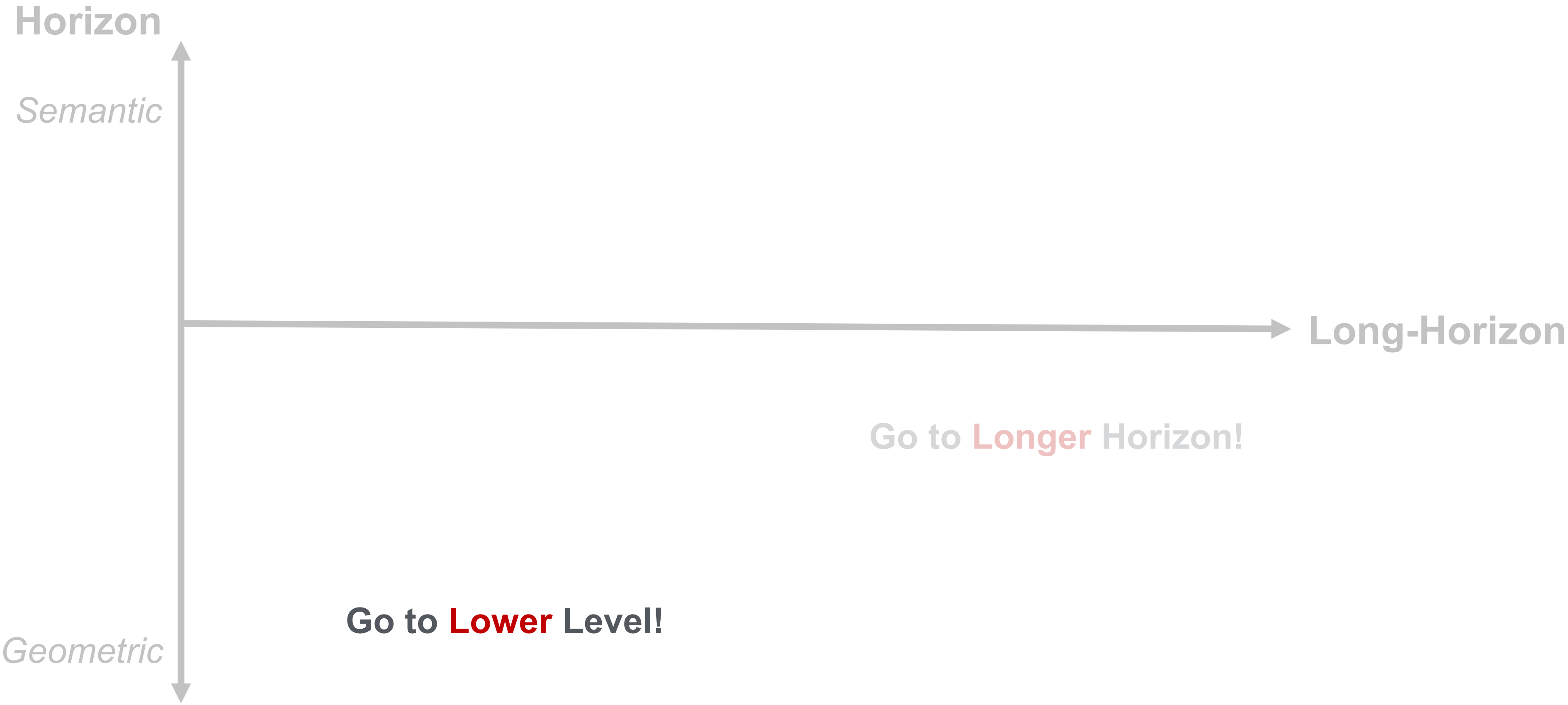
Long-Horizon

Go to Longer Horizon!

Go to Lower Level!

Geometric

My Goal: Physical World Knowledge



Horizon

Semantic

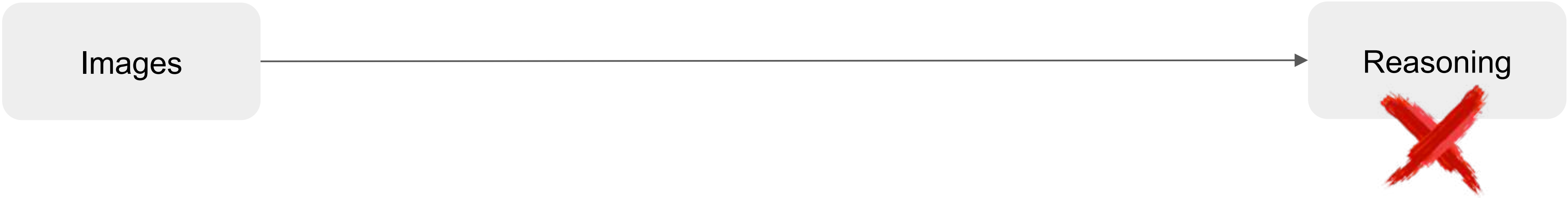
Long-Horizon

Go to **Longer** Horizon!

Go to **Lower** Level!

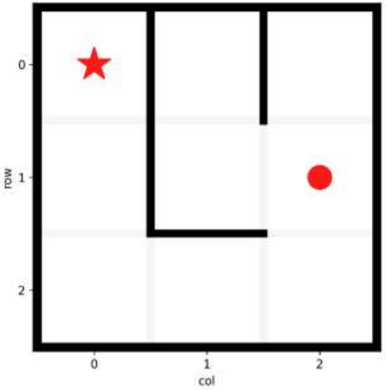
Geometric

What is the bottleneck?



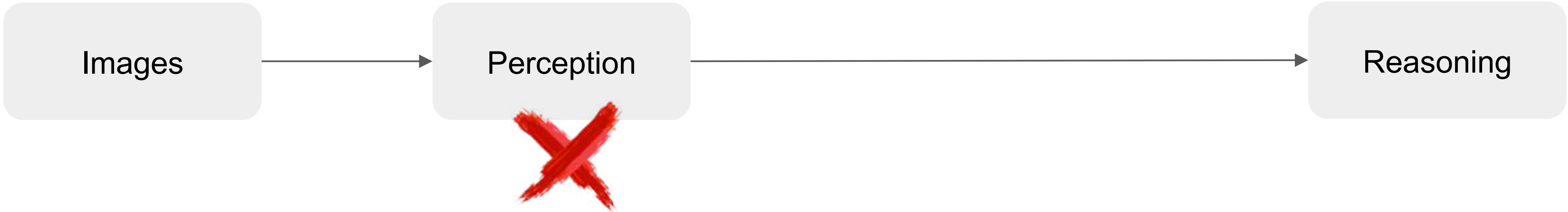
Images

Reasoning



More examples

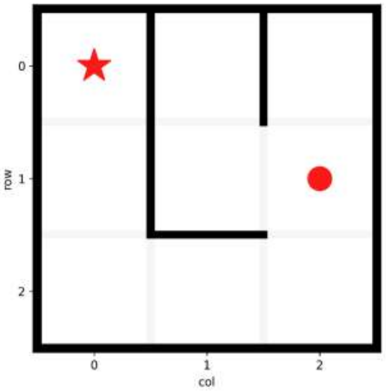
What is the bottleneck?

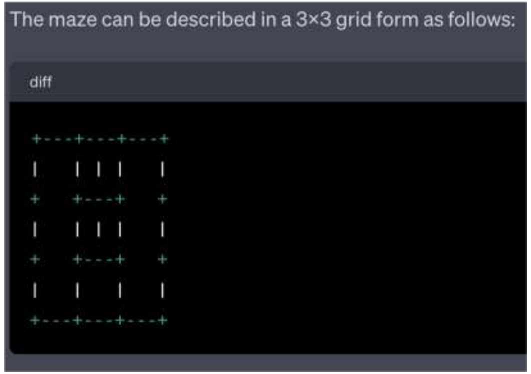


Images

Perception

Reasoning



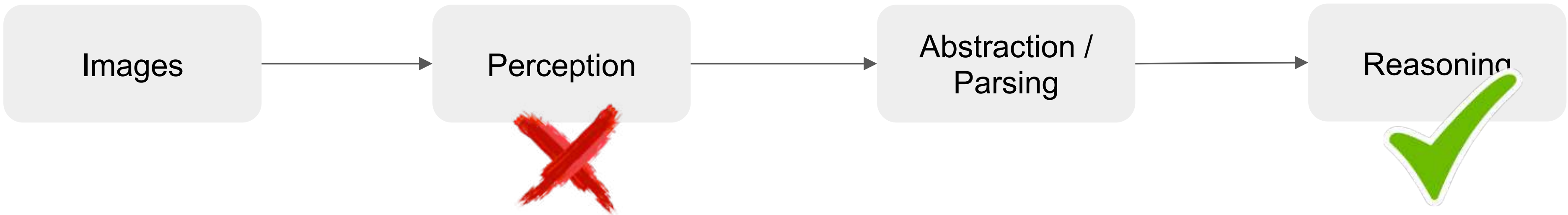


The maze can be described in a 3×3 grid form as follows:

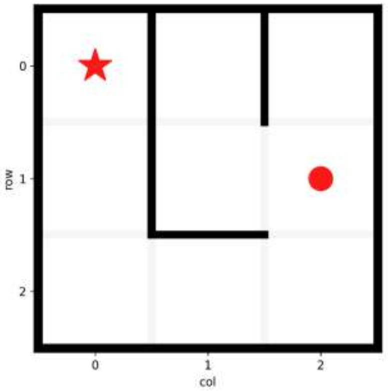
```
+---+---+---+
|   | | |   |
+   +---+   +
|   | | |   |
+   +---+   +
|   |   |   |
+---+---+---+
```

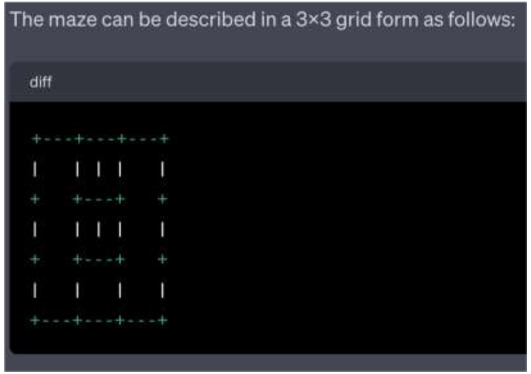
More examples

What is the bottleneck?



Images → Perception → Abstraction / Parsing → Reasoning



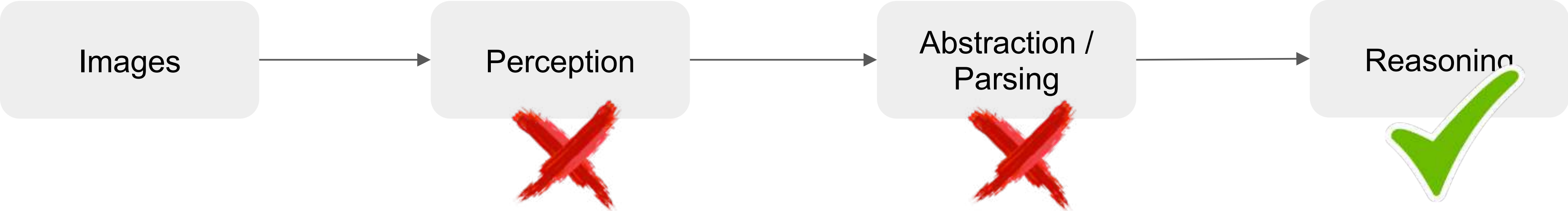


The maze can be described in a 3×3 grid form as follows:

```diff
+---+---+---+
|   | | |   |
+   +---+   +
|   | | |   |
+   +---+   +
|   |   |   |
+---+---+---+
```

More examples

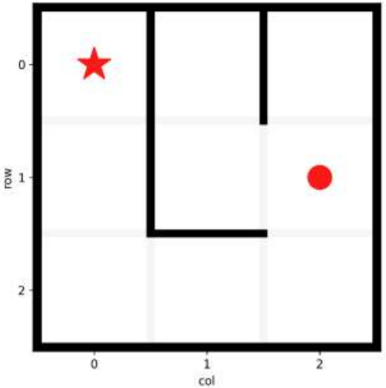
What is the bottleneck?

