

Multimodal Reasoning Intelligence

How Multimodal Models Learn (and Fail to Learn) the Physical World

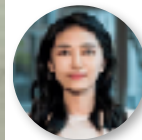
Manling Li
Northwestern University
Amazon Scholar
manling.li@northwestern.edu



NORTHWESTERN
UNIVERSITY



MLL LAB
Machine Learning and Language

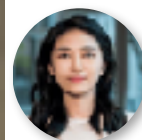


Set up the table.

...

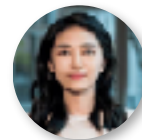
on_top_of (turkey, table)

...



Use the plates.

te

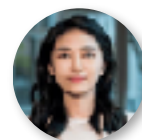


Set up the table.

...

on_top_of (turkey, table)

...



Use the plates.

...

on_top_of (turkey, plate)

on_top_of (plate, table)

...



Li, et al. Embodied Agent Interface
(**Oral** NeurIPS'24; **Best Paper** SoCal NLP'25)

Action Trajectory

...

on_top_of (turkey, plate)

on_top_of (plate, table)

...



Reasoning in the Physical World:

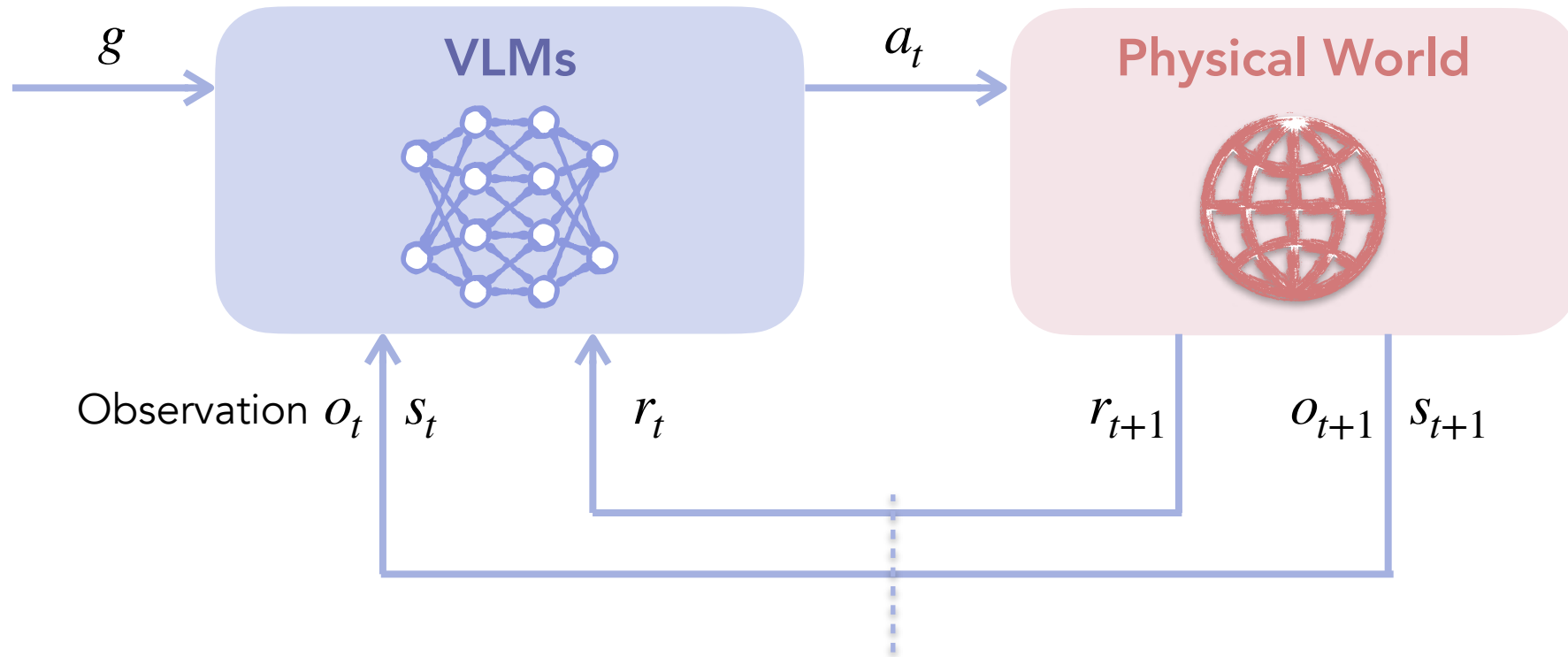
Environment **State**

Action Trajectory

Reasoning in the Physical World:

Markov Decision Process as a Principle

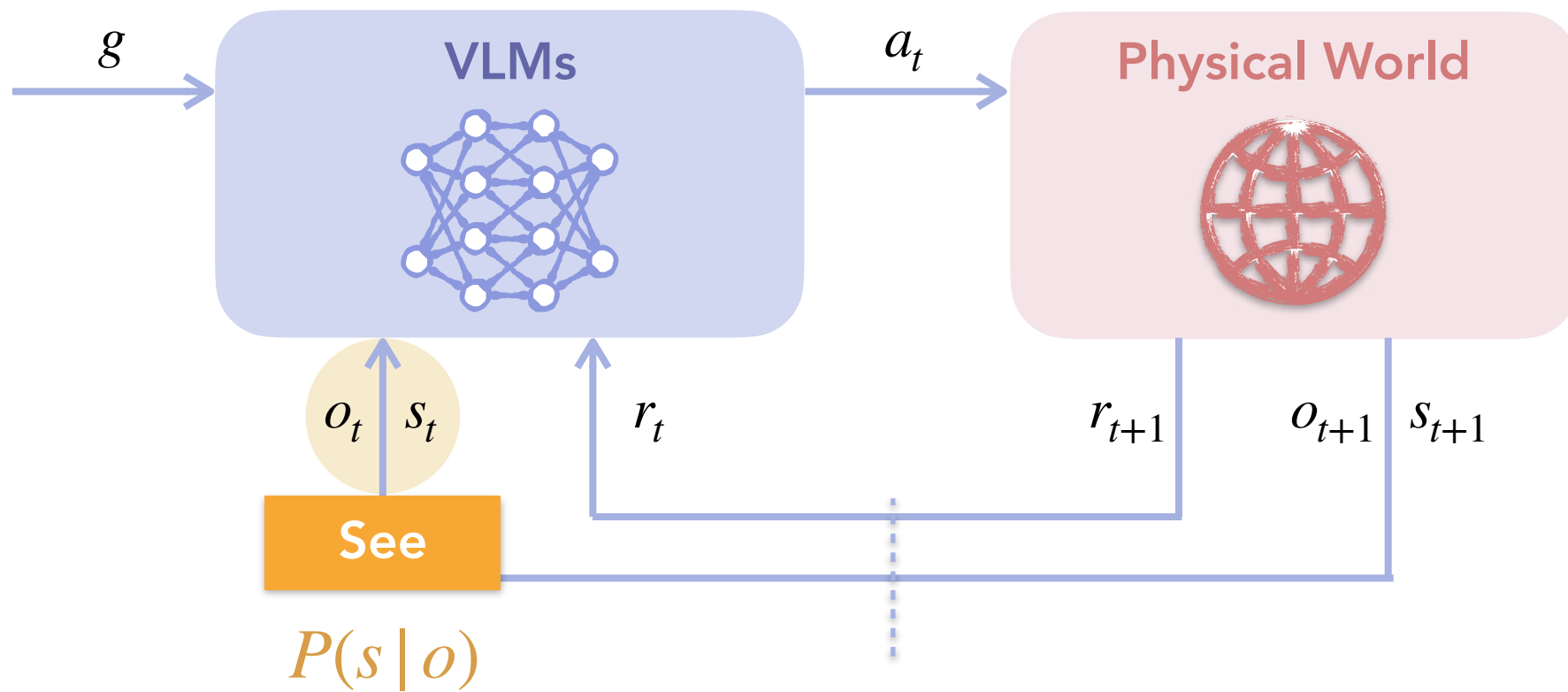
Markov Decision Process



See

Act

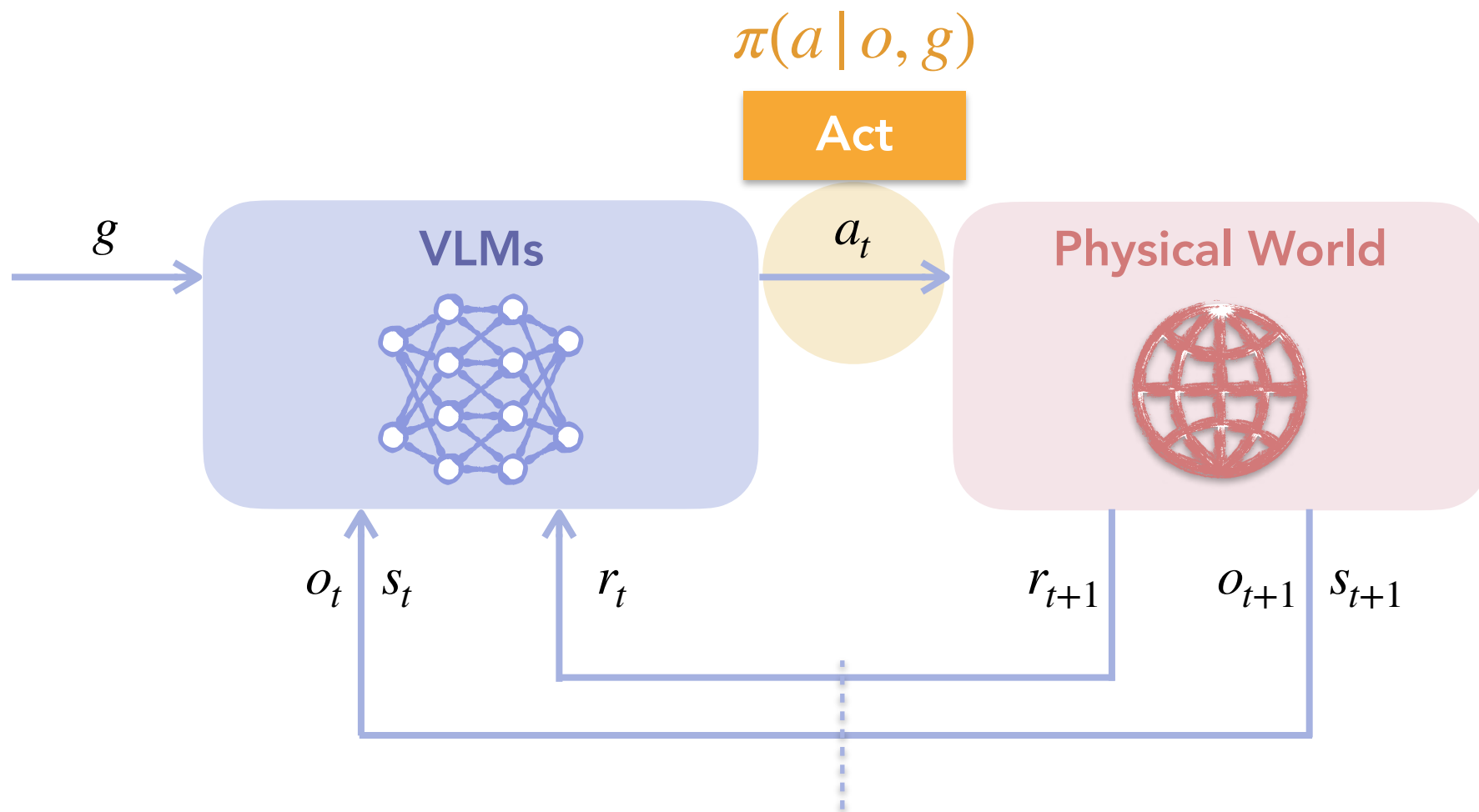
Control



See

Act

Control



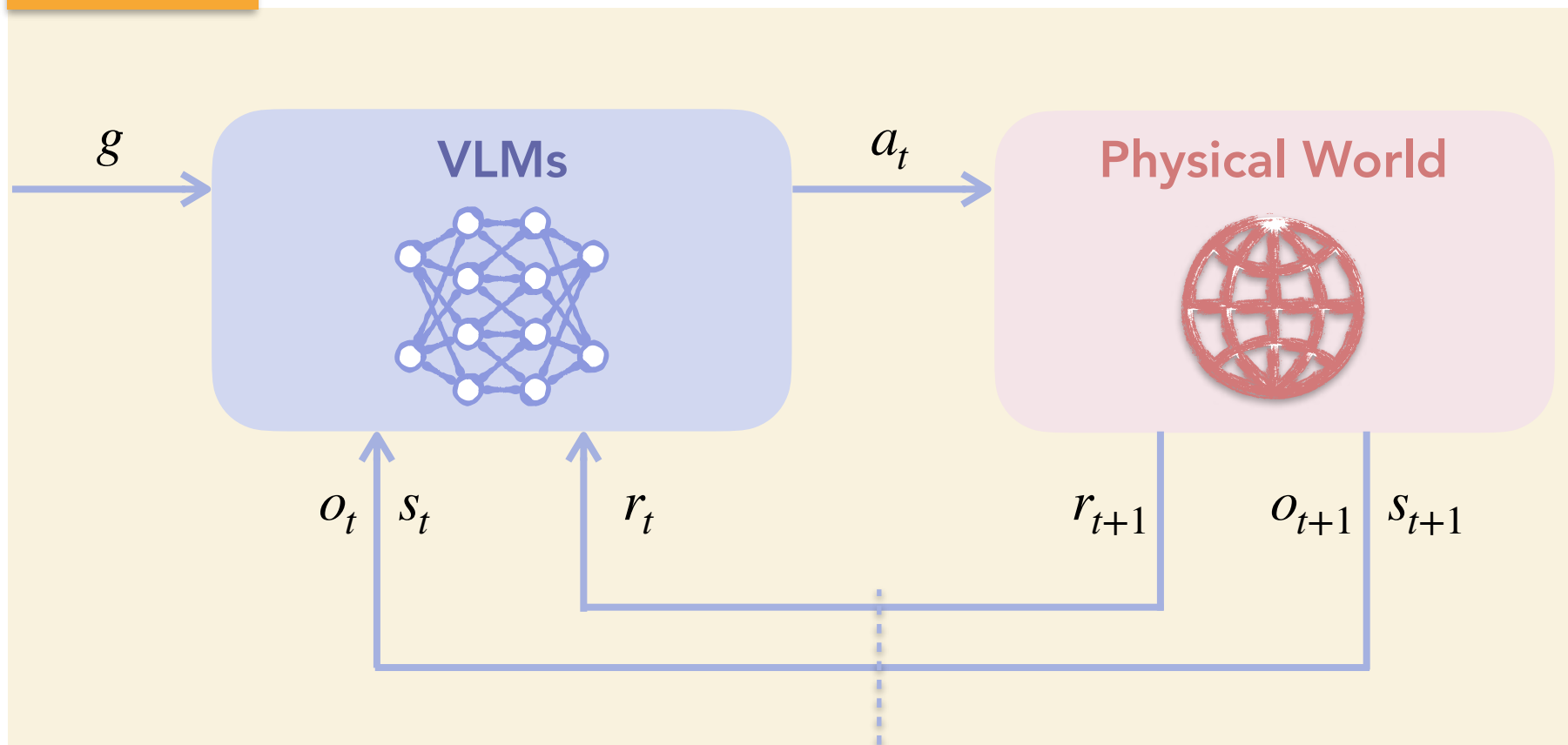
See

Act

Control

Control

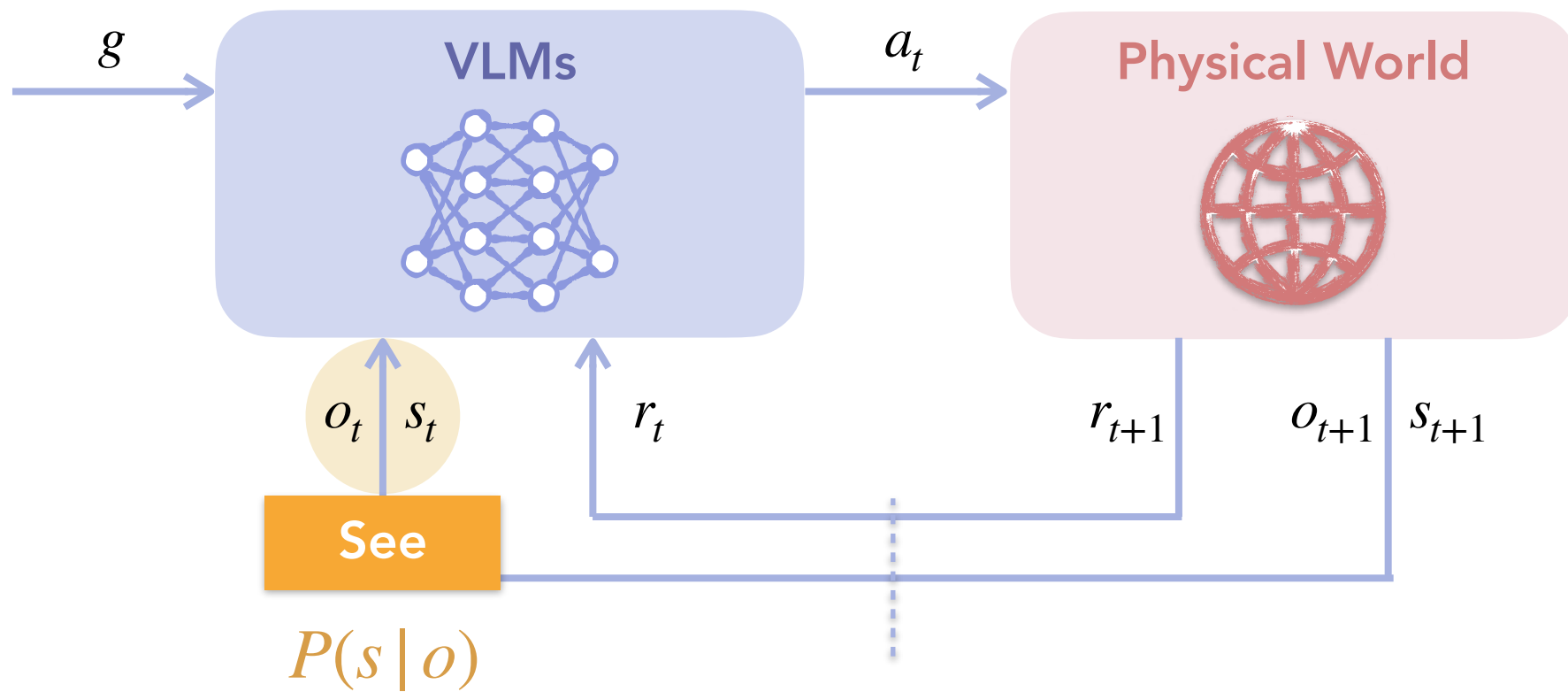
$Control(\pi)$



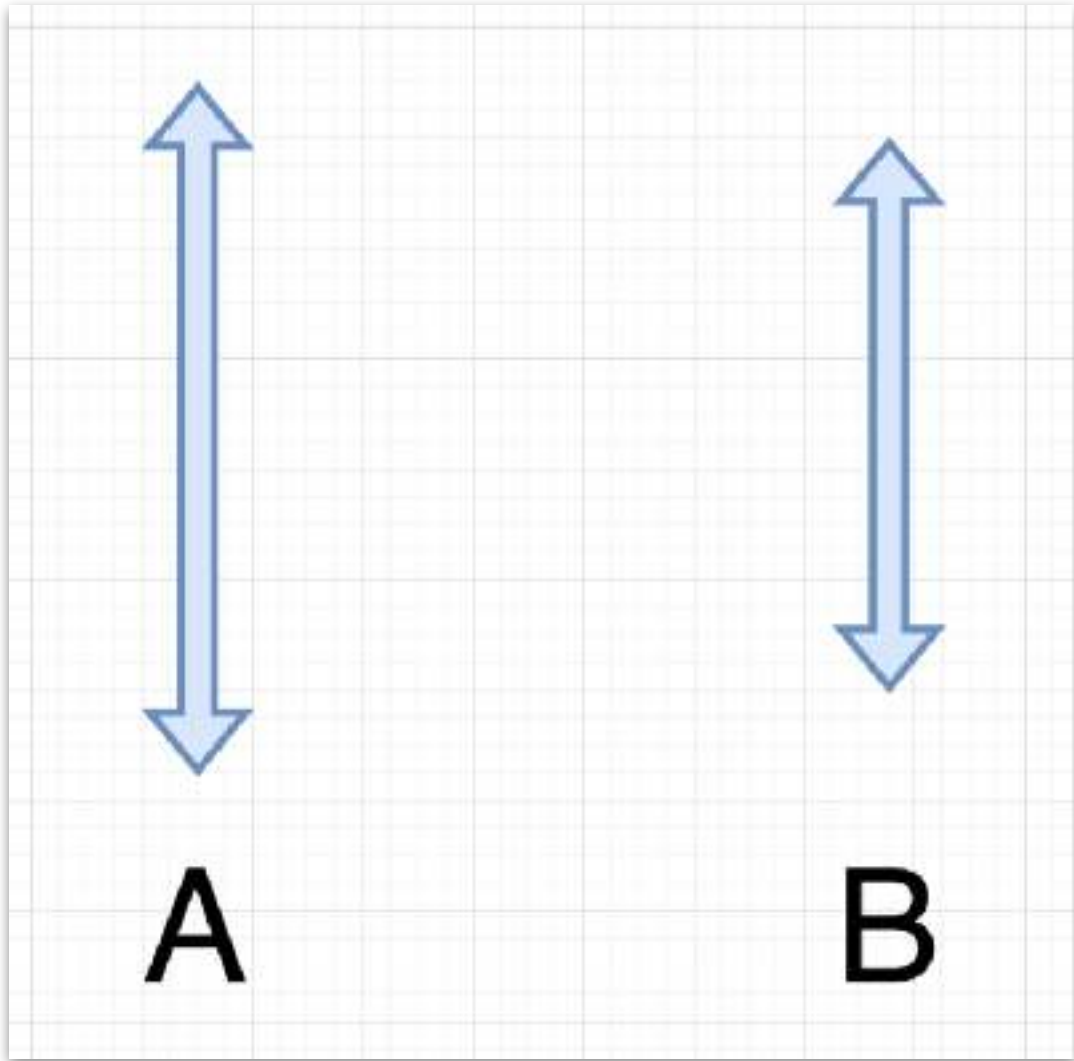
See

Act

Control



Quick Quiz



Which is longer?

Both arrows, A and B, are of the **same length.**

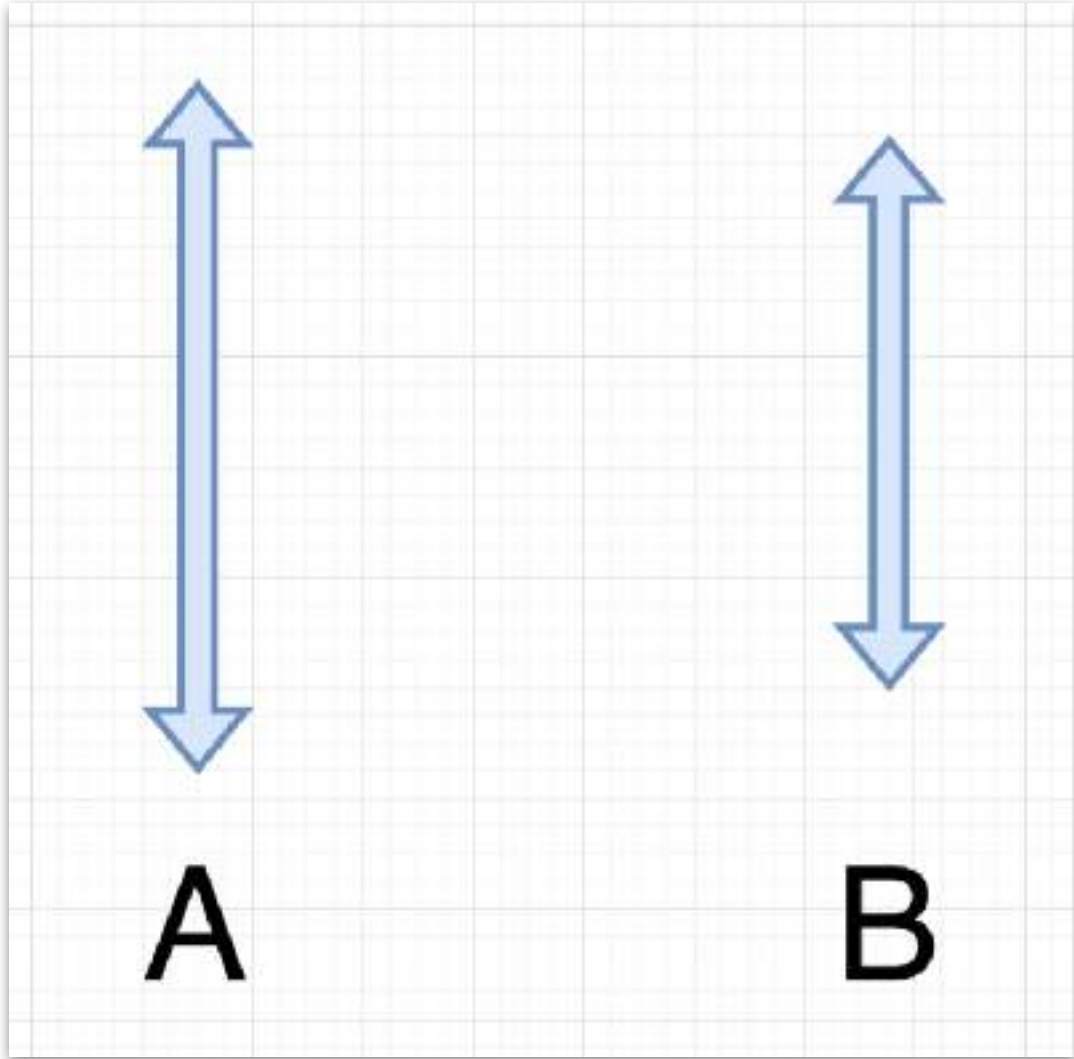
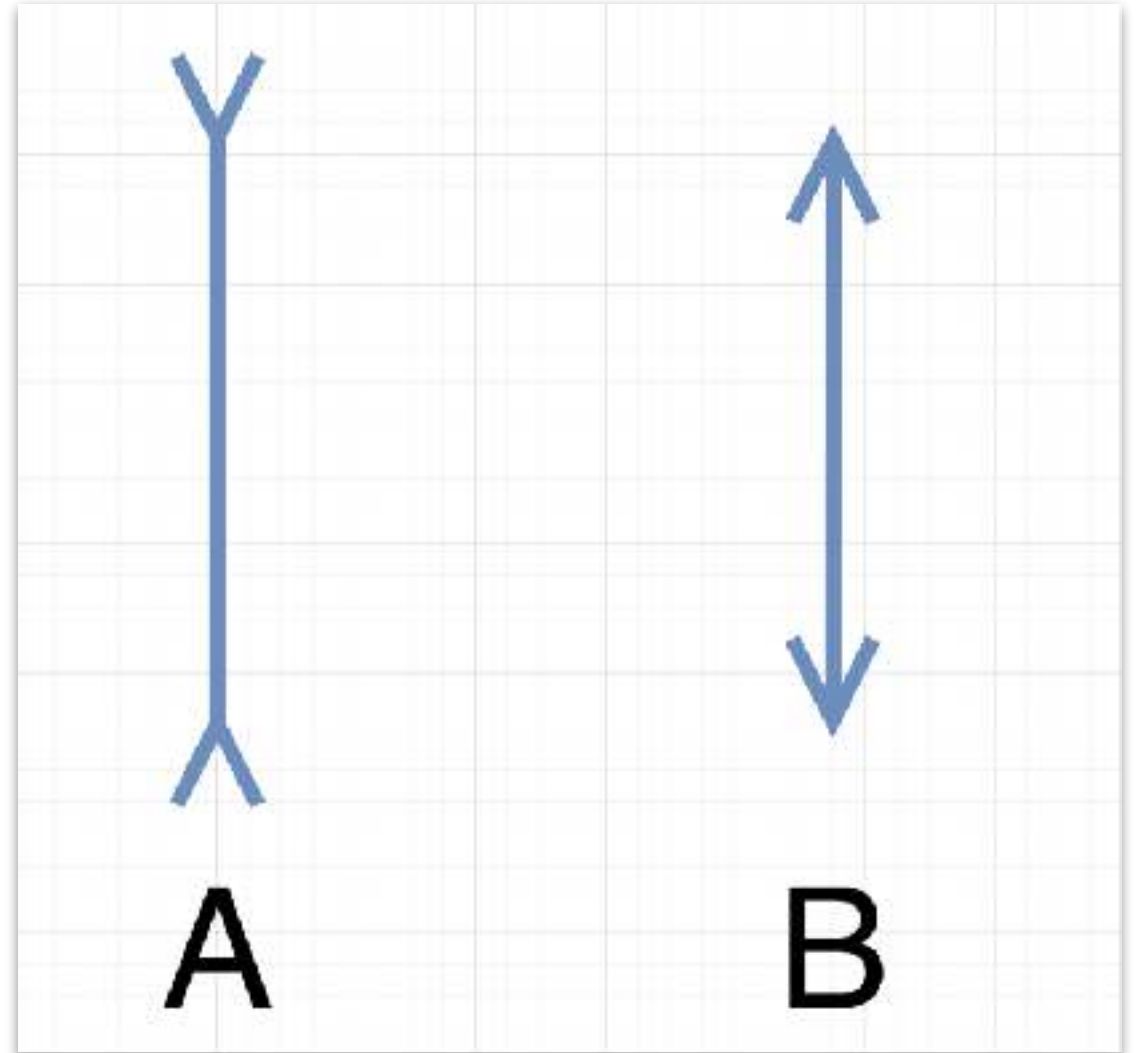


2023.09

GPT4-V



Memorization



Müller-Lyer illusion

Reasoning

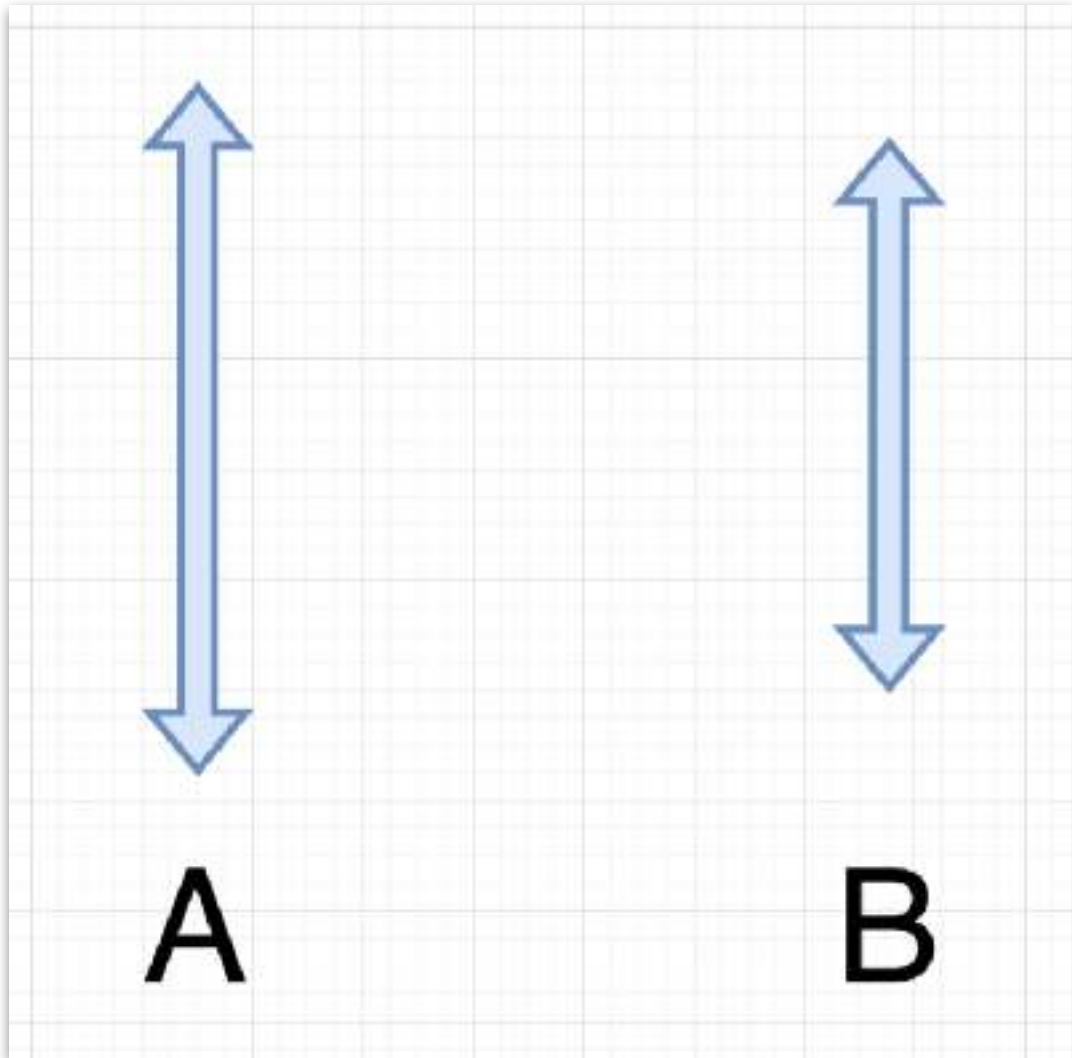
what we want

**?
==**

Memorization

what we truly learn

Reasoning $\neq$ Memorization



Which is longer?

2023.09

GPT4-V



Both arrows, A and B, are of the **same length**.



They are the **same length**.

This is a **classic visual illusion**. Because the arrows are placed in different positions on the page (A closer to the top, ...



2026.02

GPT-5.2



2019

BERT



GPT



T5



...

2023.09

GPT4-V



2026.02

GPT-5.2



Language Models (LMs)

aka. Large Language Models (LLMs), Foundation Models

$$P(y | x)$$



The cup is _____

on the table.

2019

BERT



GPT



T5



...

2023

GPT4-V



"Vision as another language"

ACL 2025 Inaugural Dissertation Award Honorable Mention 🏆

2026

GPT-5.2



2019

BERT



GPT



T5



...

2023

GPT4-V



"Vision as another language"

ACL 2025 Inaugural Dissertation Award Honorable Mention 🏆

2026

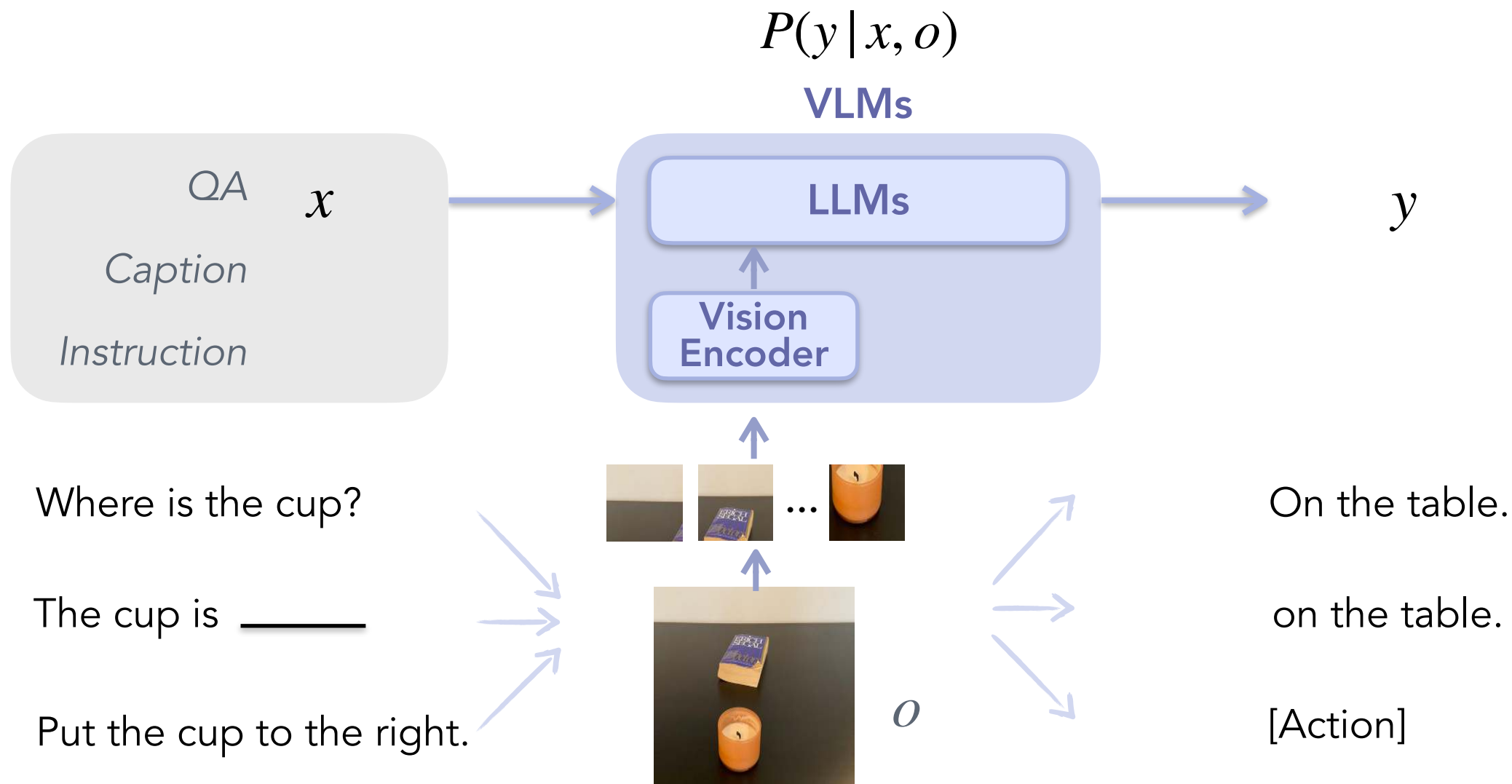
GPT-5.2



?

Vision-Language Models (VLMs)

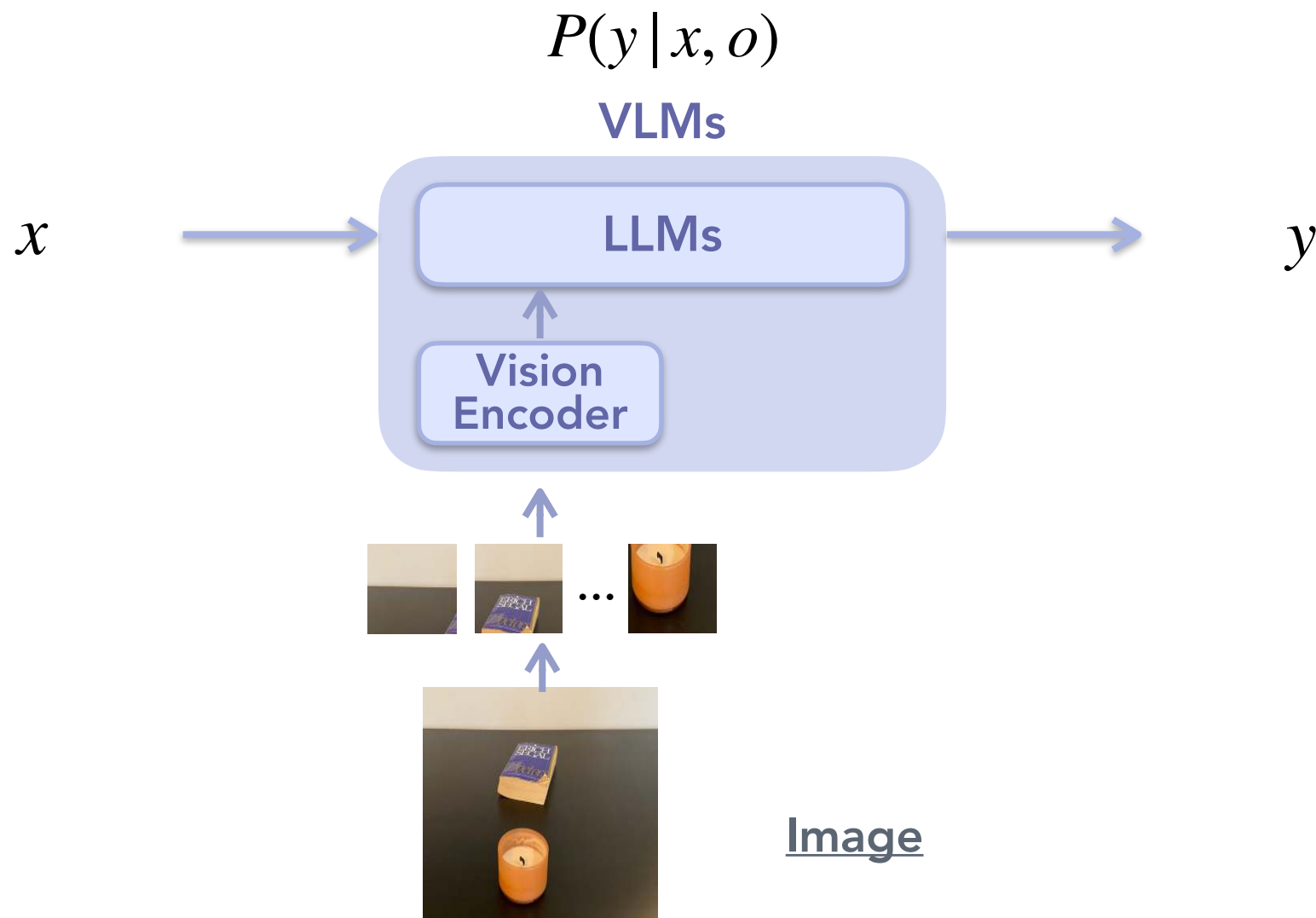
aka. Multimodal Large Language Models (MLLMs), Multimodal Foundation Models