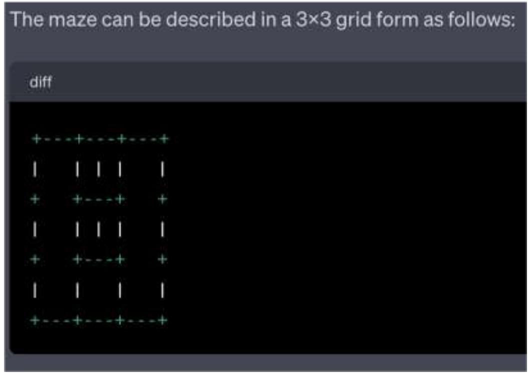

Images

Perception

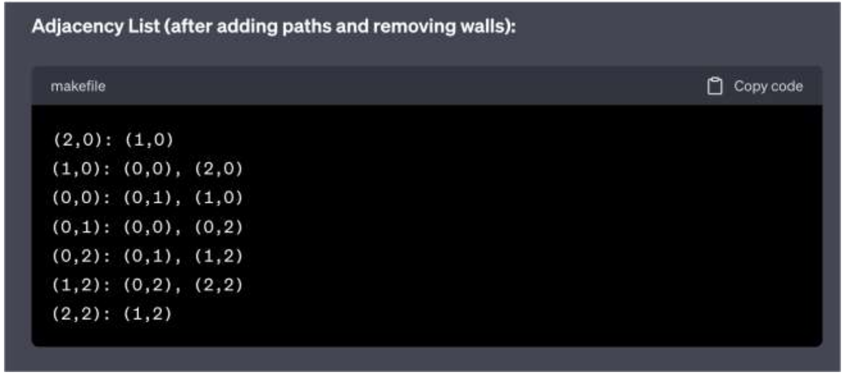
Abstraction / Parsing

Reasoning




The maze can be described in a 3×3 grid form as follows:

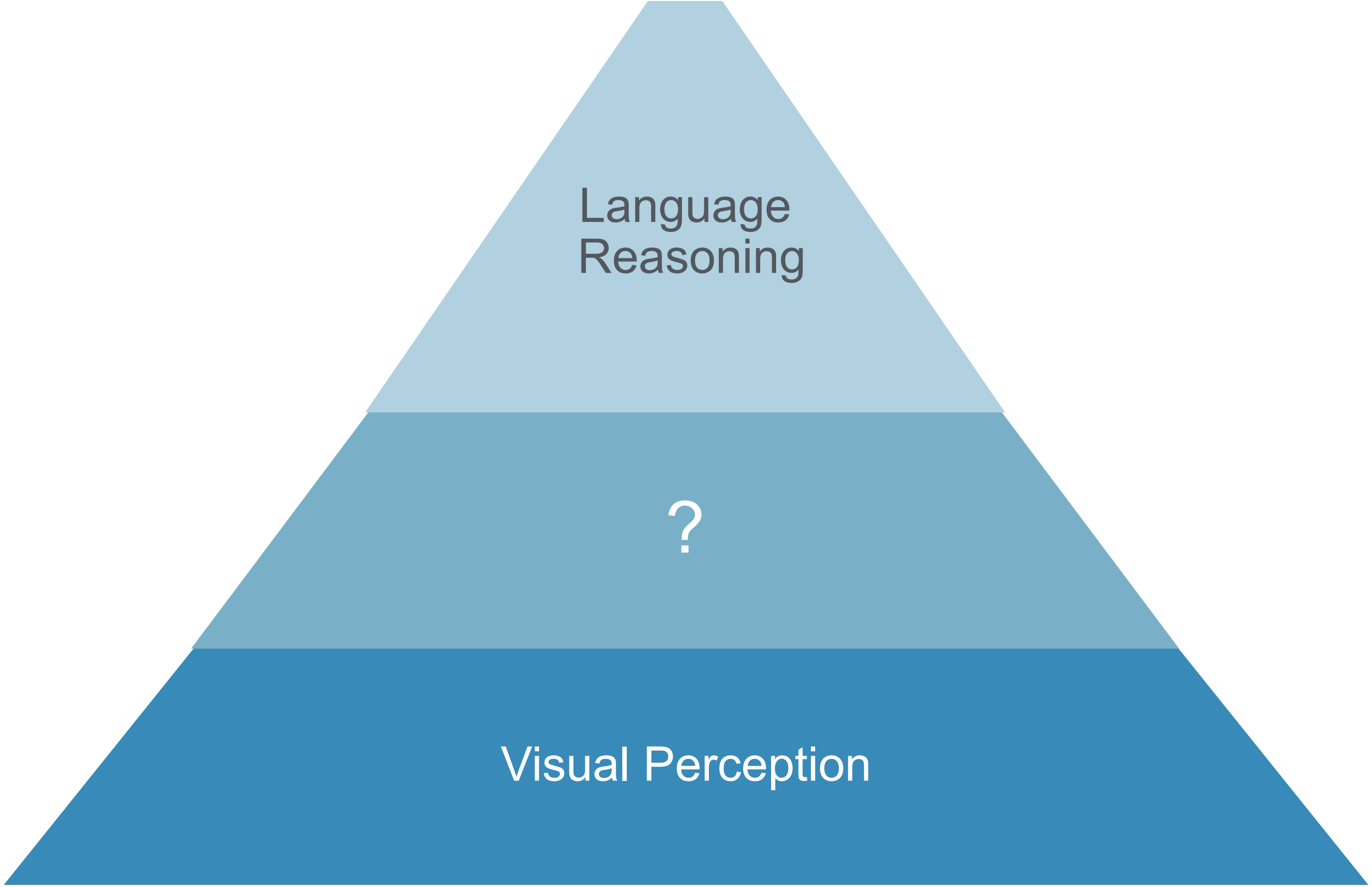
```
+---+---+---+
|   | | |   |
+   +---+   +
|   | | |   |
+   +---+   +
|   |   |   |
+---+---+---+
```


Adjacency List (after adding paths and removing walls):

```
(2,0): (1,0)
(1,0): (0,0), (2,0)
(0,0): (0,1), (1,0)
(0,1): (0,0), (0,2)
(0,2): (0,1), (1,2)
(1,2): (0,2), (2,2)
(2,2): (1,2)
```

More examples

What is Missing? Intermediate Layers in VLM Pyramid



Language Reasoning

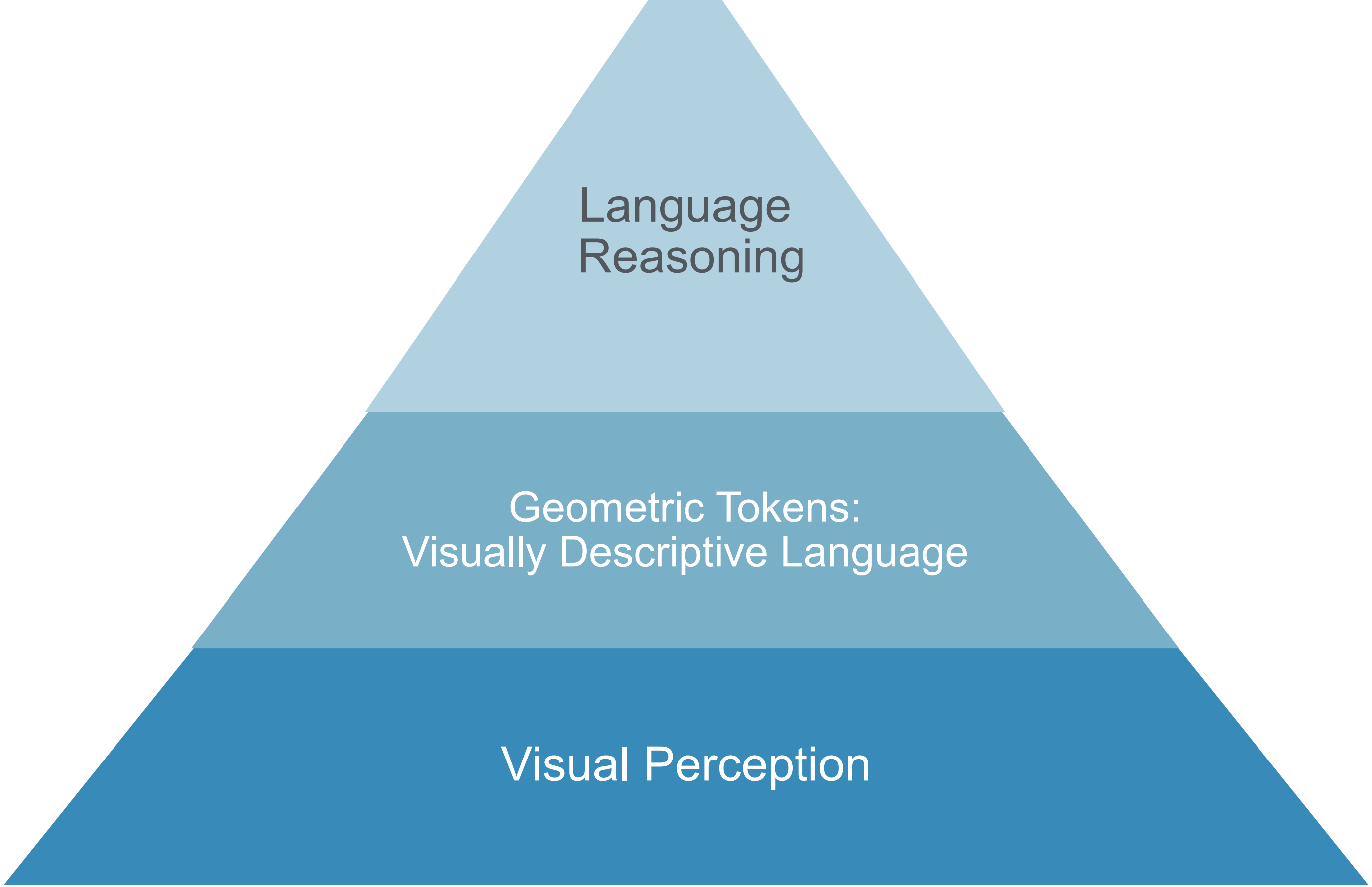
?

Visual Perception

Go to lower-level:

What is Missing? Intermediate Layers in VLM Pyramid

What is Missing? Intermediate Layers in VLM Pyramid



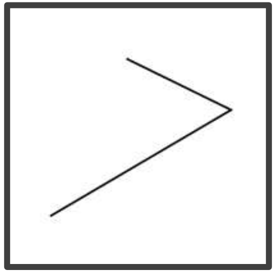
Language Reasoning

Geometric Tokens:
Visually Descriptive Language

Visual Perception

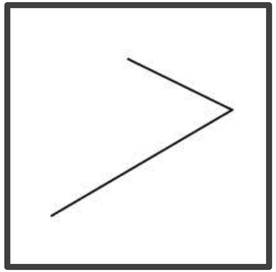
We need Abstraction (Geometric Tokens)
for positions, shapes, etc

Let us start from a very simple example



Is this a **line** or an **angle**? → **GPT-4V** → Angle ✓

high-level visual semantics

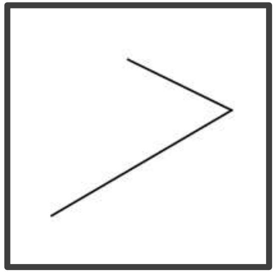


Is this an **acute** angle or an **obtuse** angle? → **GPT-4V** → Obtuse ✗

low-level visual details

Using **SVG** as intermediate representations?