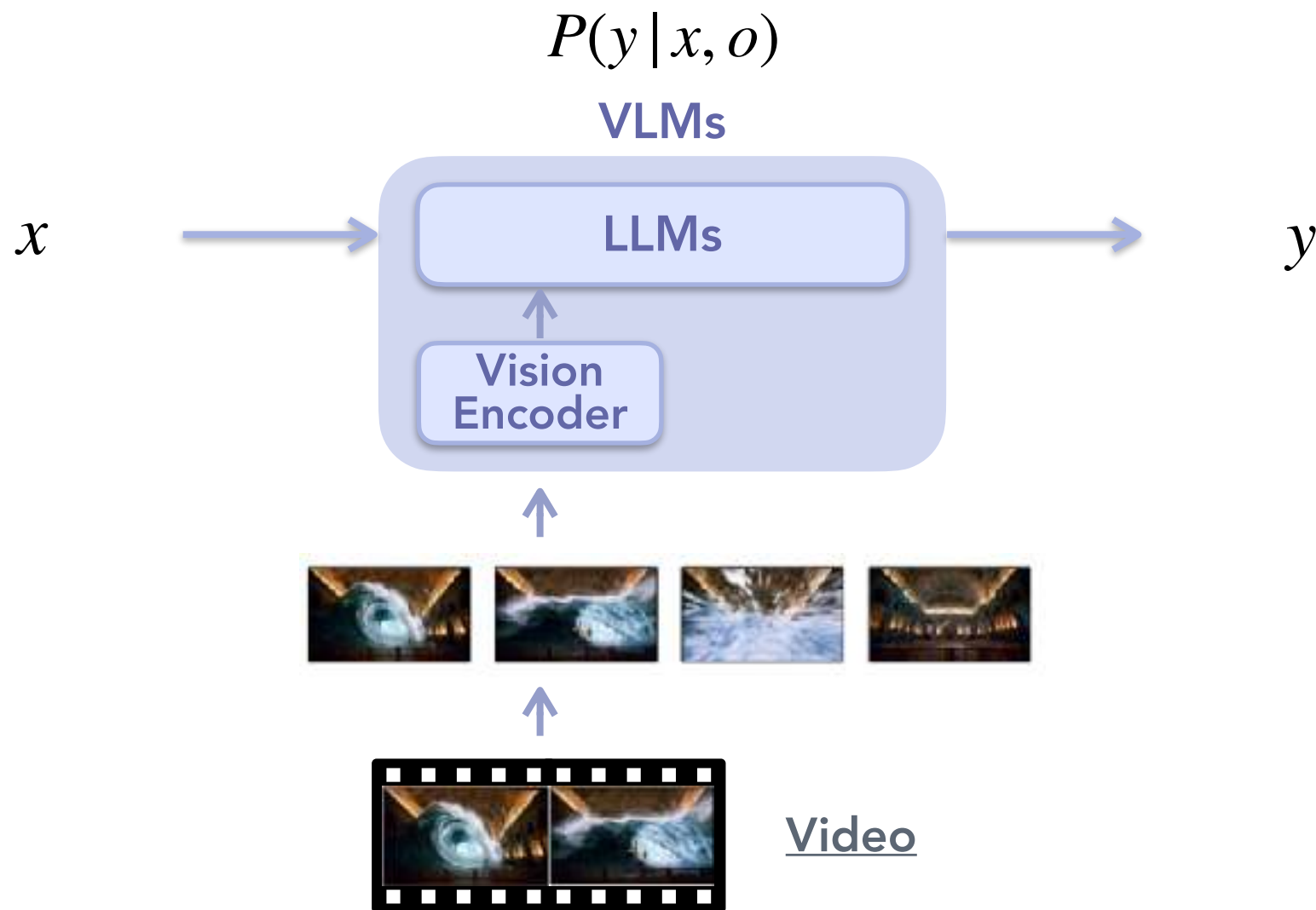
Vision as Another Language

Linearize Everything as Sequences

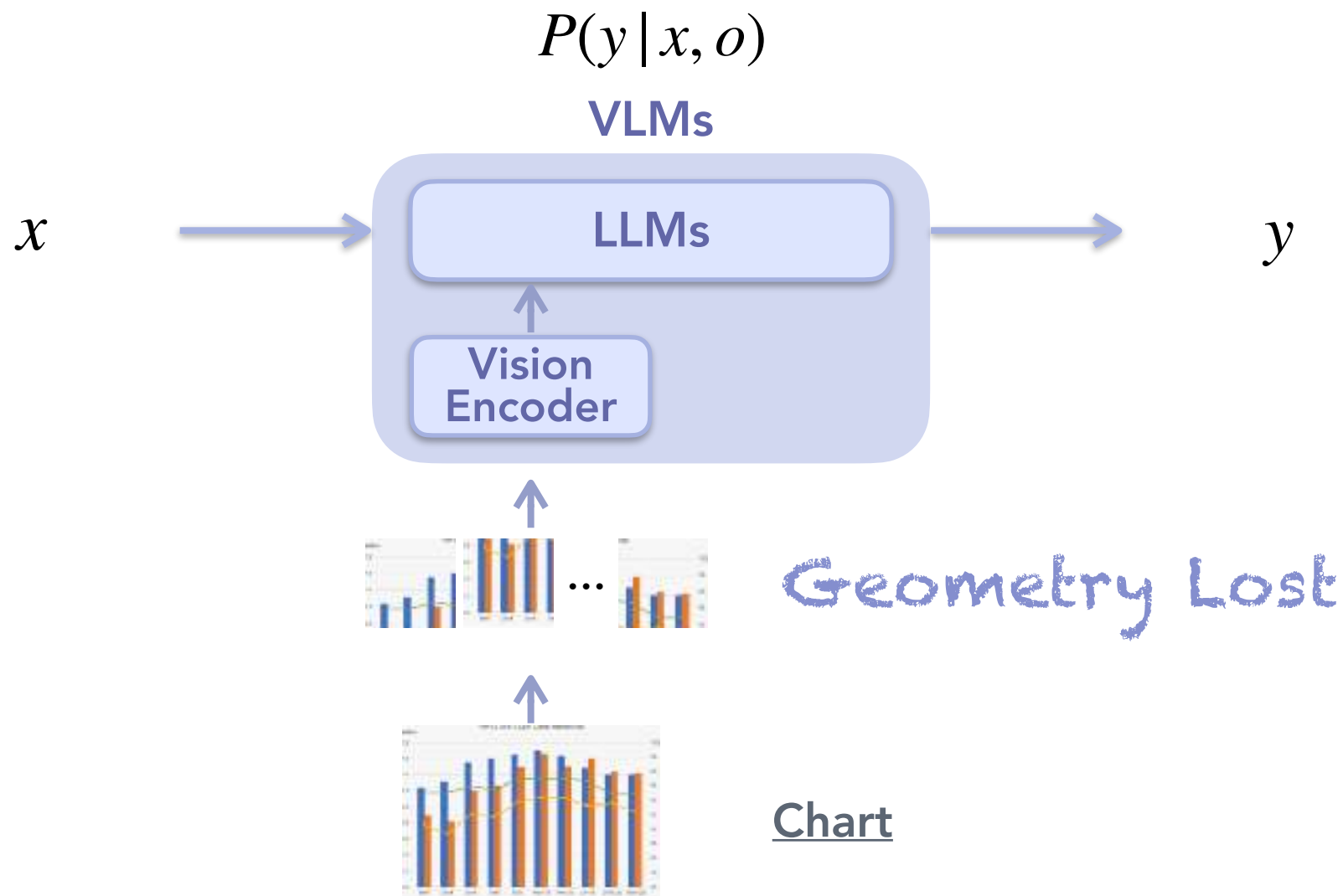
Model: Vision as Another Language



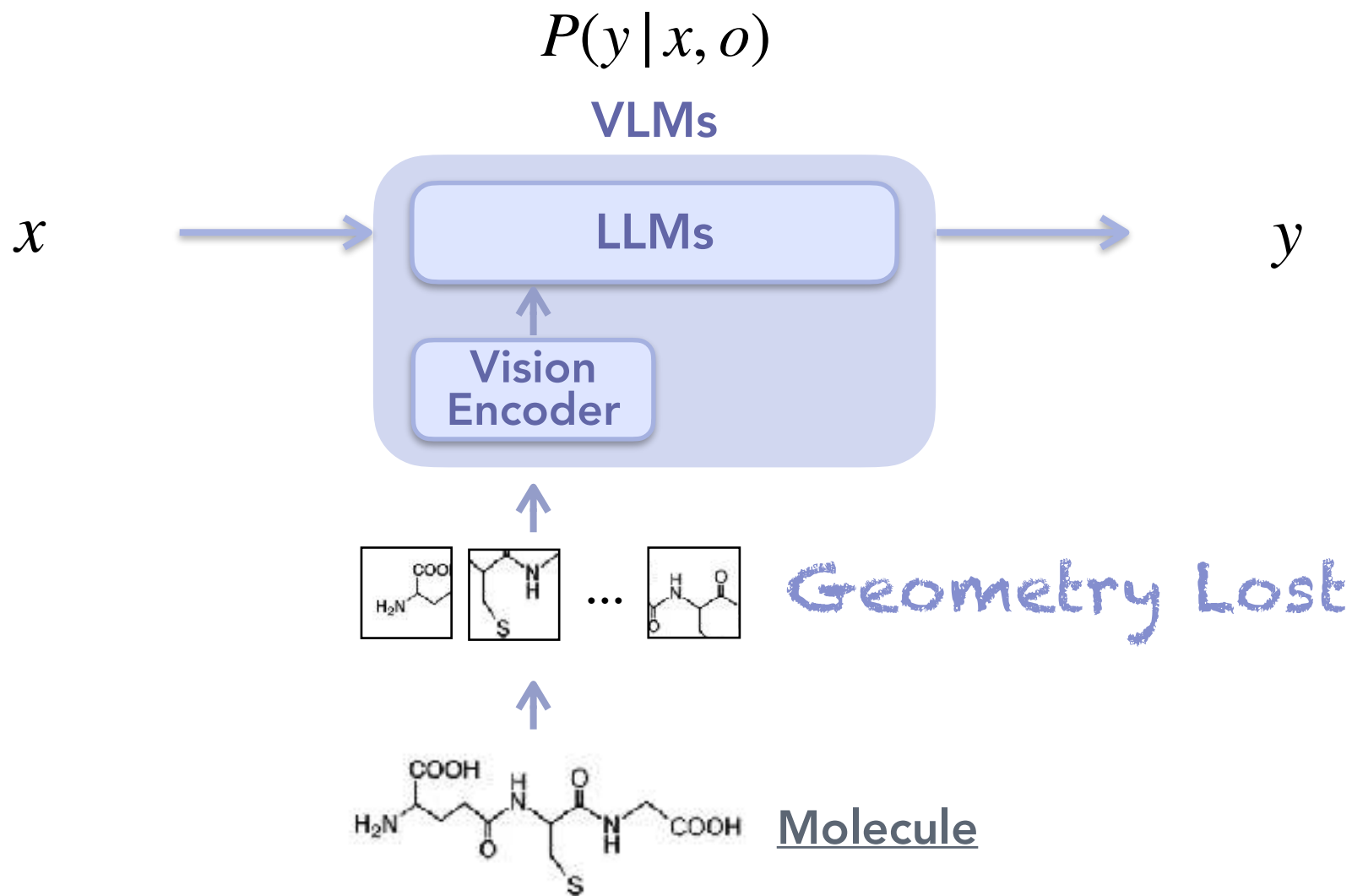
Model: Vision as Another Language



Model: Vision as Another Language



Model: Vision as Another Language



Memorization

what can be aligned

Semantics



Reasoning

what can NOT be aligned

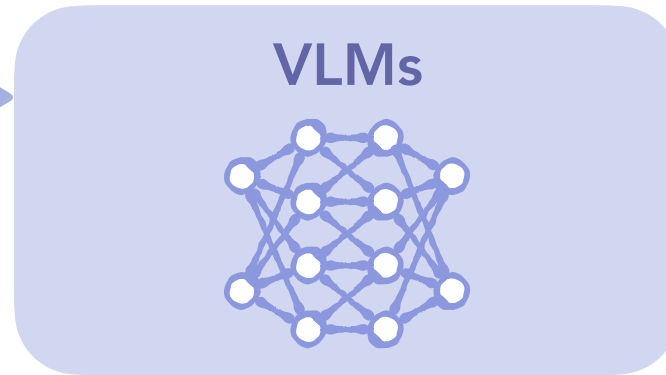
Spatial Geometry

See

Act

Control

x



y



o

Representation

Cross-modality Alignment

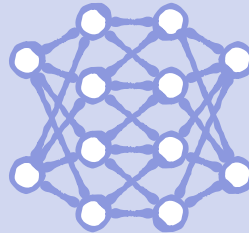
Act

Control

x



VLMs



y

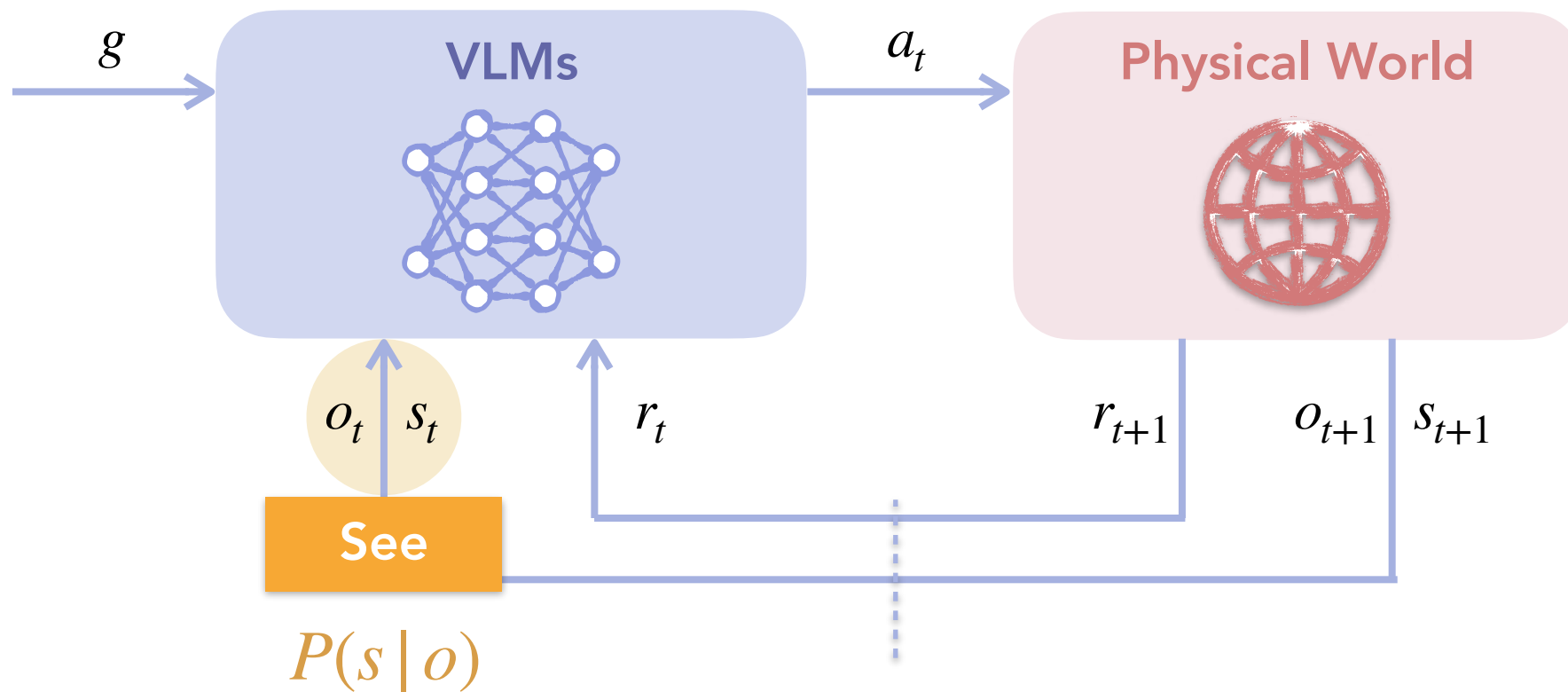


o

Representation
Cross-modality Alignment

Act

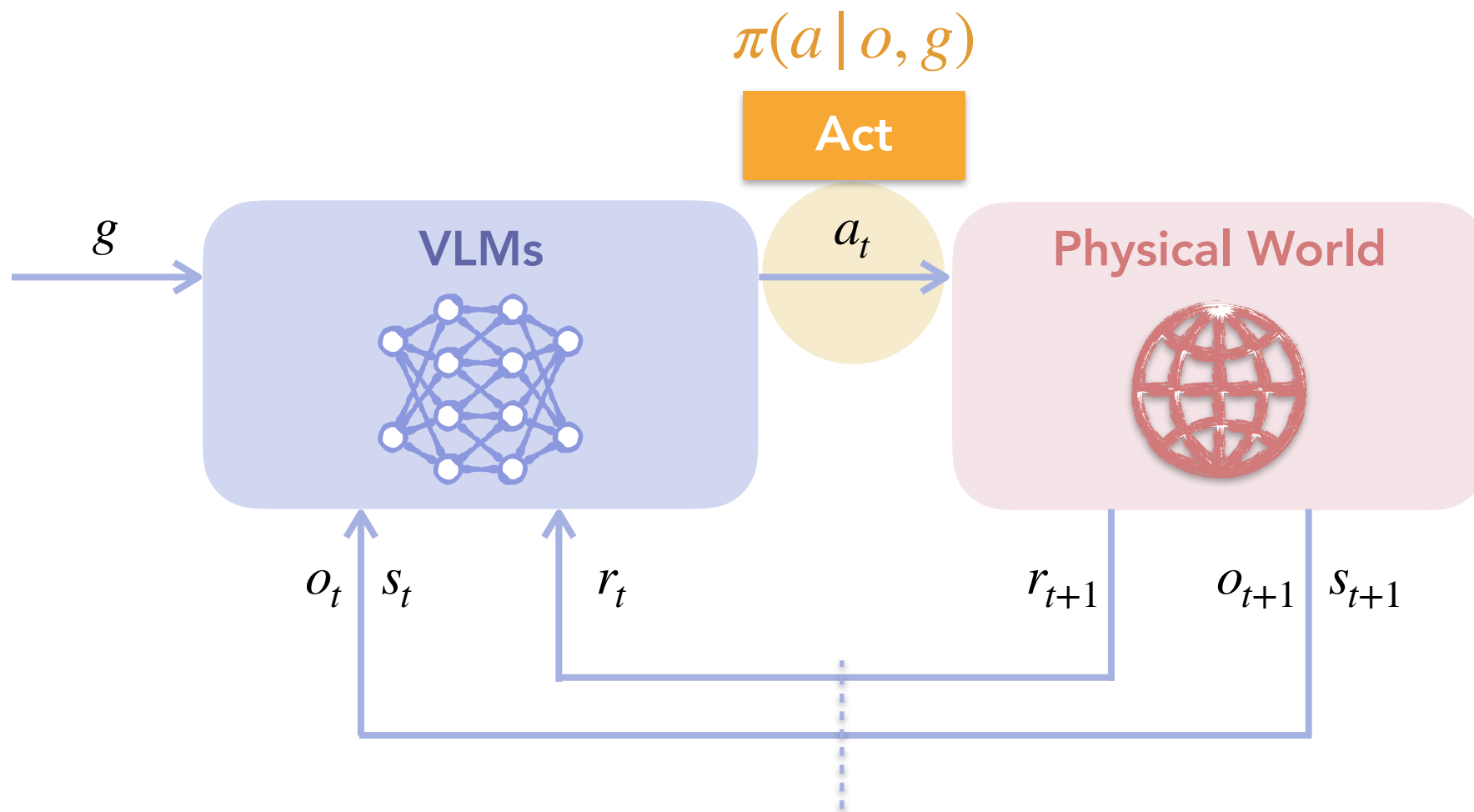
Control



Representation
Cross-modality Alignment

Act

Control





What if we take this action

on_top_of (turkey, table)



Turkey is greasy



It will make table greasy



Unwanted action

Reasoning

what if "action"

≠

Memorization

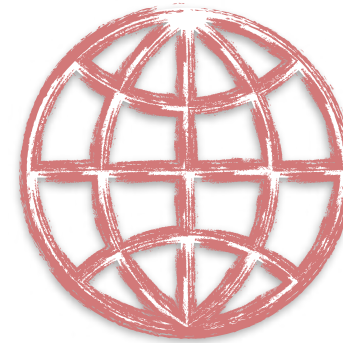
turkey $\longleftrightarrow$ table

Humans learn "actions" through **interactions**



Human

Interaction

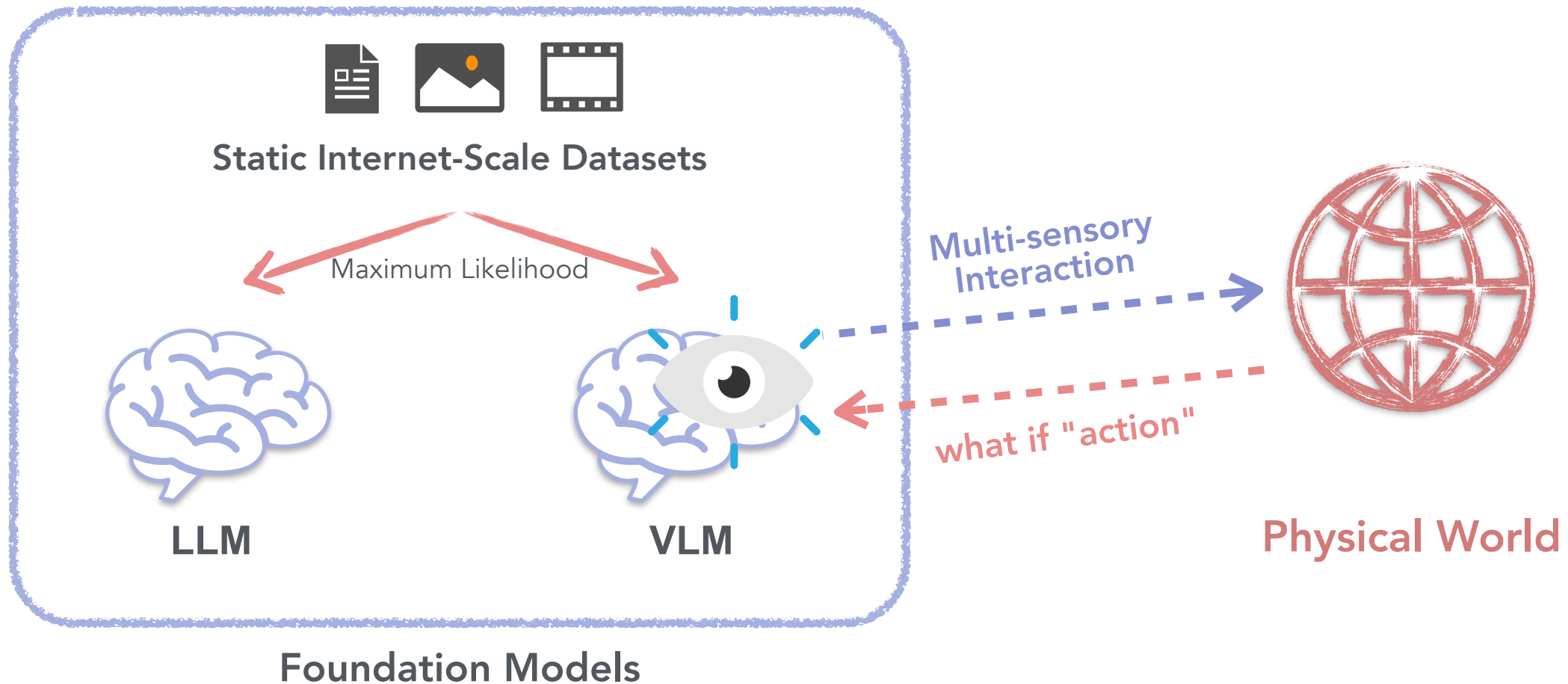


Physical World



what if "action"