We encode images with SVG for precise low-level perception



Input Image

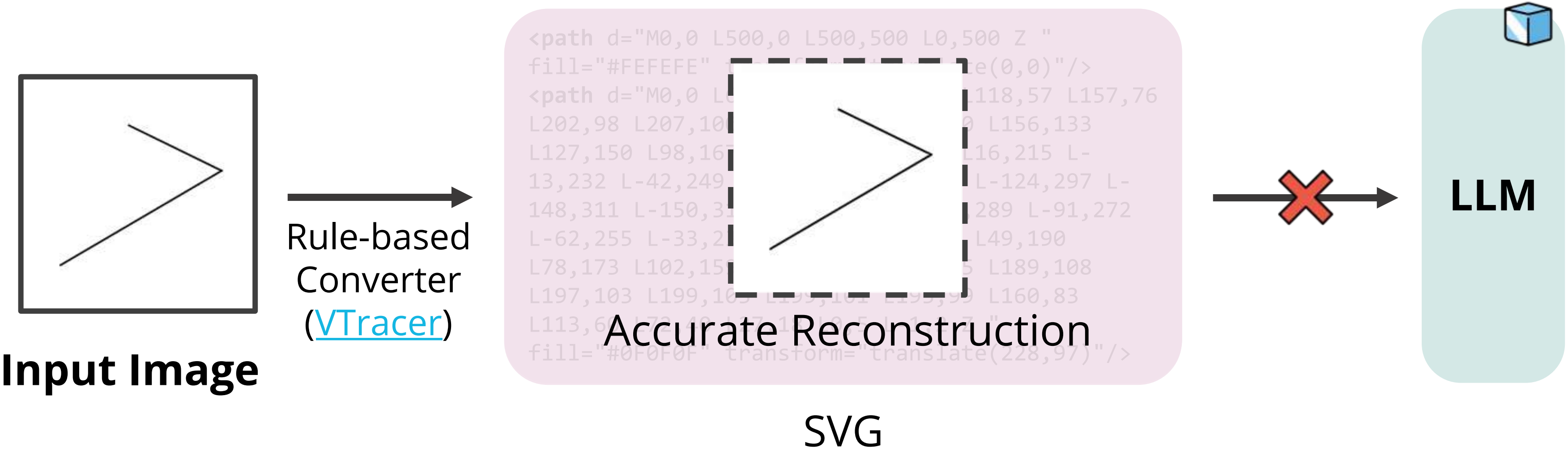
Rule-based Converter

```
<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

We encode image with SVG for precise low-level perception

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


Input Image
Rule-based Converter (VTracer)
<path d="M0,0 L500,0 L500,500 L0,500 Z "
fill="#FEFEFE"
Accurate Reconstruction
fill="#0F0F0F" transform="translate(228,97)"/>
SVG
LLM

What are properties of Geometric Tokens?

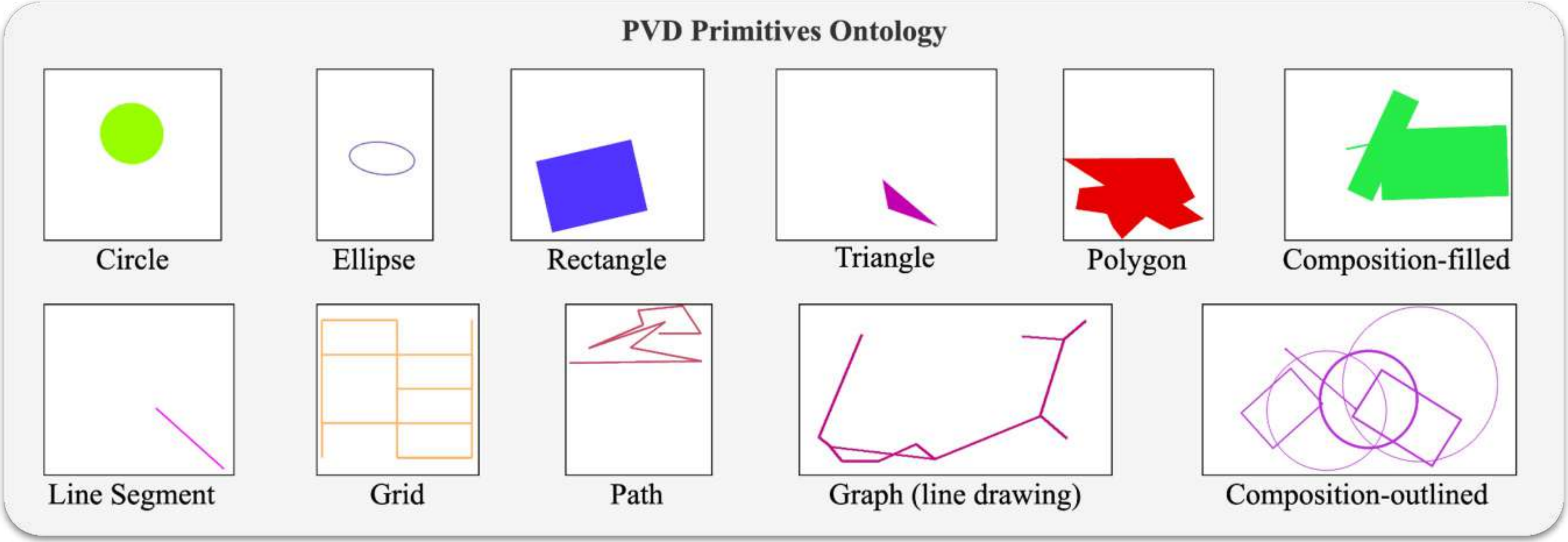
What are properties of Geometric Tokens?

Compositional
(Simple → Complicated)

Annotation-Free
(Synthetic Data)

Properties of Primal Visual Description (PVD)

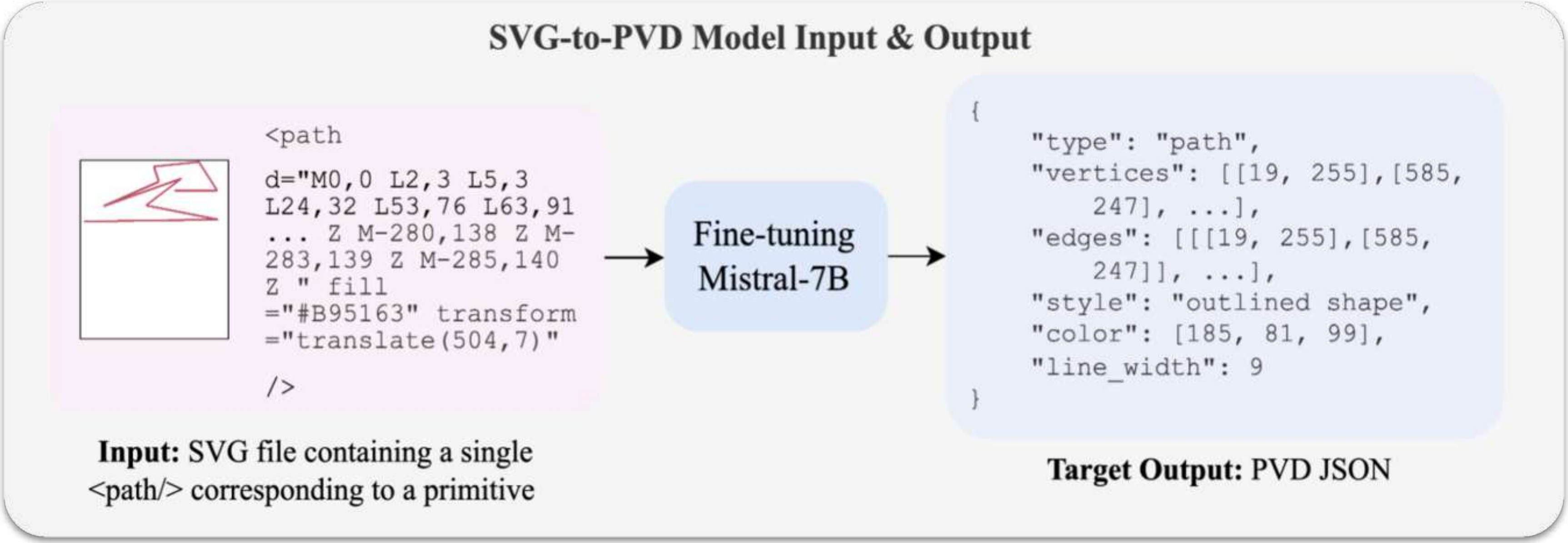
Property 1: Essential building blocks with a high coverage



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



```
<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

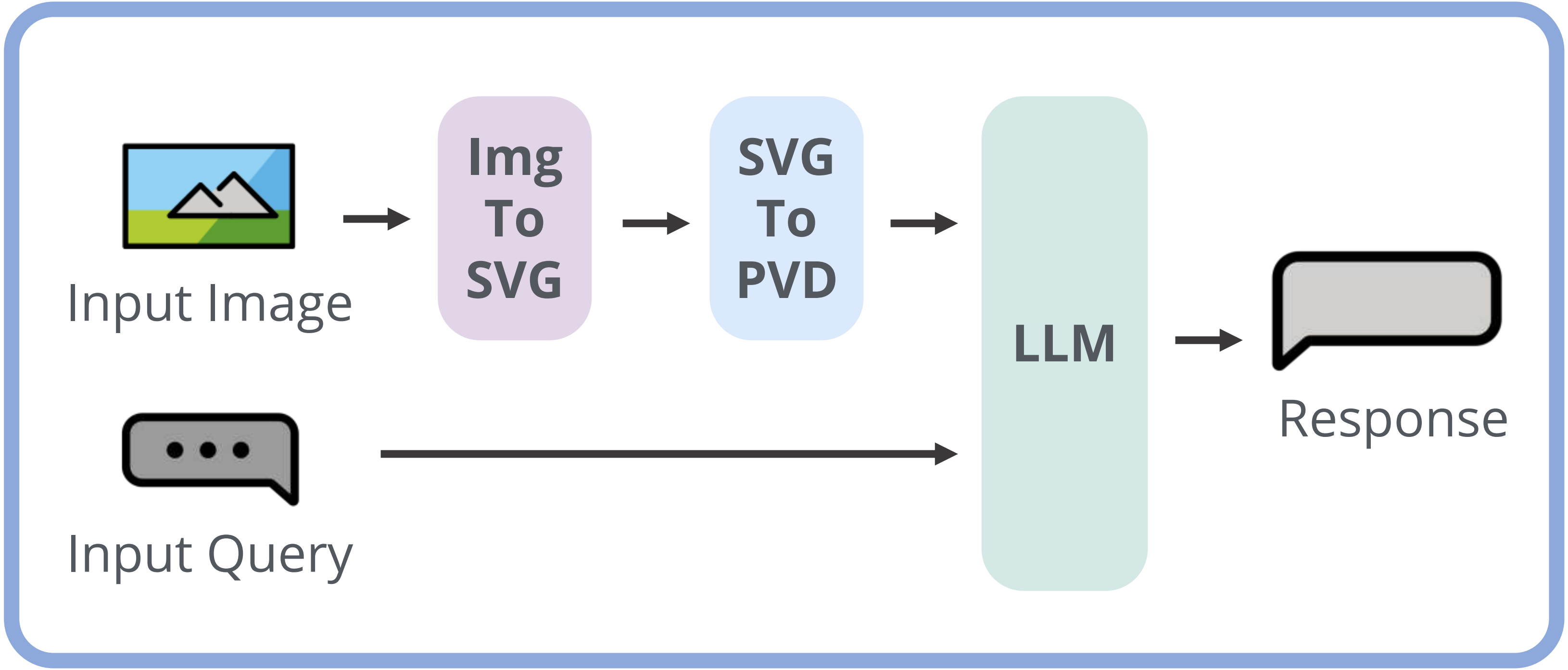
```
{
    "type": "path",
    "vertices": [[19, 255],[585,
        247], ...],
    "edges": [[[19, 255],[585,
        247]], ...],
    "style": "outlined shape",
    "color": [185, 81, 99],
    "line_width": 9
}
```