Machines learn "actions" w/o interactions



Reasoning

Exploration



Memorization

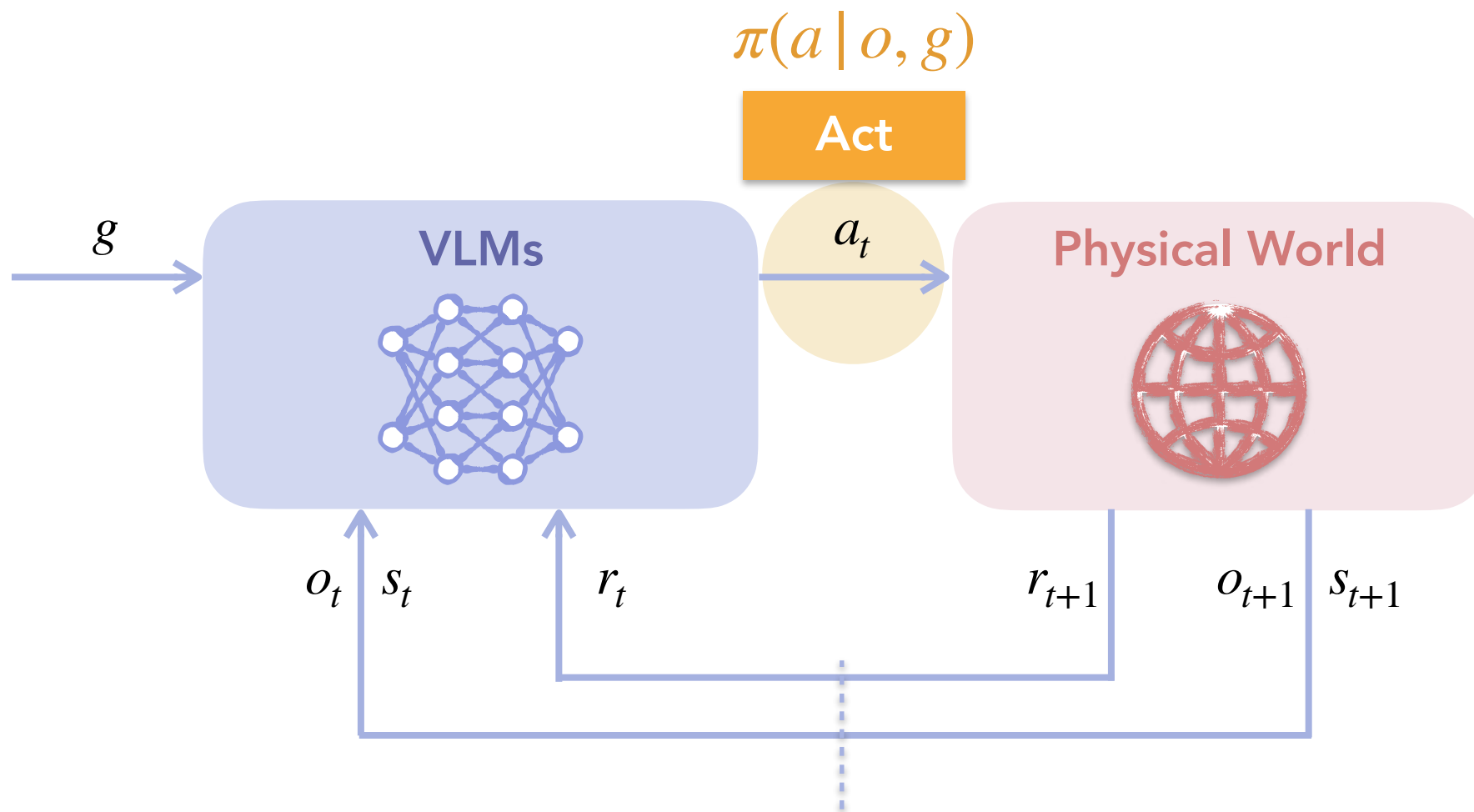
Exploitation

Active Explorer ← Passive Observer

Representation
Cross-modality Alignment

Act

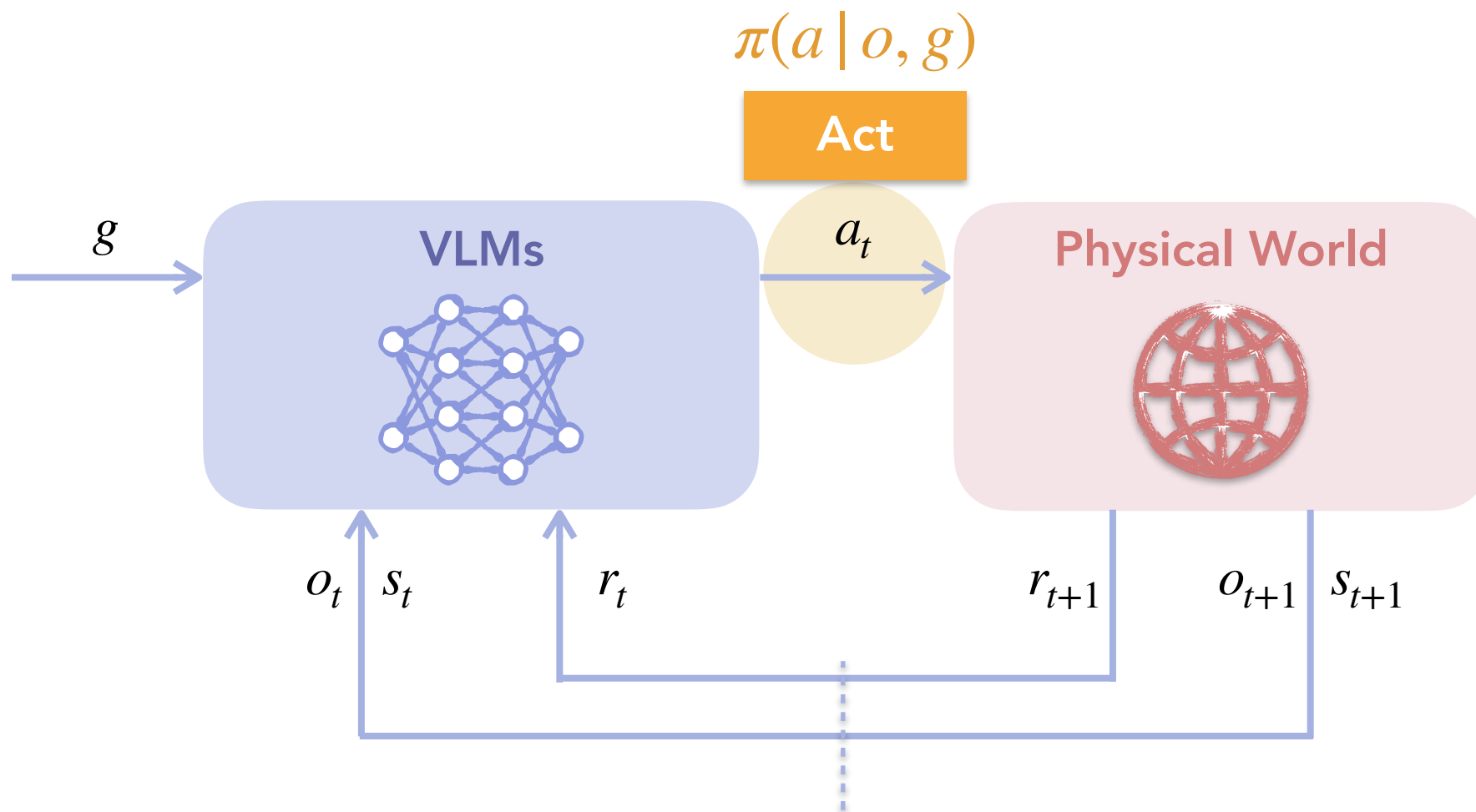
Control



Representation
Cross-modality Alignment

Learning Paradigm
Exploration over Exploitation

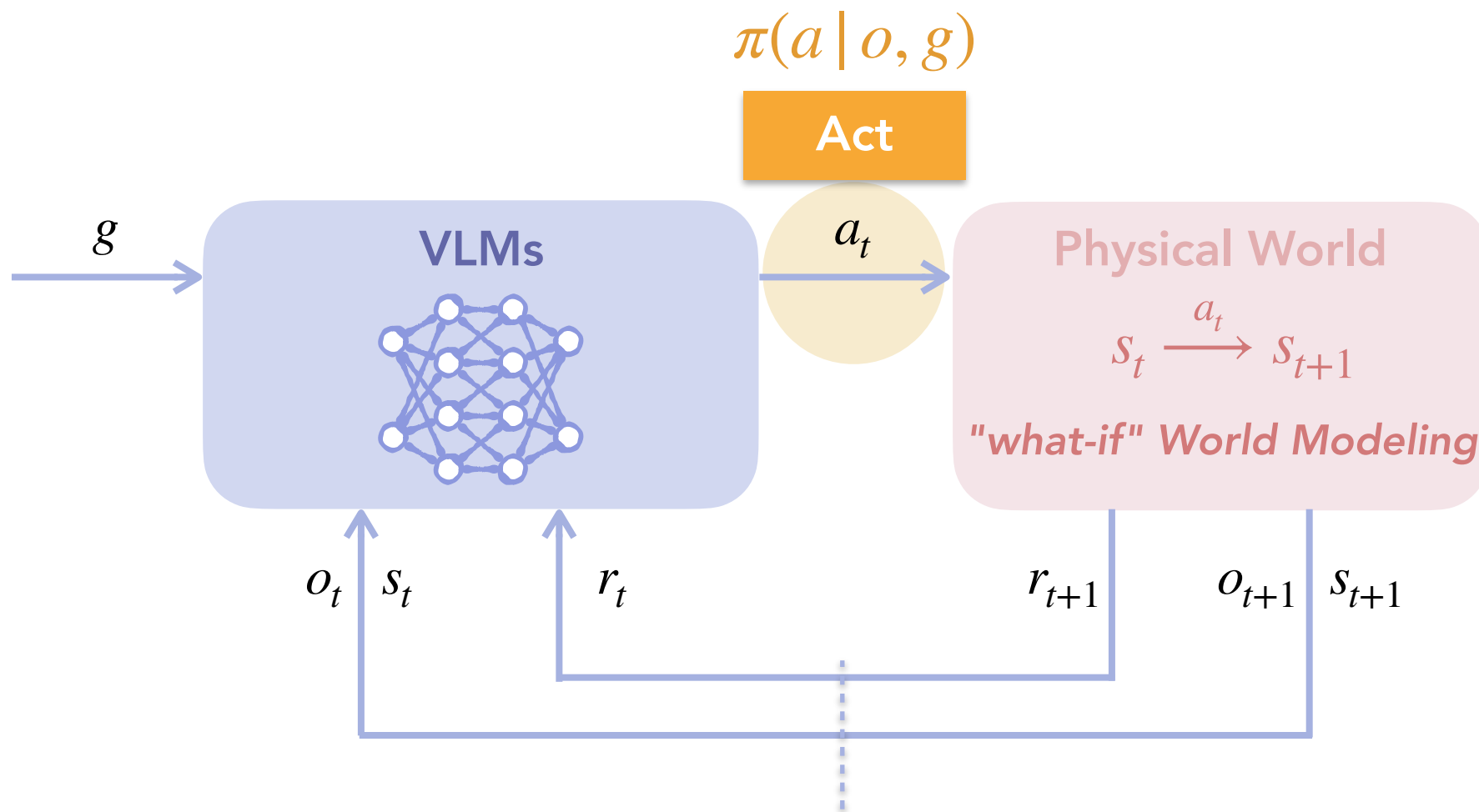
Control



Representation
Cross-modality Alignment

Learning Paradigm
Exploration over Exploitation

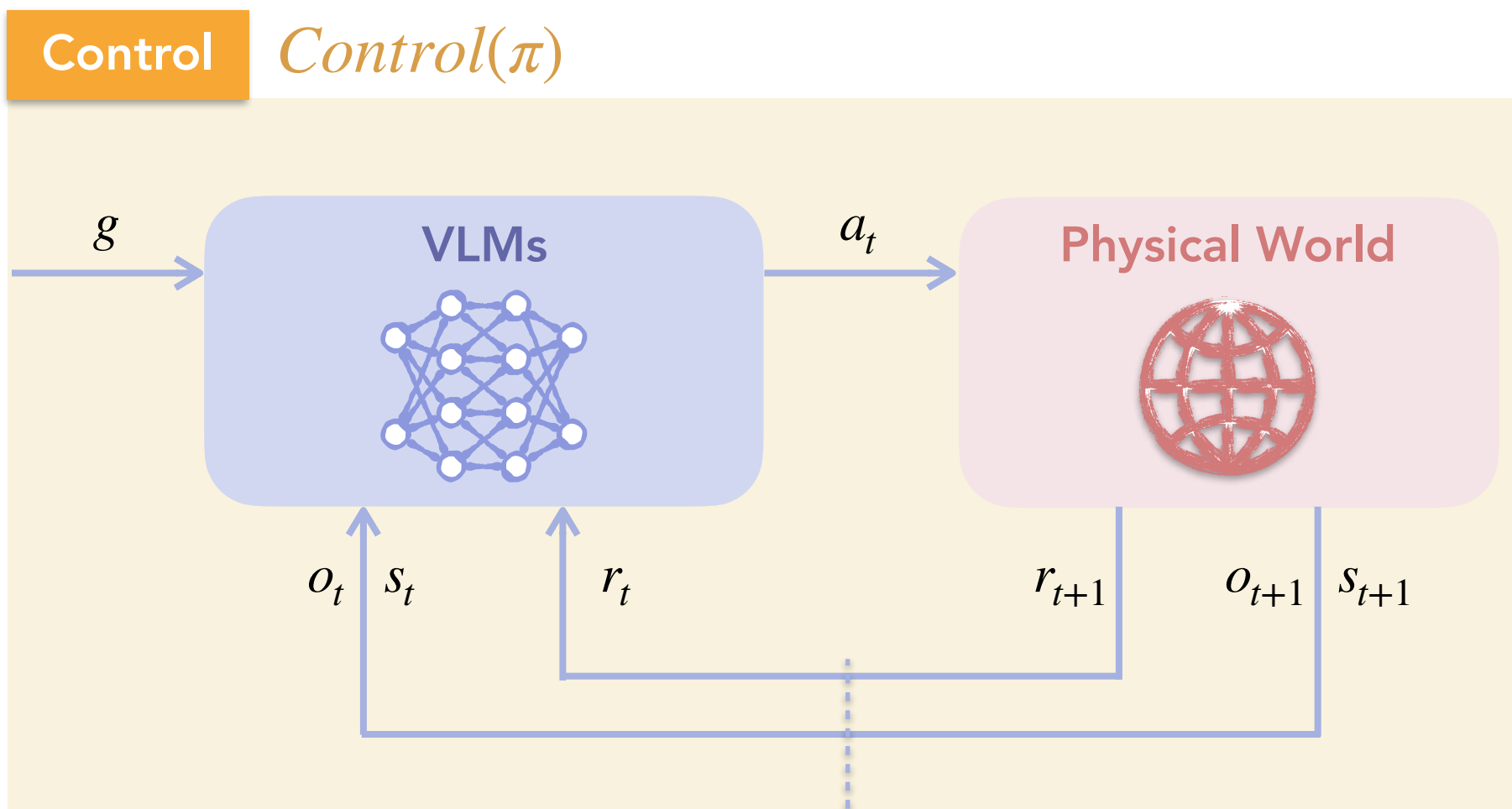
Control



Representation
Cross-modality Alignment

Learning Paradigm
Exploration over Exploitation

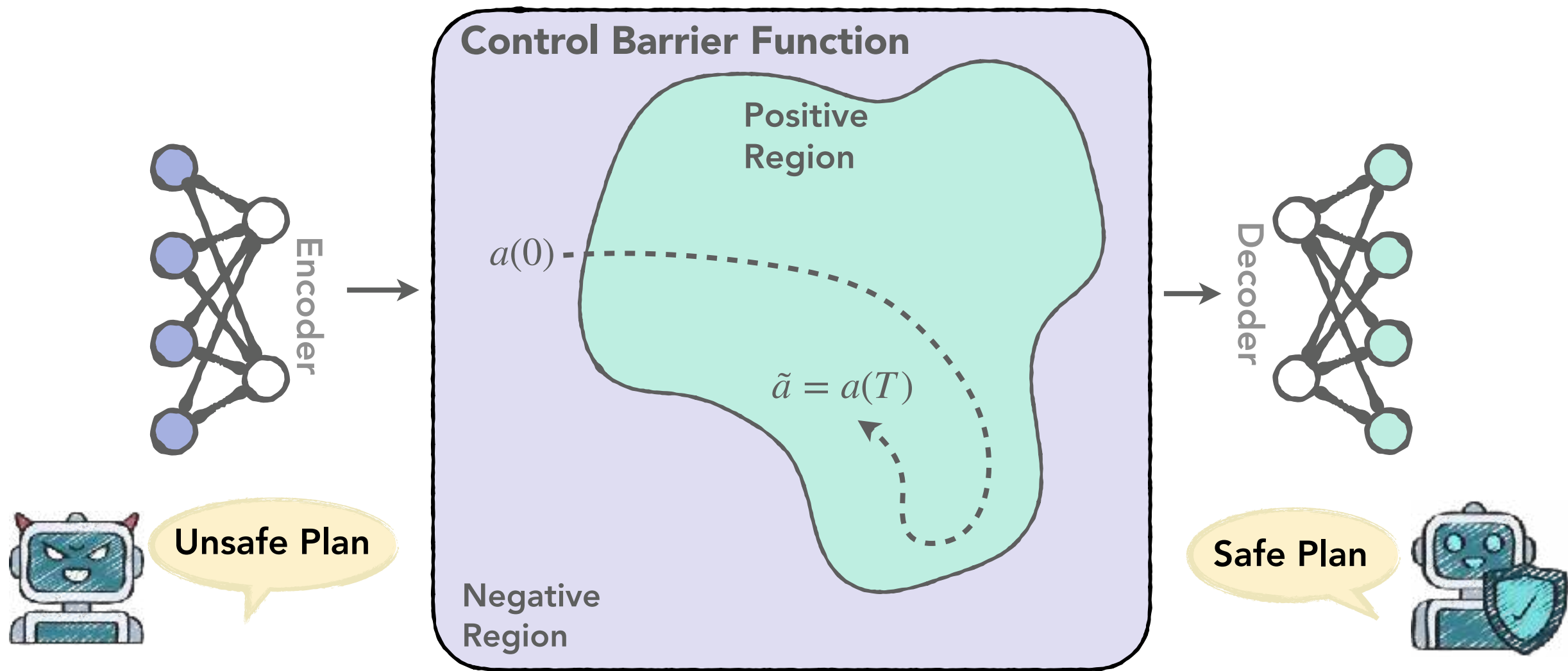
Control



Representation
Cross-modality Alignment

Learning Paradigm
Exploration over Exploitation

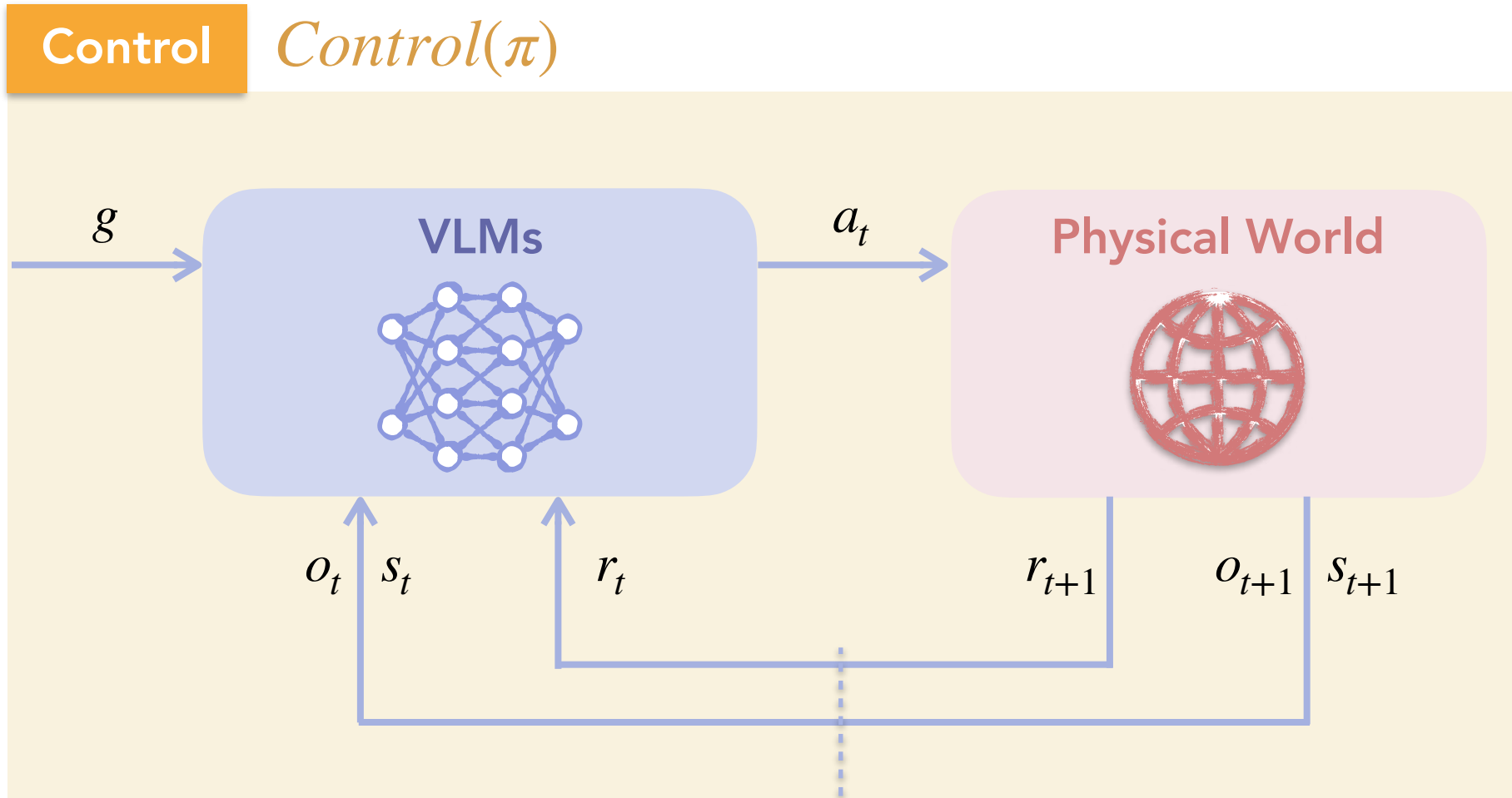
Control



Representation
Cross-modality Alignment

Learning Paradigm
Exploration over Exploitation

Control
Safety and Steering



Representation

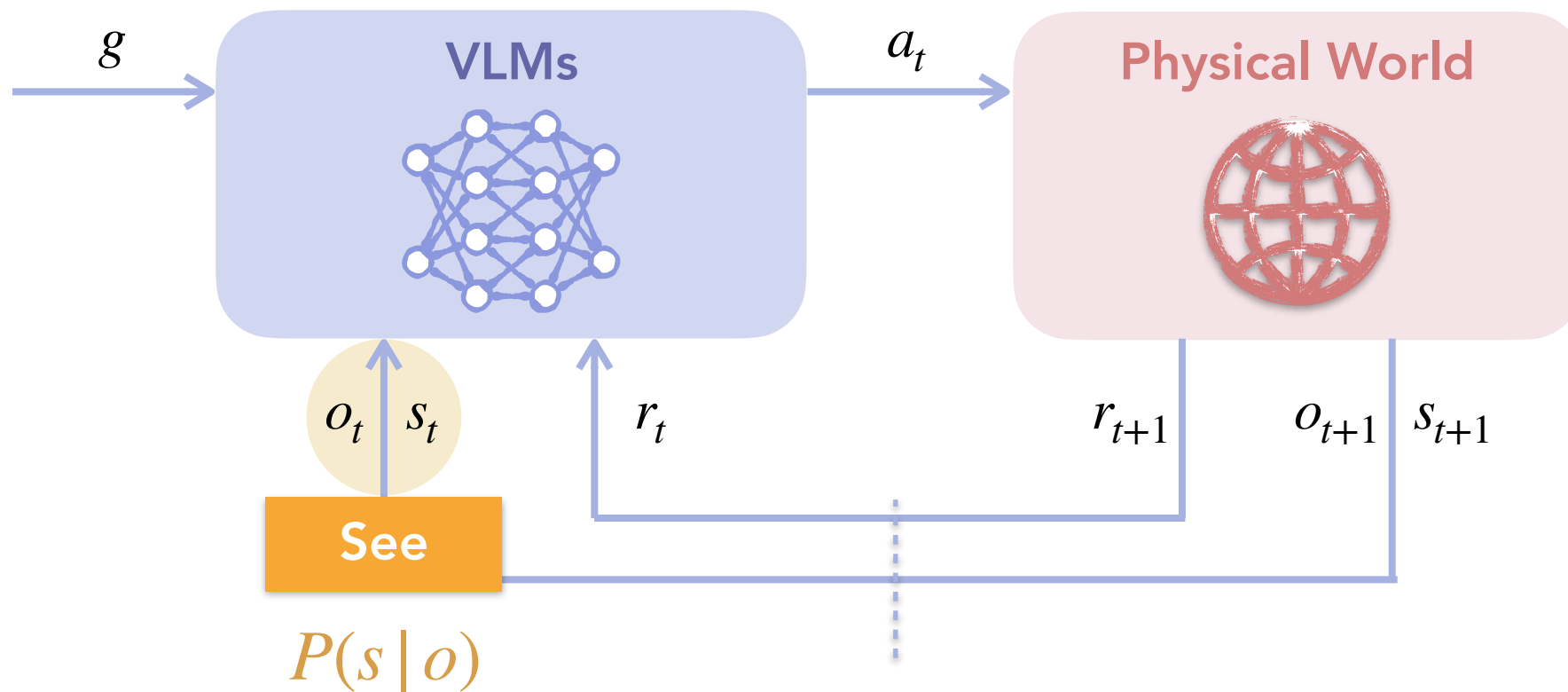
Cross-modality Alignment

Learning Paradigm

Exploration over Exploitation

Control

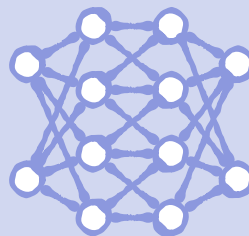
Safety and Steering



~~X~~ Above



VLMs (such as LLaVA)



Where is the **cat** in relation to the **person**?

Why Is Spatial Reasoning Hard in VLMs?

Let's open up VLMs!

~~Above~~

88.3% tokens



11.7% tokens

Where is the **cat** in relation to the **person**?

Let's open up VLMs!

89.3%
attention

Vision is barely focused on! 😞

10.7% attention

88.3% tokens

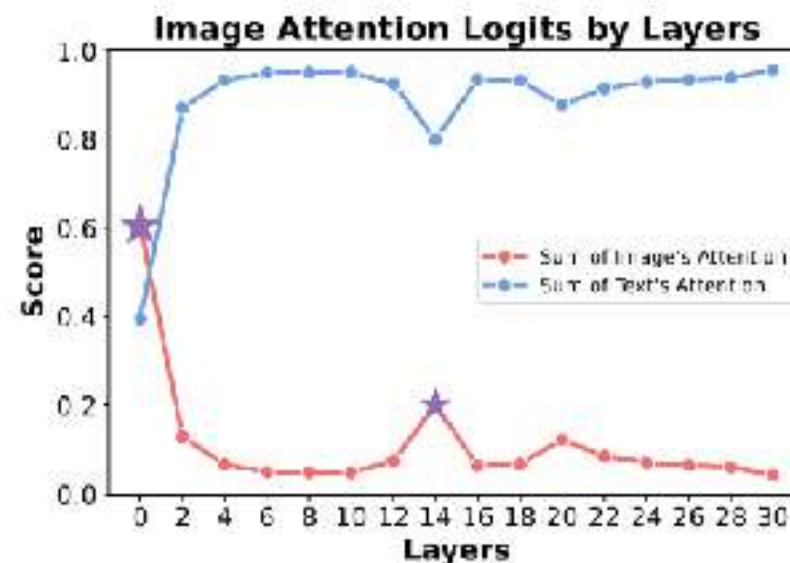
11.7%
tokens



Where is the **cat** in relation to the **person**?

Let us simply **add more** attention

Let's open up VLMs!

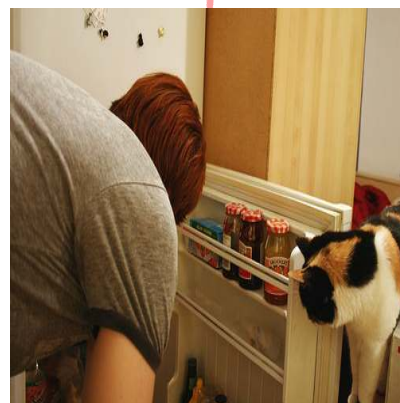


89.3%
attention

10.7% attention

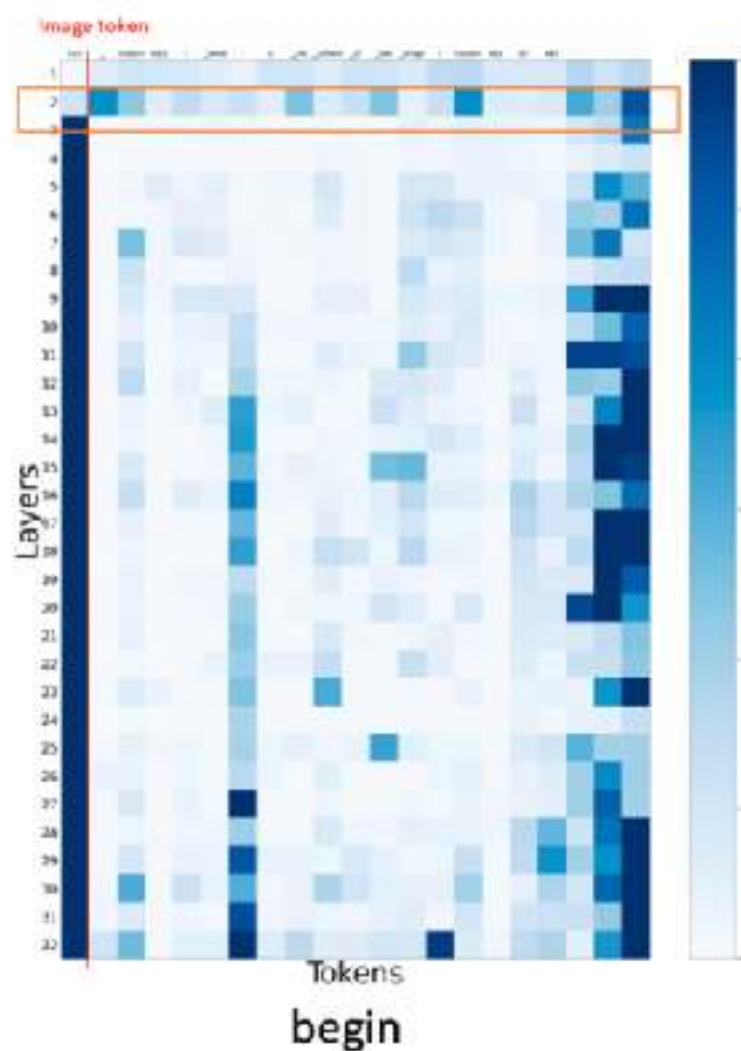
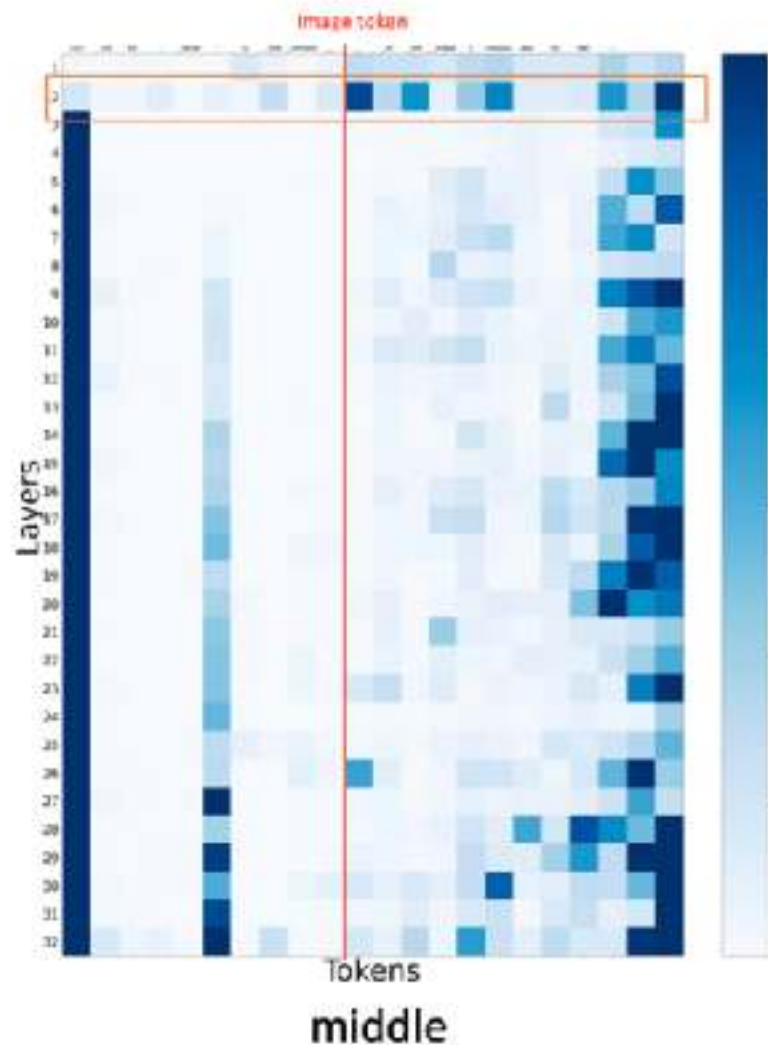
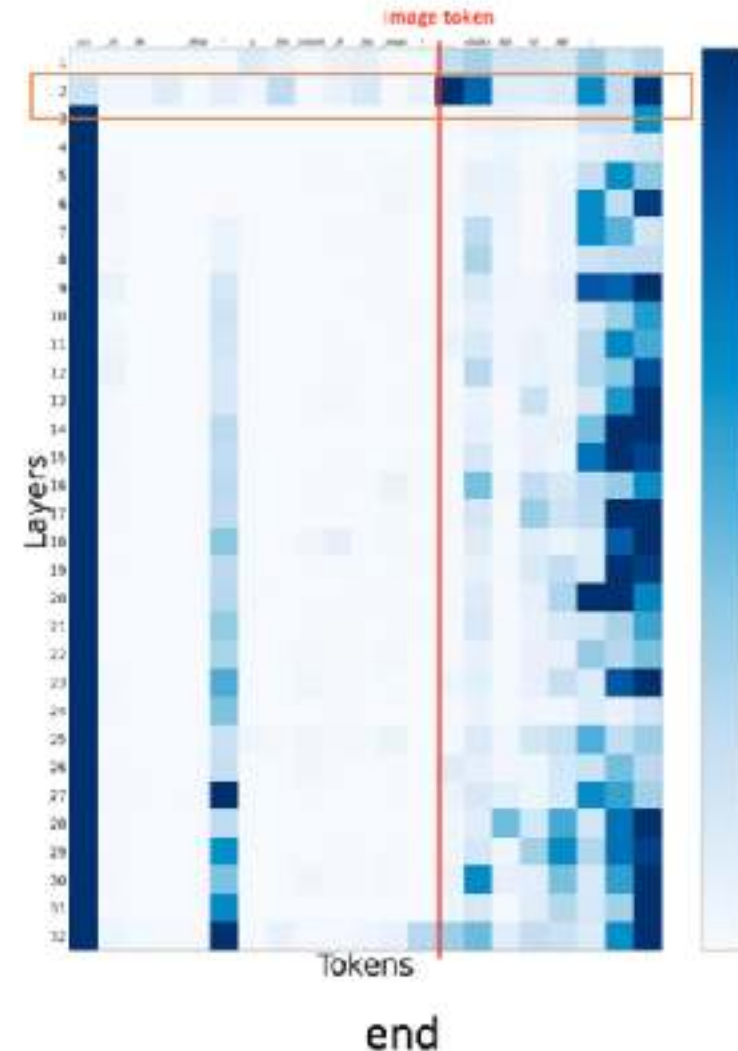
88.3% tokens

11.7% tokens

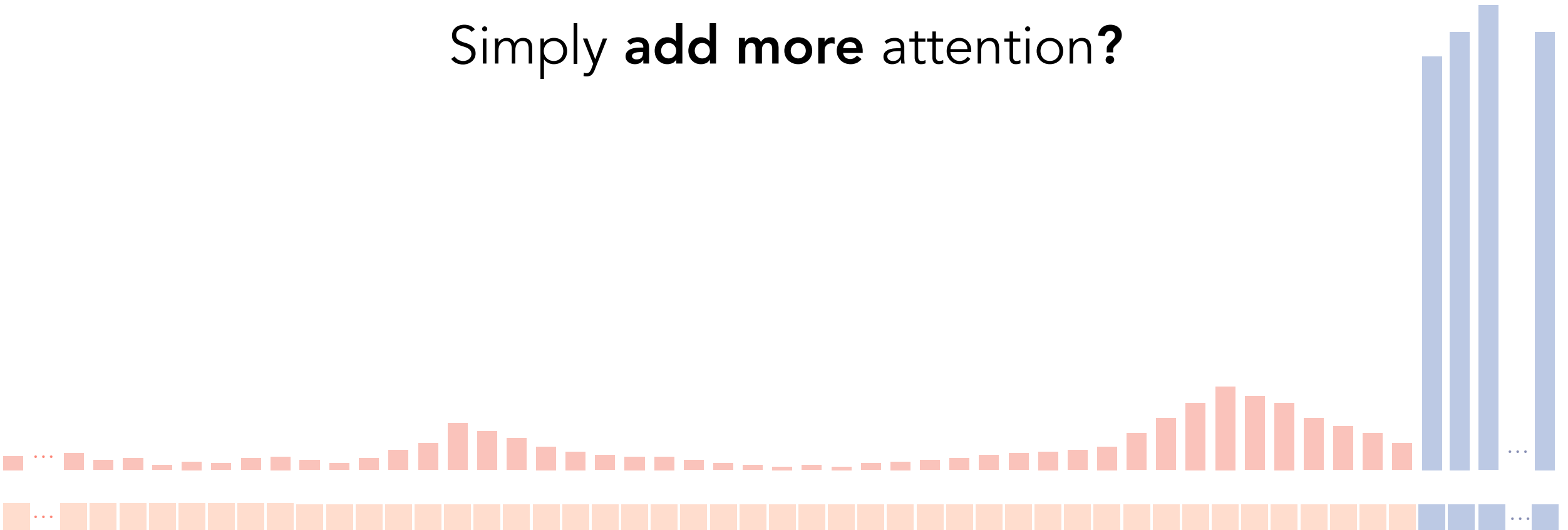


Where is the **cat** in relation to the **person**?

Input: <image>\nUSER: What's the content of the image?
\nASSISTANT:A cup to the **right** of a candle.

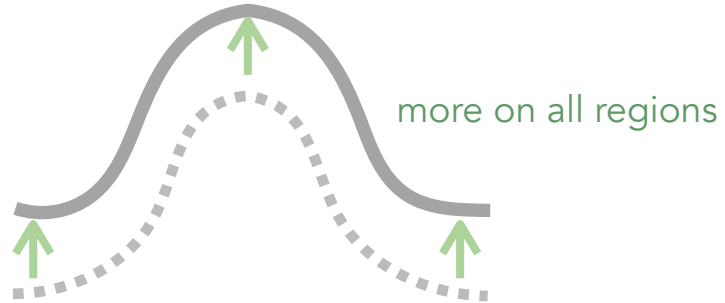
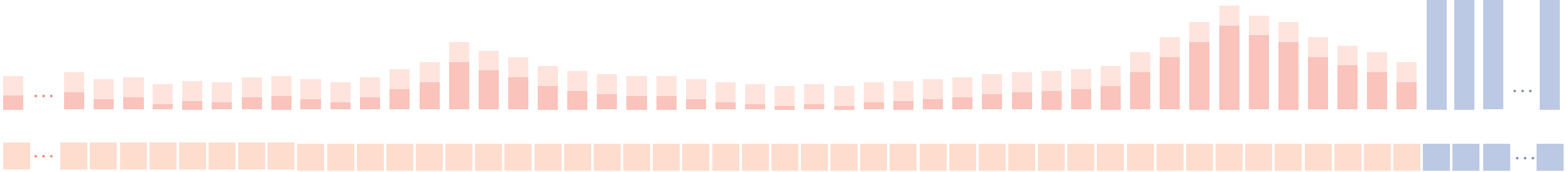


Simply **add more** attention?



Where is the **cat** in relation to the **person**?

Simply **add more** attention?



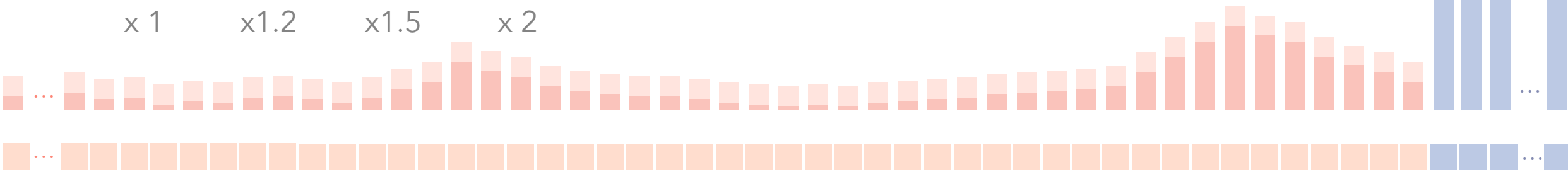
Where is the **cat** in relation to the **person**?



Simply add more attention?



Not Working 😞



Where is the **cat** in relation to the **person**?

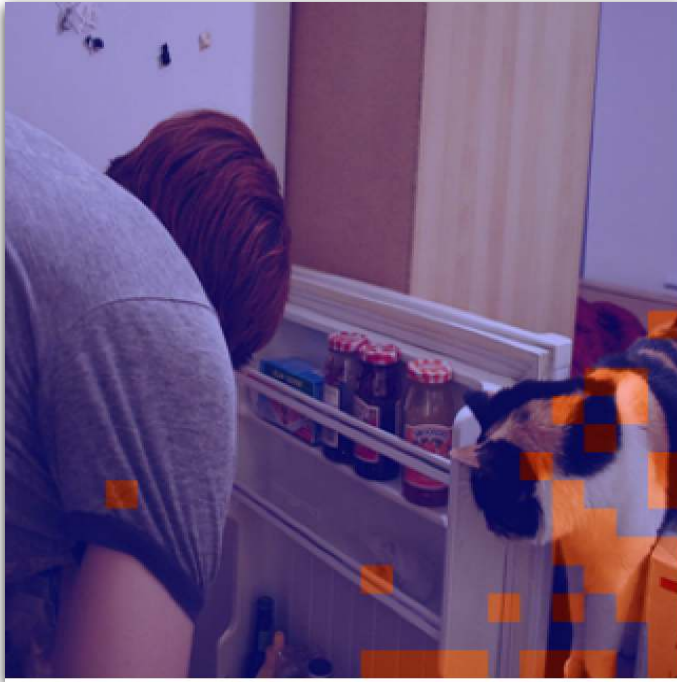
General Intervention



Geometry-Aware Intervention

Intervention over Attention Geometry

Where is the **cat** in relation to the **person**?