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



Input Image

Img To SVG

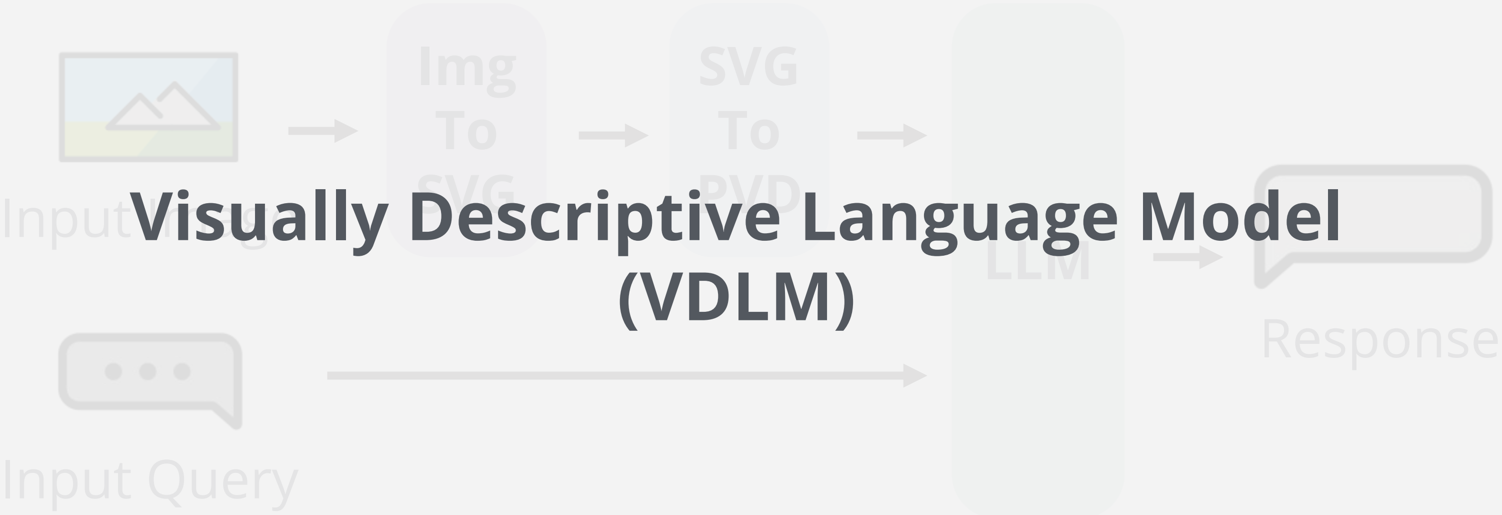
SVG To PVD

LLM

Response

Input Query

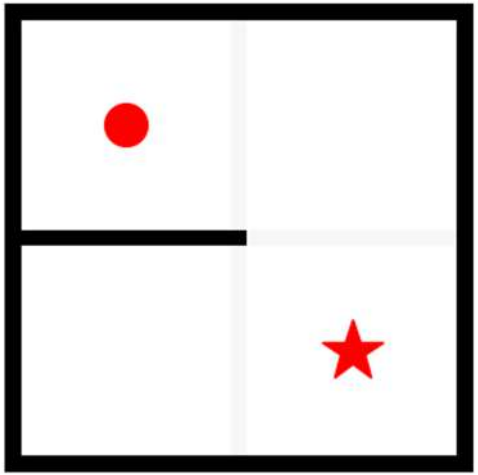
Overview



Visually Descriptive Language Model (VDLM)

Inference on Unseen Tasks

Zero-shot generalization to diverse tasks and domains

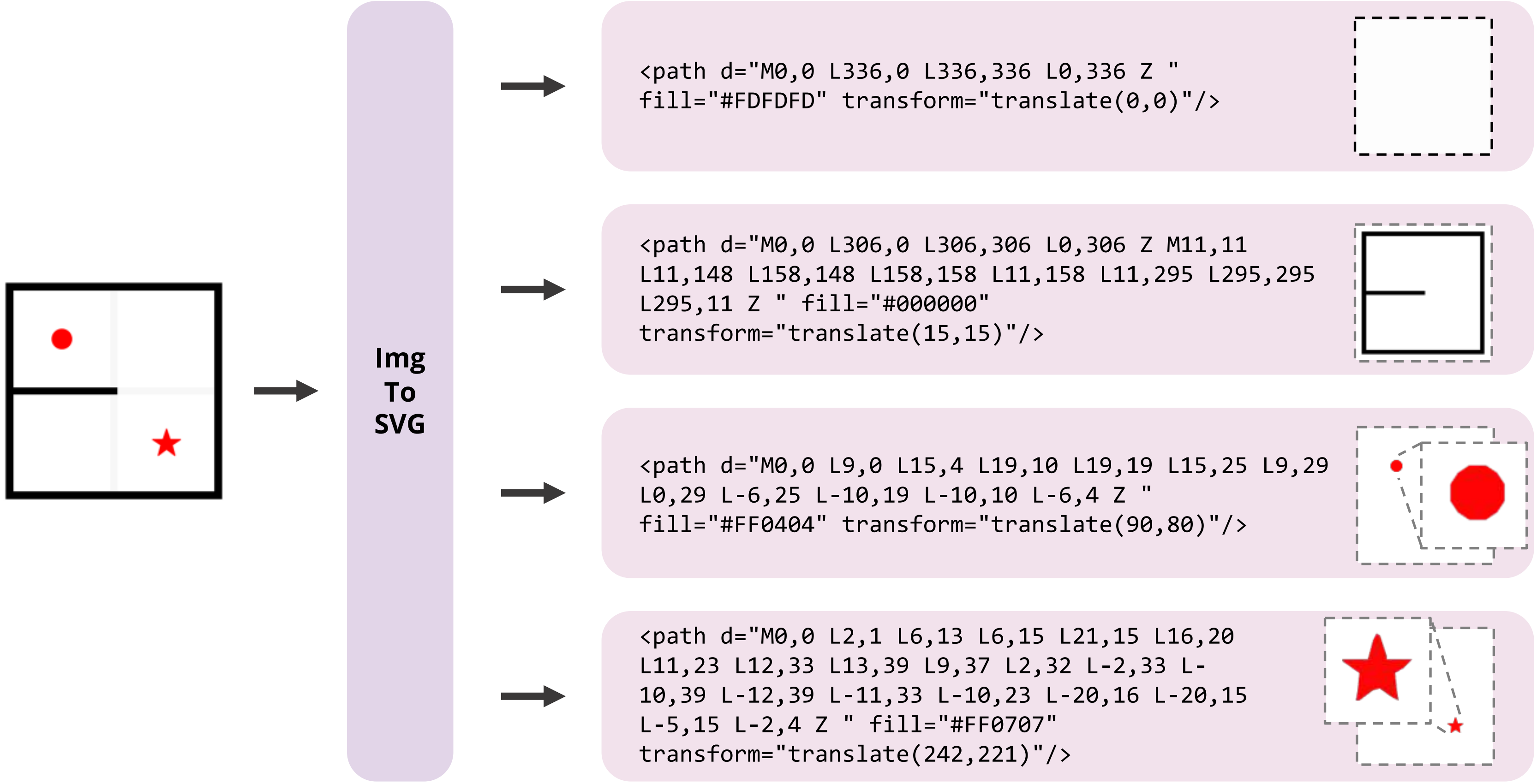


Input Image

In the 2x2 maze, walls are depicted by a grid of black lines ... The start position is marked by a red circle, and the end position by a red star.
Solve the maze ...

Input Query

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



Img To SVG

```
<path d="M0,0 L336,0 L336,336 L0,336 Z "
fill="#FDFDFD" transform="translate(0,0)"/>
```

```
<path d="M0,0 L306,0 L306,306 L0,306 Z M11,11
L11,148 L158,148 L158,158 L11,158 L11,295 L295,295
L295,11 Z " fill="#000000"
transform="translate(15,15)"/>
```

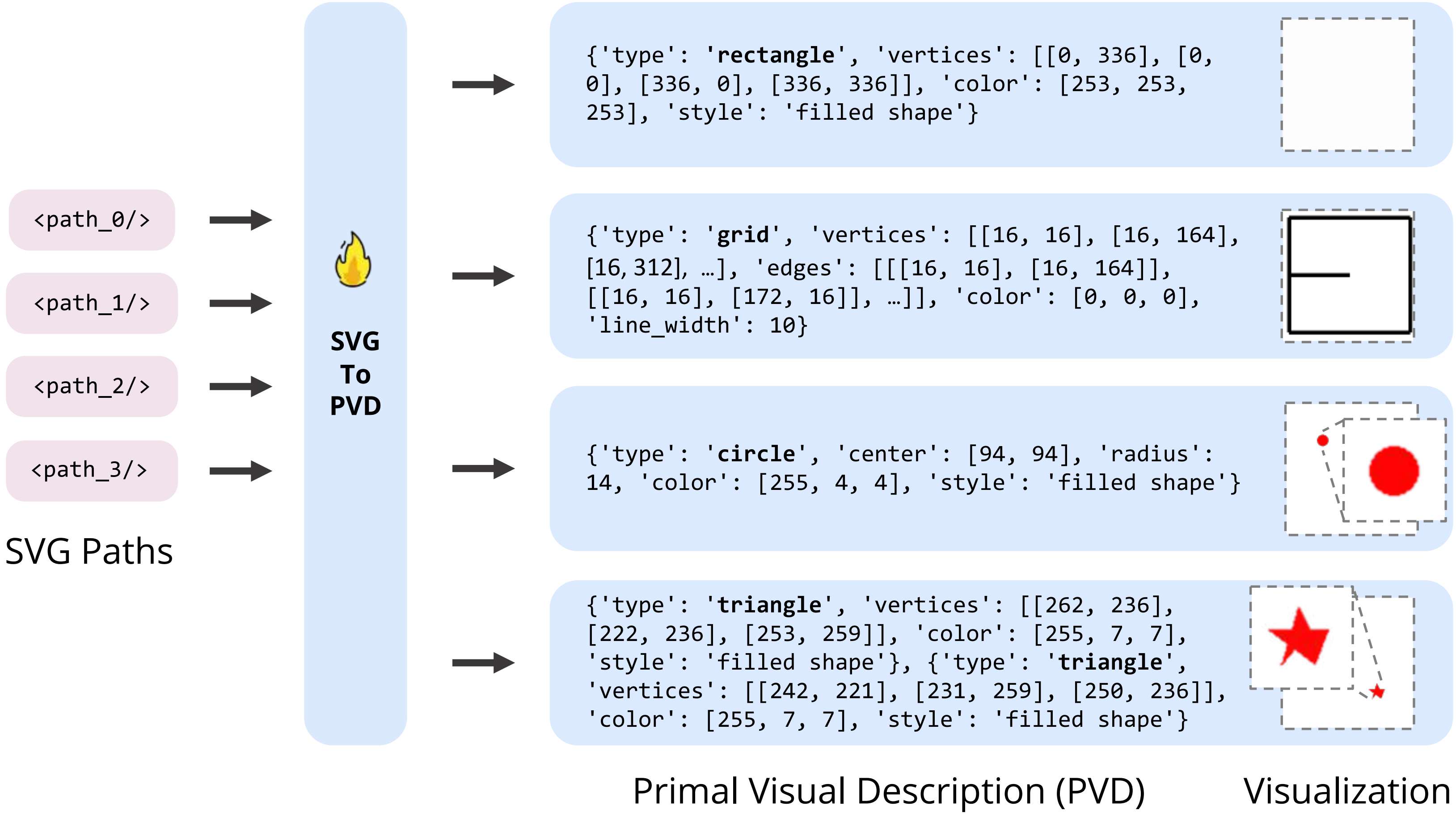
```
<path d="M0,0 L9,0 L15,4 L19,10 L19,19 L15,25 L9,29
L0,29 L-6,25 L-10,19 L-10,10 L-6,4 Z "
fill="#FF0404" transform="translate(90,80)"/>
```

```
<path d="M0,0 L2,1 L6,13 L6,15 L21,15 L16,20
L11,23 L12,33 L13,39 L9,37 L2,32 L-2,33 L-
10,39 L-12,39 L-11,33 L-10,23 L-20,16 L-20,15
L-5,15 L-2,4 Z " fill="#FF0707"
transform="translate(242,221)"/>
```

Decomposed Single SVG Paths

Visualization

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



{'type': '**rectangle**', 'vertices': [[0, 336], [0, 0], [336, 0], [336, 336]], 'color': [253, 253, 253], 'style': 'filled shape'}

{'type': '**grid**', 'vertices': [[16, 16], [16, 164], [16, 312], …], 'edges': [[[16, 16], [16, 164]], [[16, 16], [172, 16]], …]], 'color': [0, 0, 0], 'line_width': 10}

{'type': '**circle**', 'center': [94, 94], 'radius': 14, 'color': [255, 4, 4], 'style': 'filled shape'}

{'type': '**triangle**', 'vertices': [[262, 236], [222, 236], [253, 259]], 'color': [255, 7, 7], 'style': 'filled shape'}, {'type': '**triangle**', 'vertices': [[242, 221], [231, 259], [250, 236]], 'color': [255, 7, 7], 'style': 'filled shape'}

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

`{object_0}` `{object_2}`

`{object_1}` `{object_3}`

Aggregated PVD

In the 2x2 maze, walls are depicted by a grid of black lines ...
The start position is marked by a red circle, and the end position by a red star.
Solve the maze ...

Input Query

LLM

### Step 1: Infer the Connectivity of the Cells

- Vertical connections: **`[[True, True]]`**
- Horizontal connections: **`[[False], [True]]`**

### Step 2: Infer the Start and End Positions

- The **start position** is marked by a **red circle**, which is located at approximately `[94, 94]` ... in the top-left cell, or **`(0, 0)`**.
- The **end position** is marked by a **composite red shape resembling a star**, located in the bottom-right cell ... This places the end position at **`(1, 1)`**.

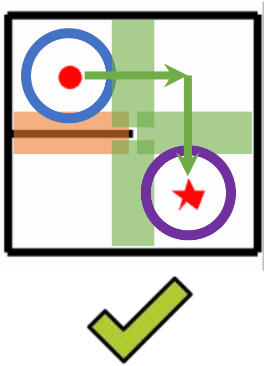
### Step 3: Solve the Maze

Given the connectivity and the start/end positions, we can find a path from the start to the end...

1. **Start at** `(0, 0)`.
2. **Move right** to `(0, 1)`.
3. **Move down** to `(1, 1)`.

Thus, the solution path is: **`[(0, 0), (0, 1), (1, 1)]`**.

GPT-4
Response



Visually Descriptive Language For Vector Graphics Reasoning

Zhenhailong Wang

Joy Hsu

Xingyao Wang

Kuan-Hao Huang

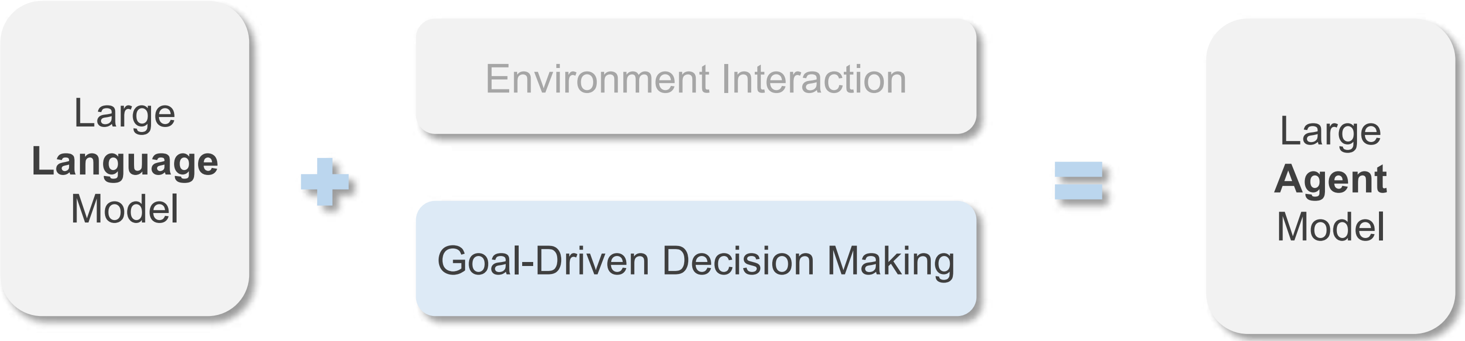
Manling Li

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

Large **Language** Models → Large **Agent** Models



Large **Language** Model

+

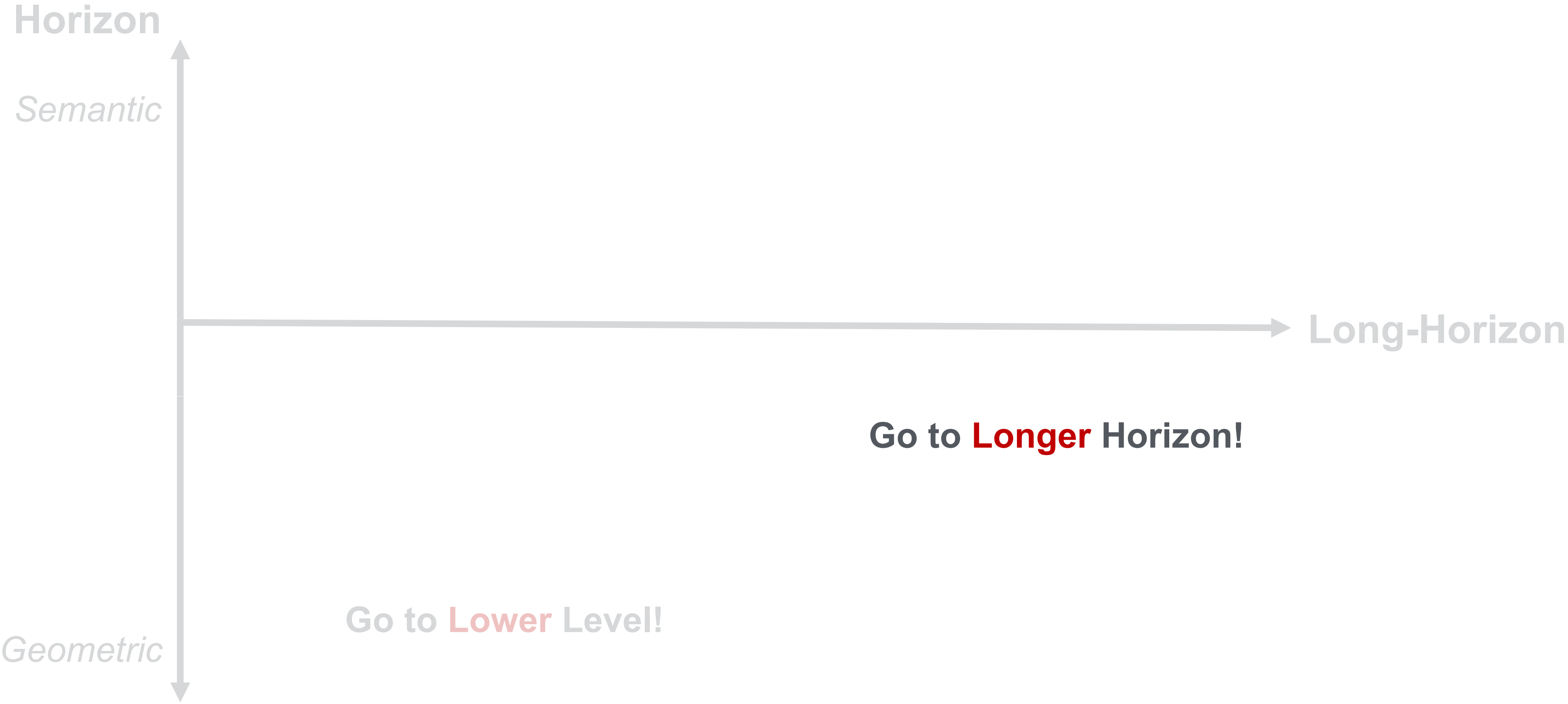
Environment Interaction

Goal-Driven Decision Making

=

Large **Agent** Model

My Goal: Physical World Knowledge



Horizon

Semantic

Long-Horizon

Go to **Longer** Horizon!

Go to **Lower** Level!

Geometric

Embodied Agent Interface:

A single line to benchmark LLMs for Embodied Decision Making

NeurIPS 2024 DB (Oral, 1%)

Manling Li

Shiyu Zhao*

Qineng Wang*

Kangrui Wang*

Bryan Yu Zhou*

Sanjana Srivastava

Tony Lee

Erran Li

Ruohan Zhang

Weiyu Liu

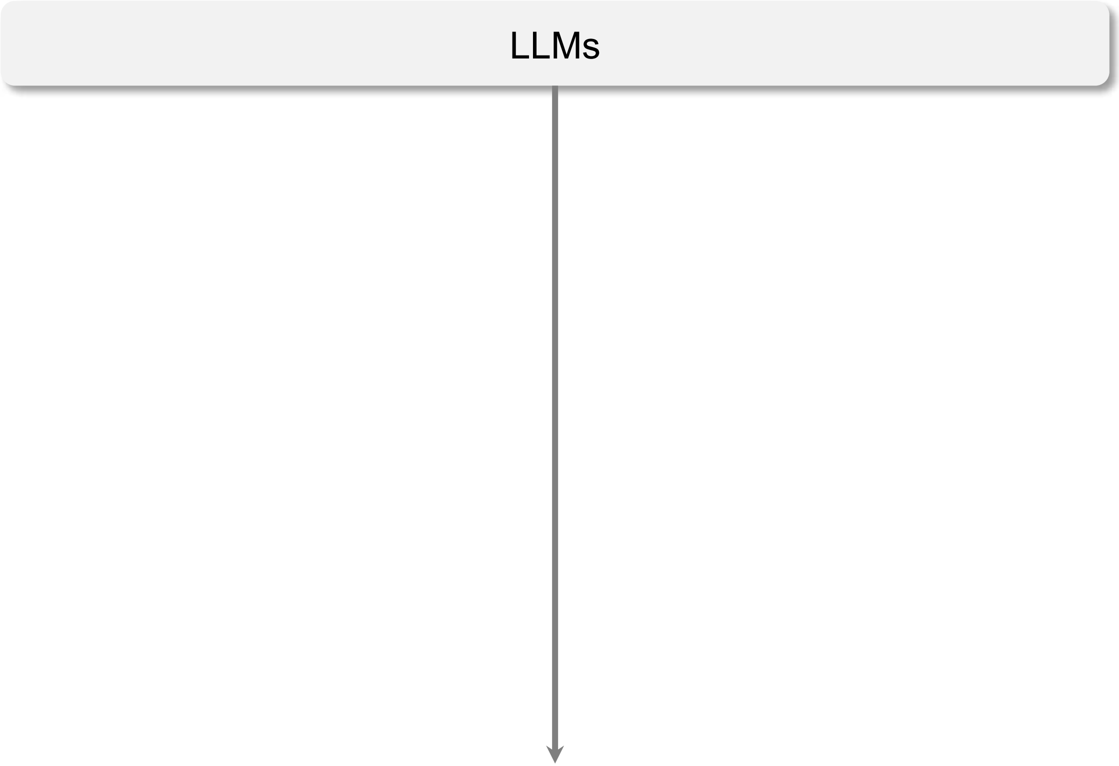
Percy Liang

Fei-Fei Li

Jiayuan Mao

Jiajun Wu

LLMs for Embodied Decision Making

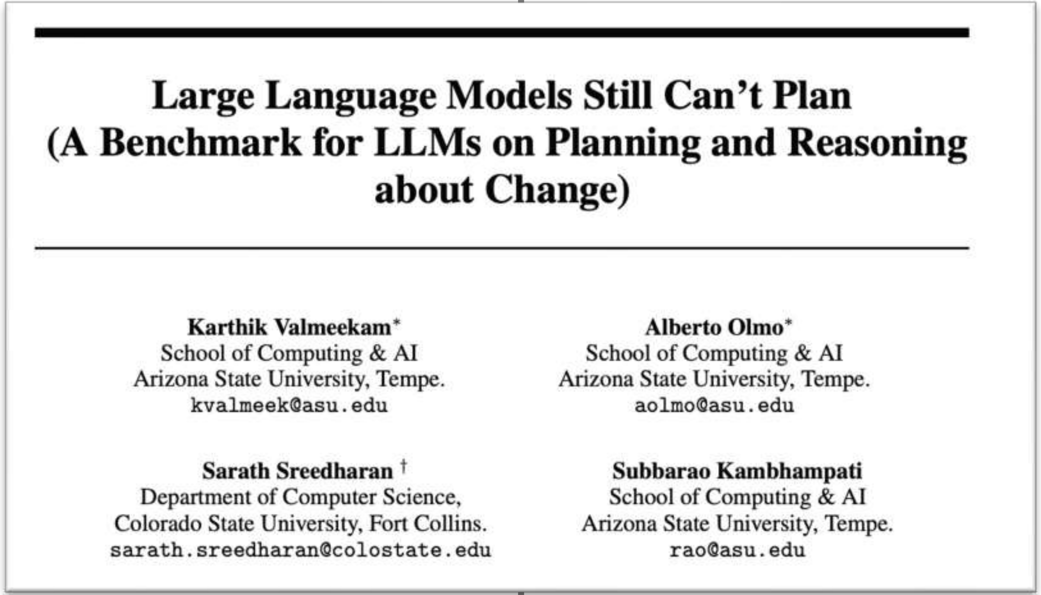


LLMs

Embodied Decision Making

LLMs still cannot plan (Valmeekam, et al, 2024)

LLMs



Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)

Karthik Valmeekam[*]
School of Computing & AI
Arizona State University, Tempe.
kvalmeek@asu.edu

Alberto Olmo[*]
School of Computing & AI
Arizona State University, Tempe.
aolmo@asu.edu

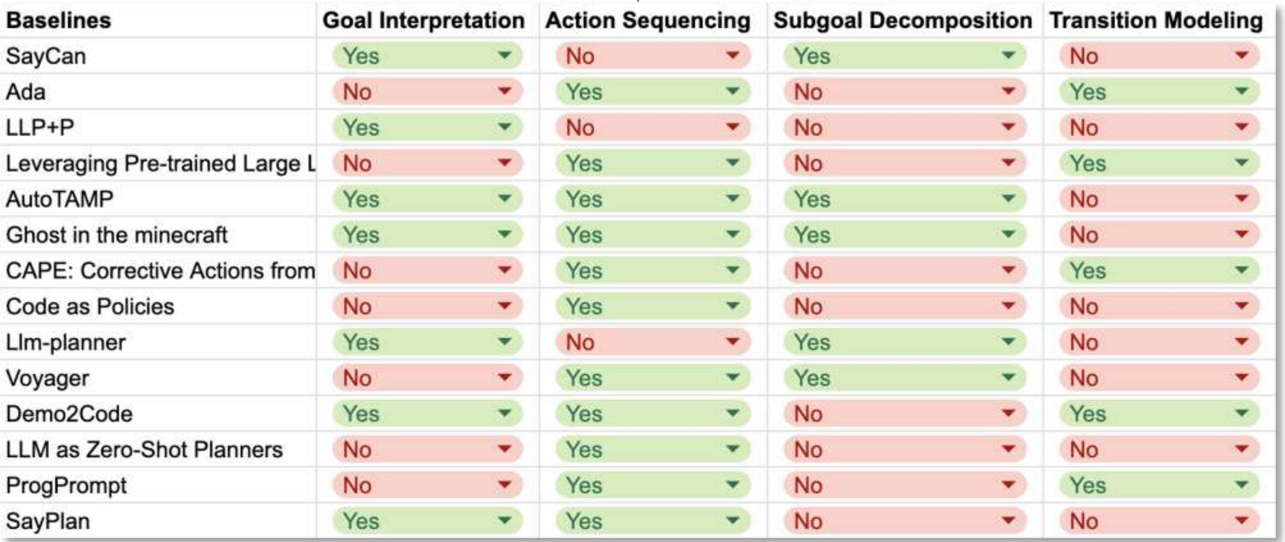
Sarath Sreedharan [†]
Department of Computer Science,
Colorado State University, Fort Collins.
sarath.sreedharan@colostate.edu

Subbarao Kambhampati
School of Computing & AI
Arizona State University, Tempe.
rao@asu.edu

Embodied Decision Making

So many different ways of using LLMs

LLMs



Baselines	Goal Interpretation	Action Sequencing	Subgoal Decomposition	Transition Modeling
SayCan	Yes	No	Yes	No
Ada	No	Yes	No	Yes
LLP+P	Yes	No	No	No
Leveraging Pre-trained Large L	No	Yes	No	Yes
AutoTAMP	Yes	Yes	Yes	No
Ghost in the minecraft	Yes	Yes	Yes	No
CAPE: Corrective Actions from	No	Yes	No	Yes
Code as Policies	No	Yes	No	No
Llm-planner	Yes	No	Yes	No
Voyager	No	Yes	Yes	No
Demo2Code	Yes	Yes	No	Yes
LLM as Zero-Shot Planners	No	Yes	No	No
ProgPrompt	No	Yes	No	Yes
SayPlan	Yes	Yes	No	No

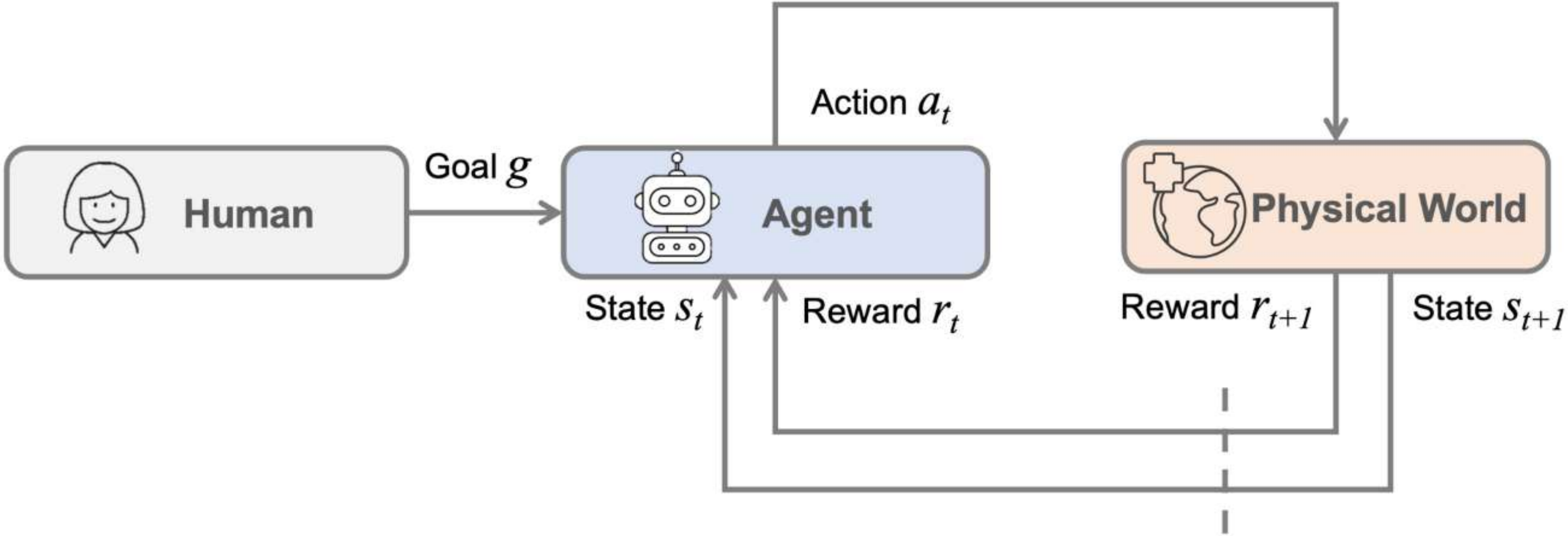
…

Embodied Decision Making

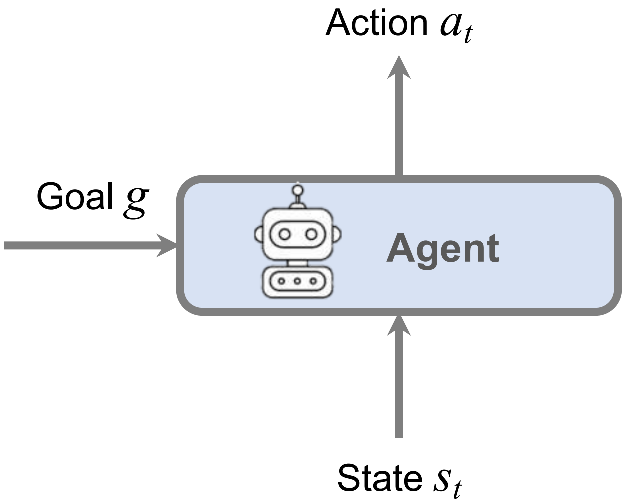
So we need **Standardization!**

We need an Embodied Agent Interface**!**

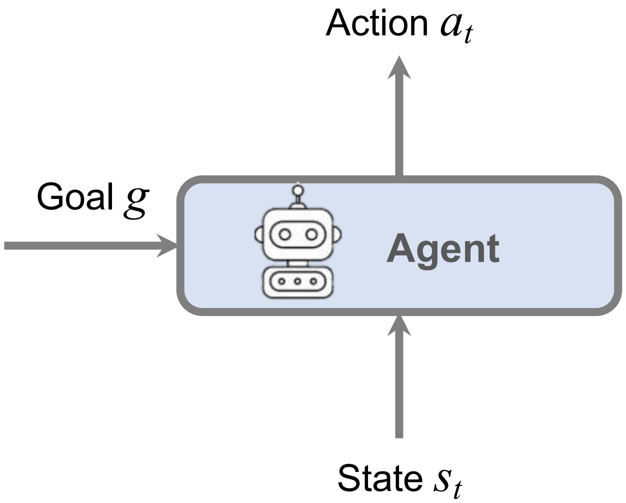
Trajectory: Let us go back to MDP



Trajectory: essentially modeling a MDP



Trajectory: essentially modeling a MDP

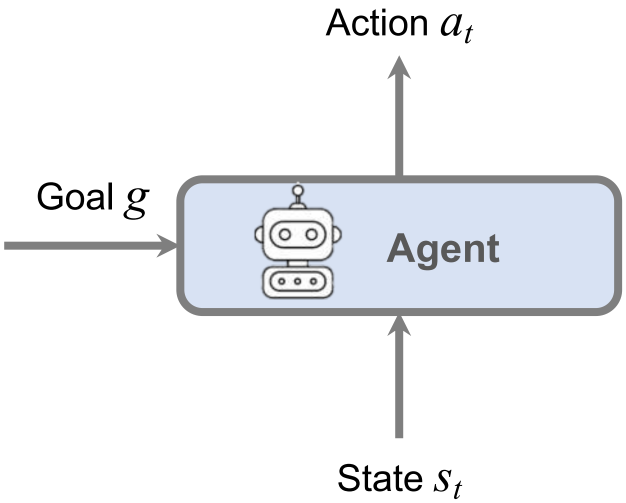


Input

$<g, s_0>$

Goal Interpretation

Trajectory: MDP



Output

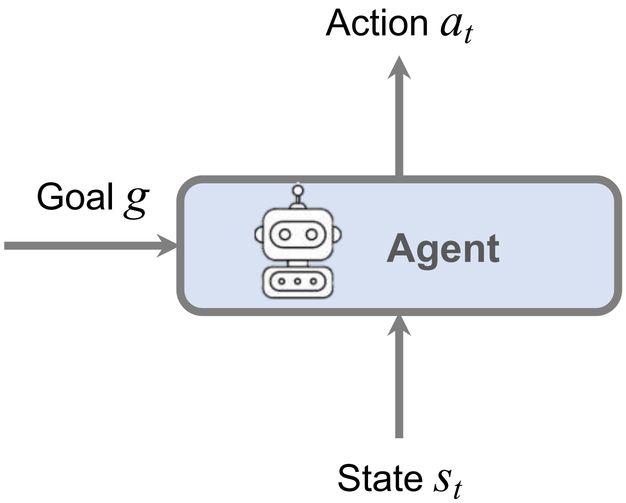
$<a_1, a_2, \ldots, a_t>$ **Action Sequencing**

$<s_1, s_2, \ldots, s_t>$ **Subgoal Decomposition**

Input

$<g, s_0>$ **Goal Interpretation**

Trajectory: MDP



Action a_t

Goal g

Agent

State s_t

Output	$<a_1, a_2, \ldots, a_t>$	**Action Sequencing**
	$<s_1, s_2, \ldots, s_t>$	**Subgoal Decomposition**
Learn	$(s_t, a_t) \rightarrow s_{t+1}$	**Transition Modeling**
Input	$<g, s_0>$	**Goal Interpretation**

Standardization based on MDP

Action Sequencing

Subgoal Decomposition

Transition Modeling

Goal Interpretation

Embodied Agent Interface:

So we treat them as ability modules in our interface

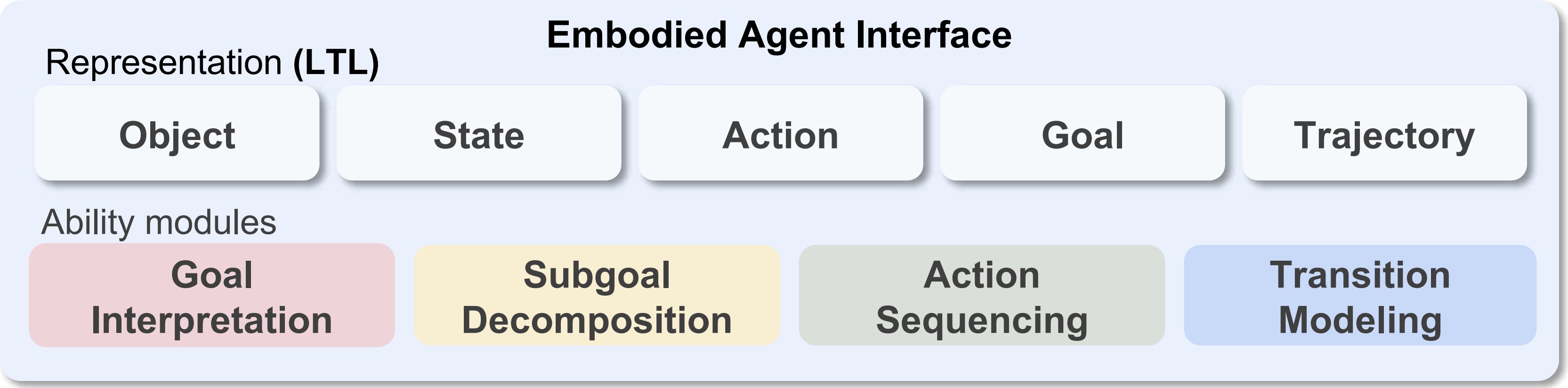
Goal Interpretation

Subgoal Decomposition

Action Sequencing

Transition Modeling

Embodied Agent Interface



Embodied Agent Interface

Representation **(LTL)**

Object

State

Action

Goal

Trajectory

Ability modules

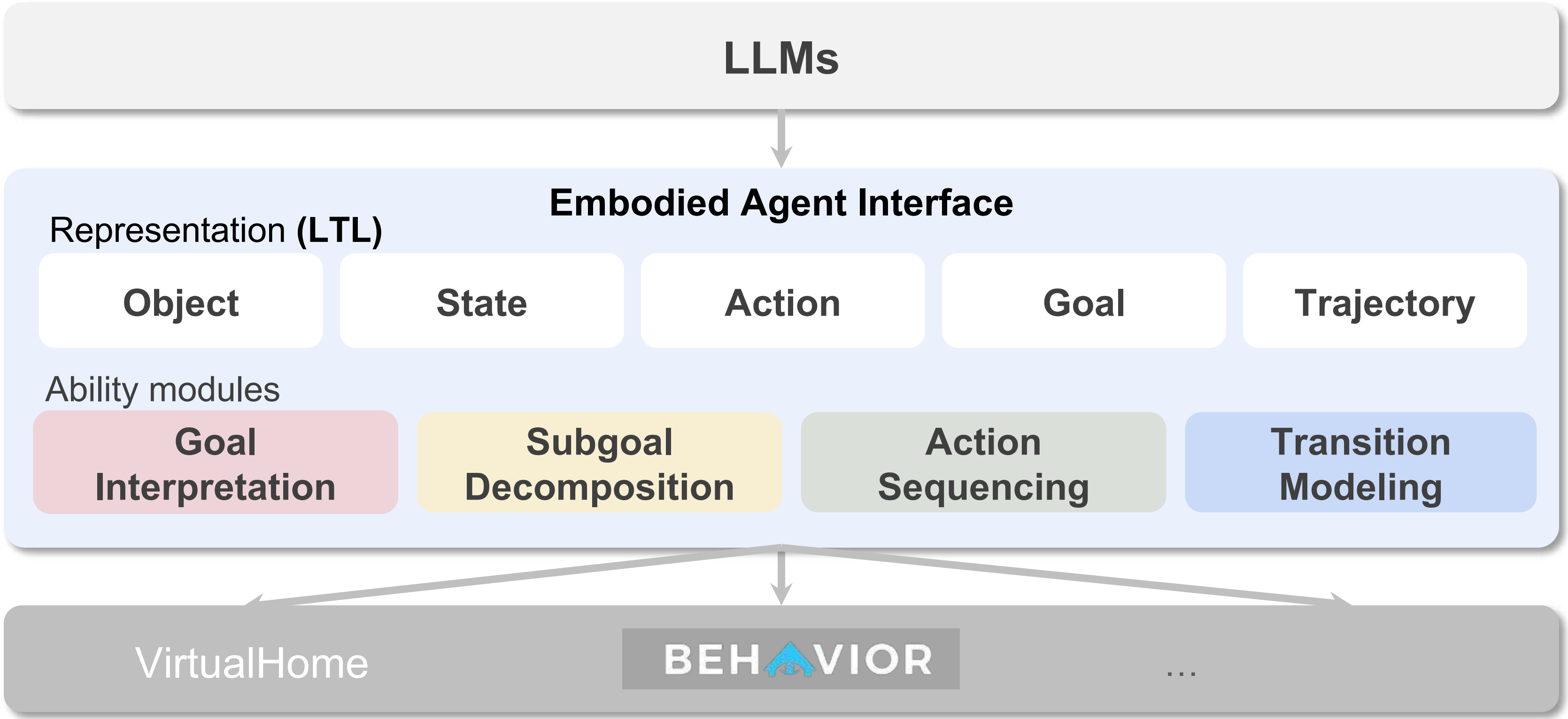
Goal Interpretation

Subgoal Decomposition

Action Sequencing

Transition Modeling

Embodied Agent Interface



LLMs

Embodied Agent Interface

Representation **(LTL)**

Object | **State** | **Action** | **Goal** | **Trajectory**

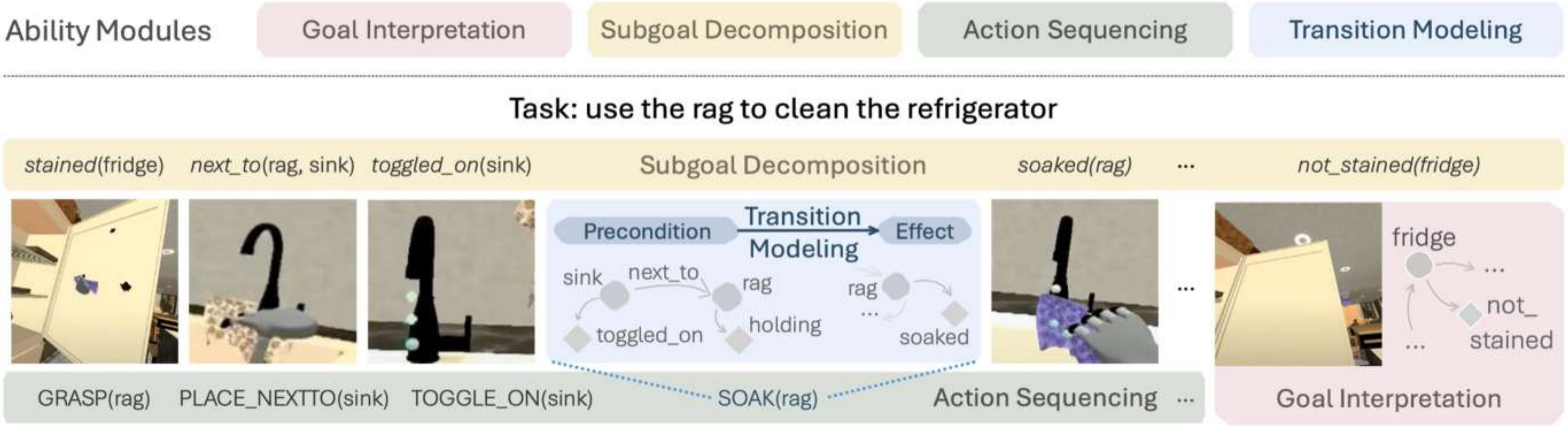
Ability modules

Goal Interpretation | **Subgoal Decomposition** | **Action Sequencing** | **Transition Modeling**

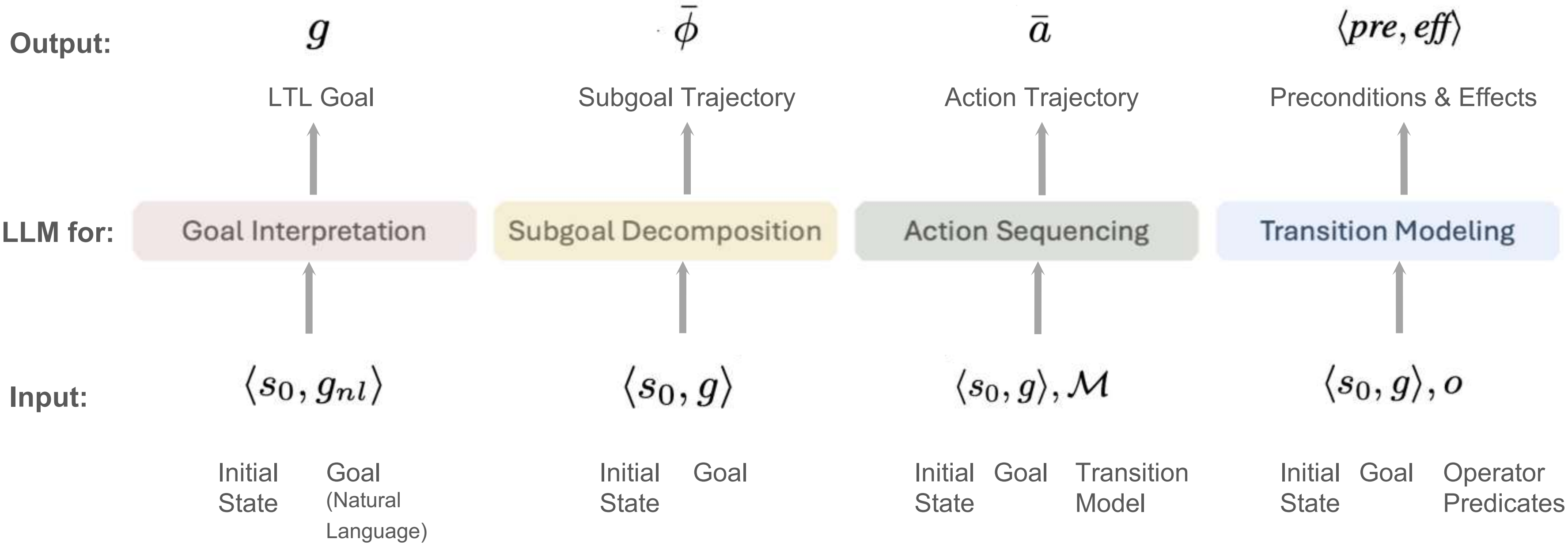
VirtualHome | BEHAVIOR | …

Embodied Decision Making

Fine-grained evaluation → Four modules

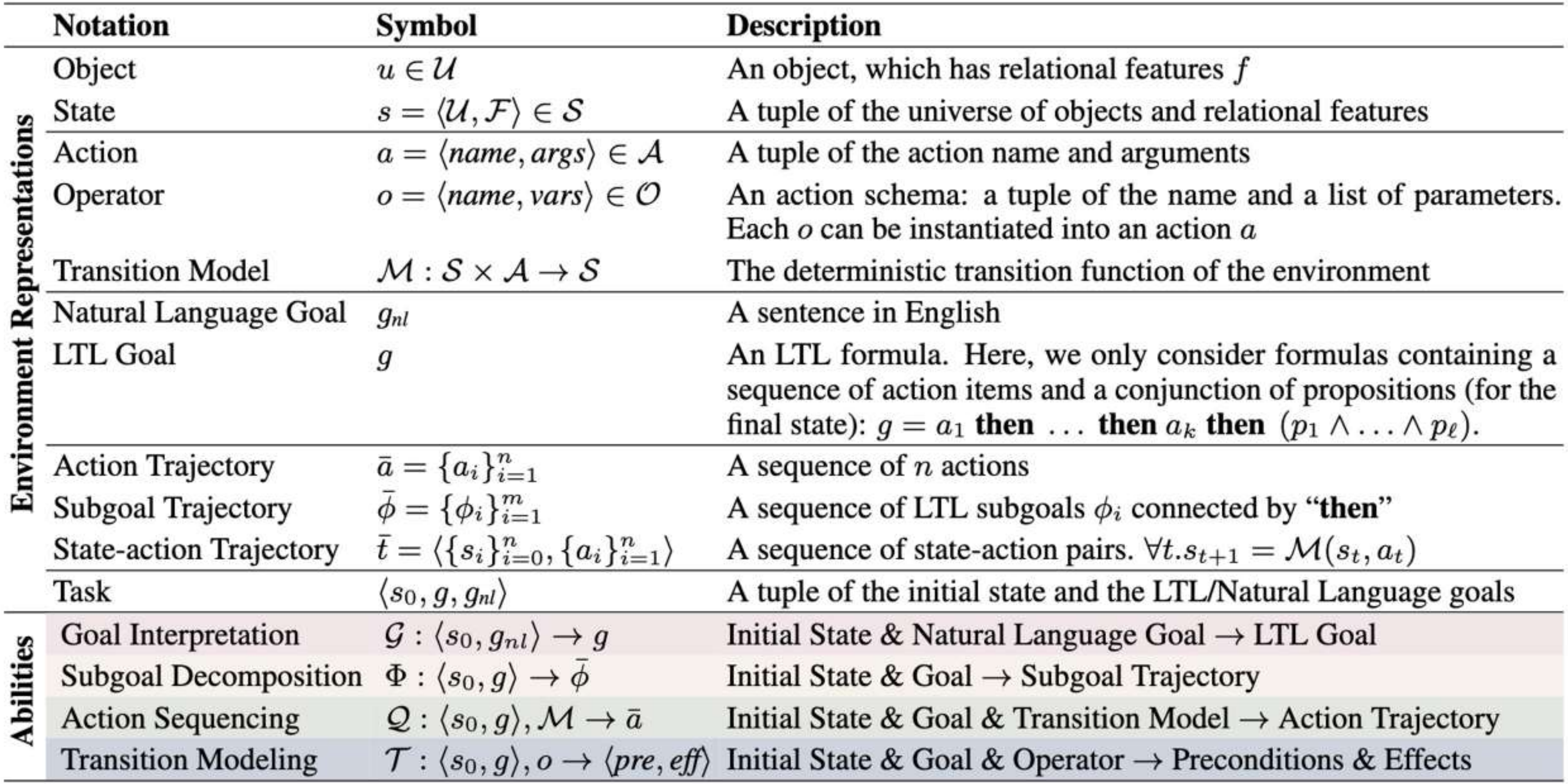


Standardization of modules and interfaces



Output:	g	$\bar{\phi}$	$\bar{a}$	$\langle pre, \mathit{eff} \rangle$
	LTL Goal	Subgoal Trajectory	Action Trajectory	Preconditions & Effects
LLM for:	Goal Interpretation	Subgoal Decomposition	Action Sequencing	Transition Modeling
Input:	$\langle s_0, g_{nl} \rangle$	$\langle s_0, g \rangle$	$\langle s_0, g \rangle, \mathcal{M}$	$\langle s_0, g \rangle, o$
	Initial State Goal (Natural Language)	Initial State Goal	Initial State Goal Transition Model	Initial State Goal Operator Predicates

Embodied Agent Interface

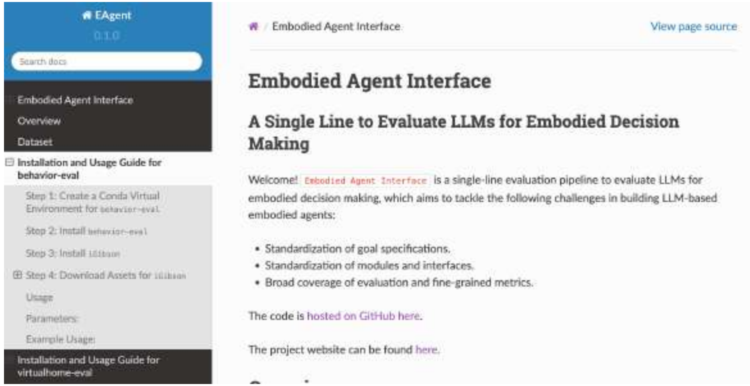


	Notation	Symbol	Description
Environment Representations	Object	$u \in \mathcal{U}$	An object, which has relational features f
	State	$s = \langle \mathcal{U}, \mathcal{F} \rangle \in \mathcal{S}$	A tuple of the universe of objects and relational features
	Action	$a = \langle name, args \rangle \in \mathcal{A}$	A tuple of the action name and arguments
	Operator	$o = \langle name, vars \rangle \in \mathcal{O}$	An action schema: a tuple of the name and a list of parameters. Each o can be instantiated into an action a
	Transition Model	$\mathcal{M} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$	The deterministic transition function of the environment
	Natural Language Goal	g_{nl}	A sentence in English
	LTL Goal	g	An LTL formula. Here, we only consider formulas containing a sequence of action items and a conjunction of propositions (for the final state): $g = a_1$ **then** $\ldots$ **then** a_k **then** $(p_1 \wedge \ldots \wedge p_\ell)$.
	Action Trajectory	$\bar{a} = \{a_i\}_{i=1}^{n}$	A sequence of n actions
	Subgoal Trajectory	$\bar{\phi} = \{\phi_i\}_{i=1}^{m}$	A sequence of LTL subgoals ϕ_i connected by "**then**"
	State-action Trajectory	$\bar{t} = \langle \{s_i\}_{i=0}^{n}, \{a_i\}_{i=1}^{n} \rangle$	A sequence of state-action pairs. $\forall t . s_{t+1} = \mathcal{M}(s_t, a_t)$
	Task	$\langle s_0, g, g_{nl} \rangle$	A tuple of the initial state and the LTL/Natural Language goals
Abilities	Goal Interpretation	$\mathcal{G} : \langle s_0, g_{nl} \rangle \to g$	Initial State & Natural Language Goal → LTL Goal
	Subgoal Decomposition	$\Phi : \langle s_0, g \rangle \to \bar{\phi}$	Initial State & Goal → Subgoal Trajectory
	Action Sequencing	$\mathcal{Q} : \langle s_0, g \rangle, \mathcal{M} \to \bar{a}$	Initial State & Goal & Transition Model → Action Trajectory
	Transition Modeling	$\mathcal{T} : \langle s_0, g \rangle, o \to \langle pre, eff \rangle$	Initial State & Goal & Operator → Preconditions & Effects

A single line

to evaluate LLMs for embodied agents!

(get rid of simulator installation, execution, debugging, etc)


EAgent
0.1.0
Search docs
Embodied Agent Interface
Overview
Dataset
Installation and Usage Guide for behavior-eval
Step 1: Create a Conda Virtual Environment for behavior-eval
Step 2: Install behavior-eval
Step 3: Install iGibson
Step 4: Download Assets for iGibson
Usage
Parameters:
Example Usage:
Installation and Usage Guide for virtualhome-eval
Embodied Agent Interface
View page source

Embodied Agent Interface

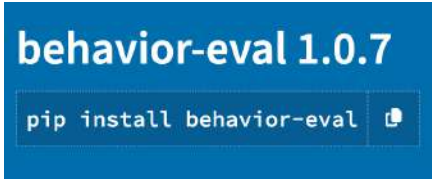
A Single Line to Evaluate LLMs for Embodied Decision Making

Welcome! `Embodied Agent Interface` is a single-line evaluation pipeline to evaluate LLMs for embodied decision making, which aims to tackle the following challenges in building LLM-based embodied agents:

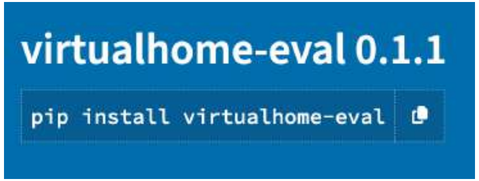
- Standardization of goal specifications.
- Standardization of modules and interfaces.
- Broad coverage of evaluation and fine-grained metrics.

The code is hosted on GitHub here.

The project website can be found here.


behavior-eval 1.0.7

```
pip install behavior-eval
```

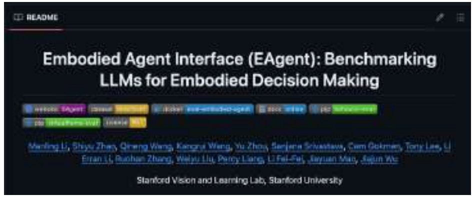

virtualhome-eval 0.1.1

```
pip install virtualhome-eval
```

jameskrw/eval-embodied-agent

By jameskrw · Updated 10 days ago

The docker file for eval-embodied-agent project

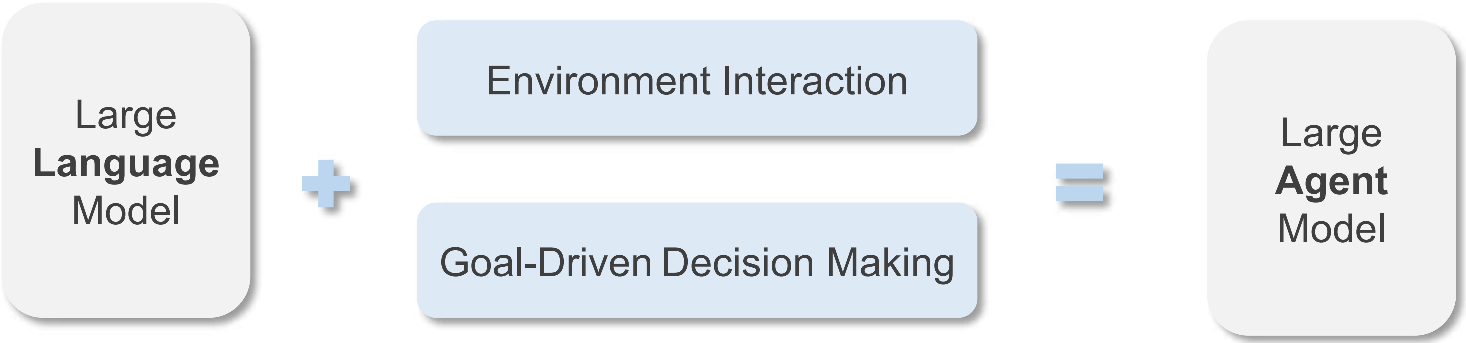


Embodied Agent Interface (EAgent): Benchmarking LLMs for Embodied Decision Making

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, Jiajun Wu

Stanford Vision and Learning Lab, Stanford University

Large **Language** Models → Large **Agent** Models



Large **Language** Model

+

Environment Interaction

Goal-Driven Decision Making

=

Large **Agent** Model

Physical World Knowledge

Long-Horizon

Go to Longer Horizon Reasoning!

Geometric

Go to Lower Level!

Please feel free to drop an email if you would like to talk!
manling.li@northwestern.edu