Above

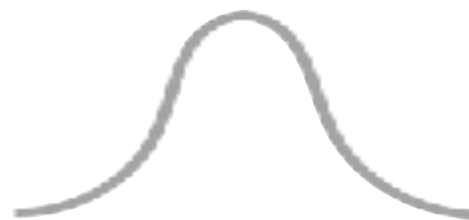


Intervention over Attention Geometry

Where is the **cat** in relation to the **person**?

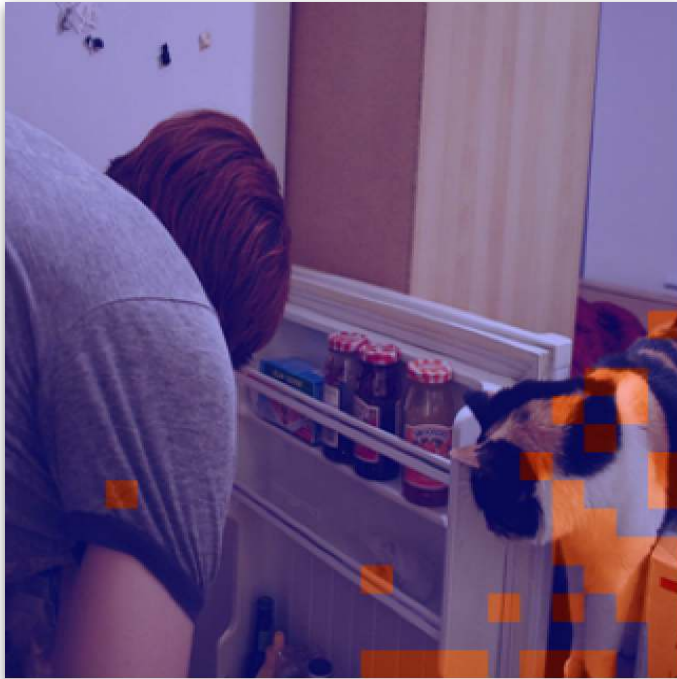


Above

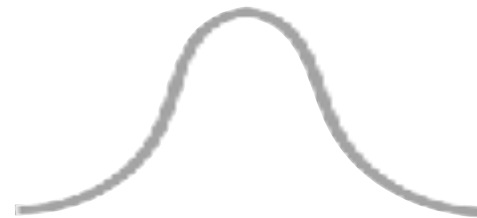


Intervention over Attention Geometry

Where is the **cat** in relation to the **person**?

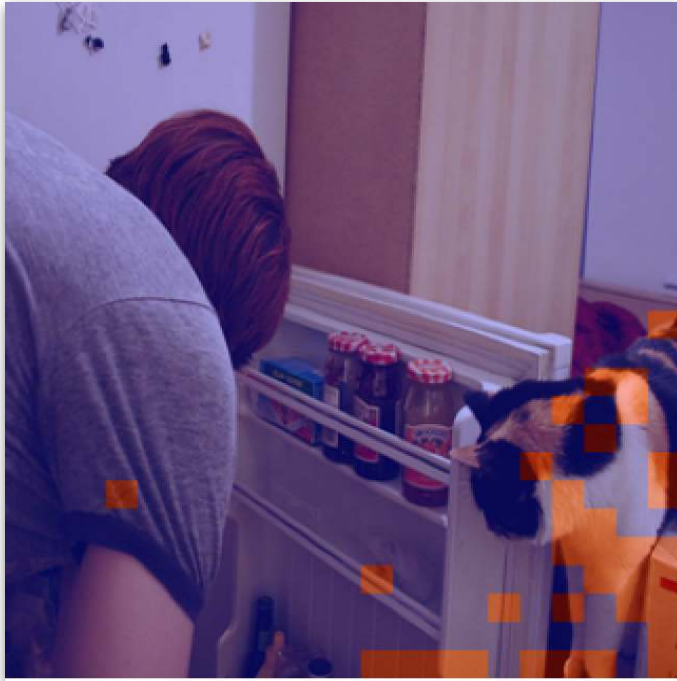


Above

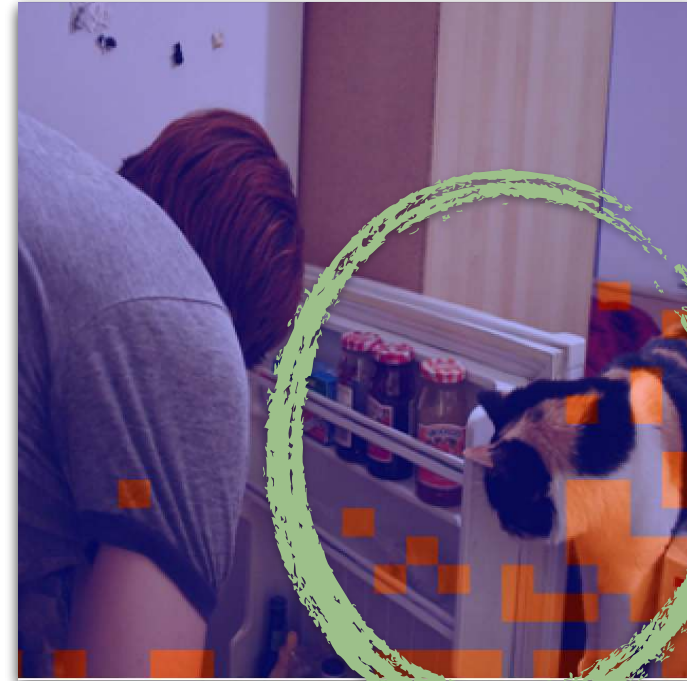


Intervention over Attention Geometry

Where is the **cat** in relation to the **person**?



Above



Right



more on high-attention regions

less on low-attention regions

Compared to **adding more** attention



How about **missing regions**?

Intervention over Attention Geometry

Where is the **book** in relation to the **candle**?



Intervention over Attention Geometry

Where is the **book** in relation to the **candle**?



Left



Intervention over Attention Geometry

Where is the **book** in relation to the **candle**?



Left



Intervention over Attention Geometry

Where is the **book** in relation to the **candle**?



Left



Intervention over Attention Geometry

Where is the **book** in relation to the **candle**?



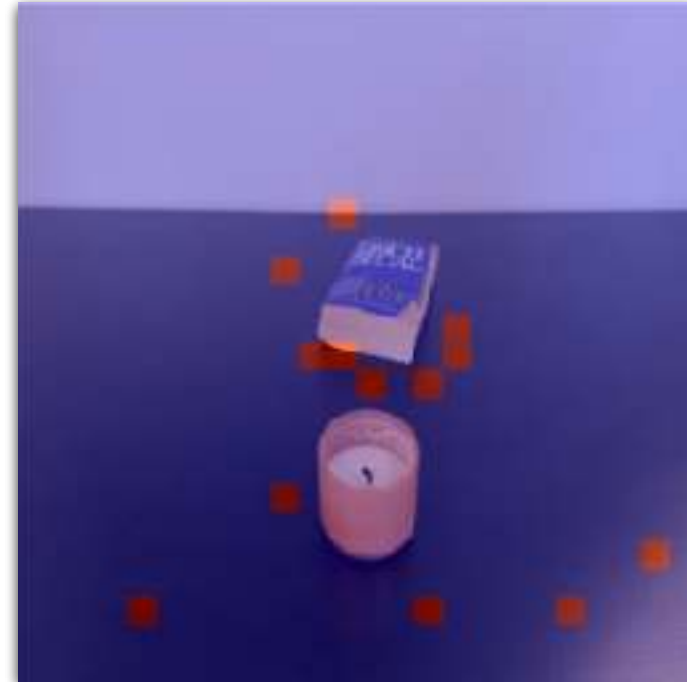
→ Left ❌



→ Behind ✅



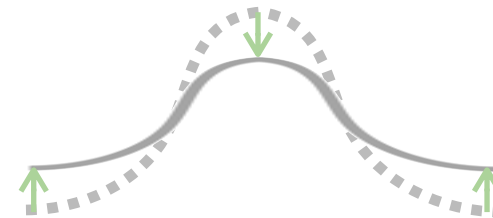
Intervention over Attention Geometry



Intervention over Attention Geometry



Sharpen



Smoothen

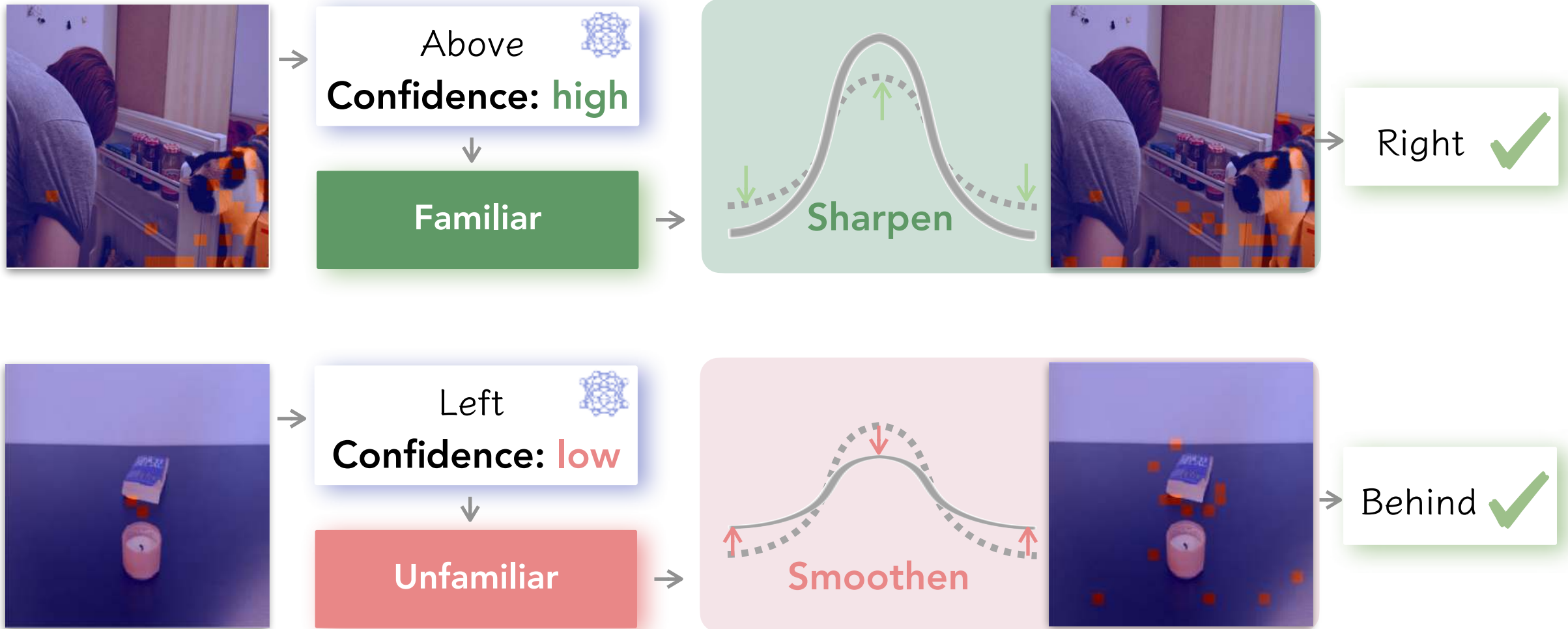
When **sharpen**? When **smoothen**?

When **sharpen**? When **smoothen**?

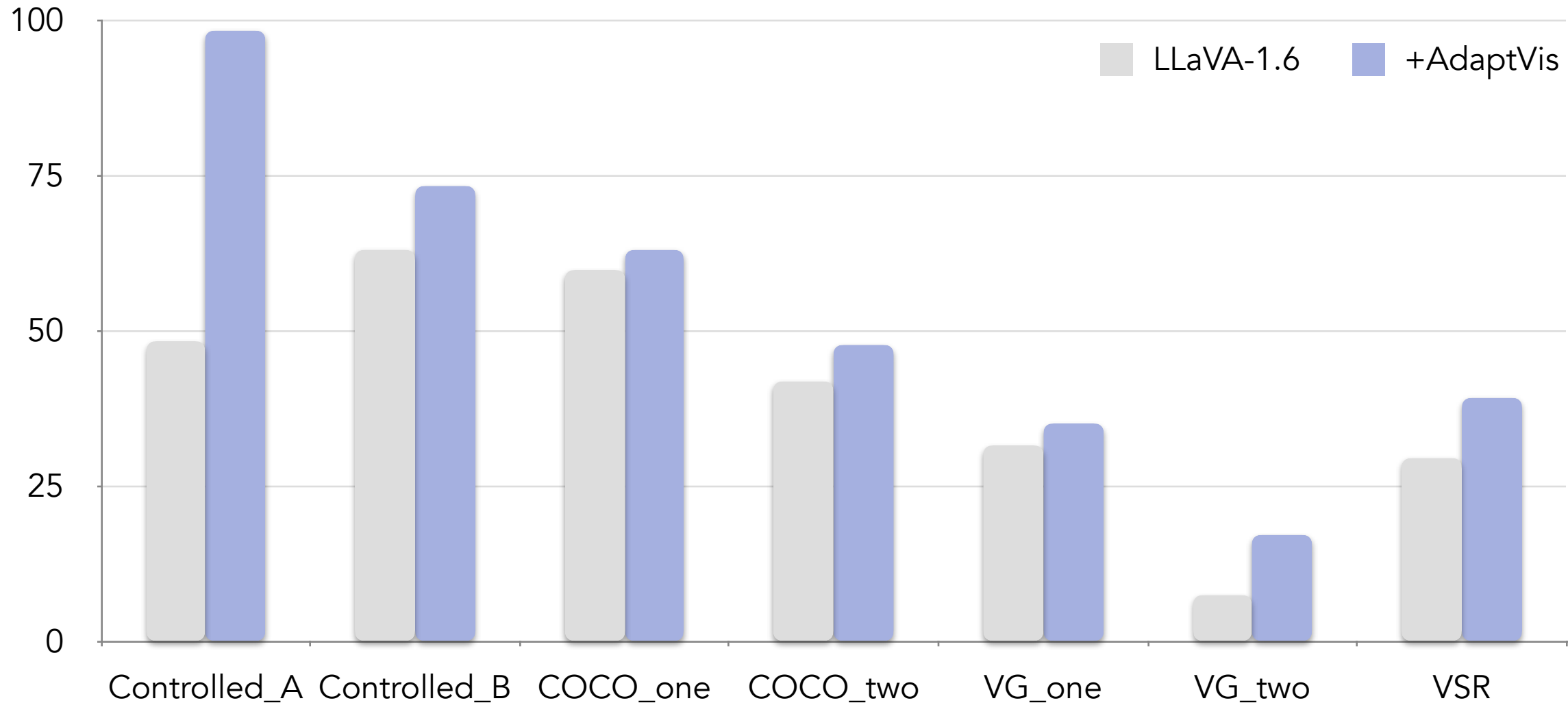
Can we trust a model's attention pattern?

→ **Confidence** score.

When **sharpen**? When **smoothen**?



Such simple interventions have impressive results



Attention behavior of VLMs from a **mechanism interpretability** lens

1. What **causes** these failures? How do they manifest **internally**?

→ **Overshadow** of attention distributions

2. How to **mitigate these failures** internally?

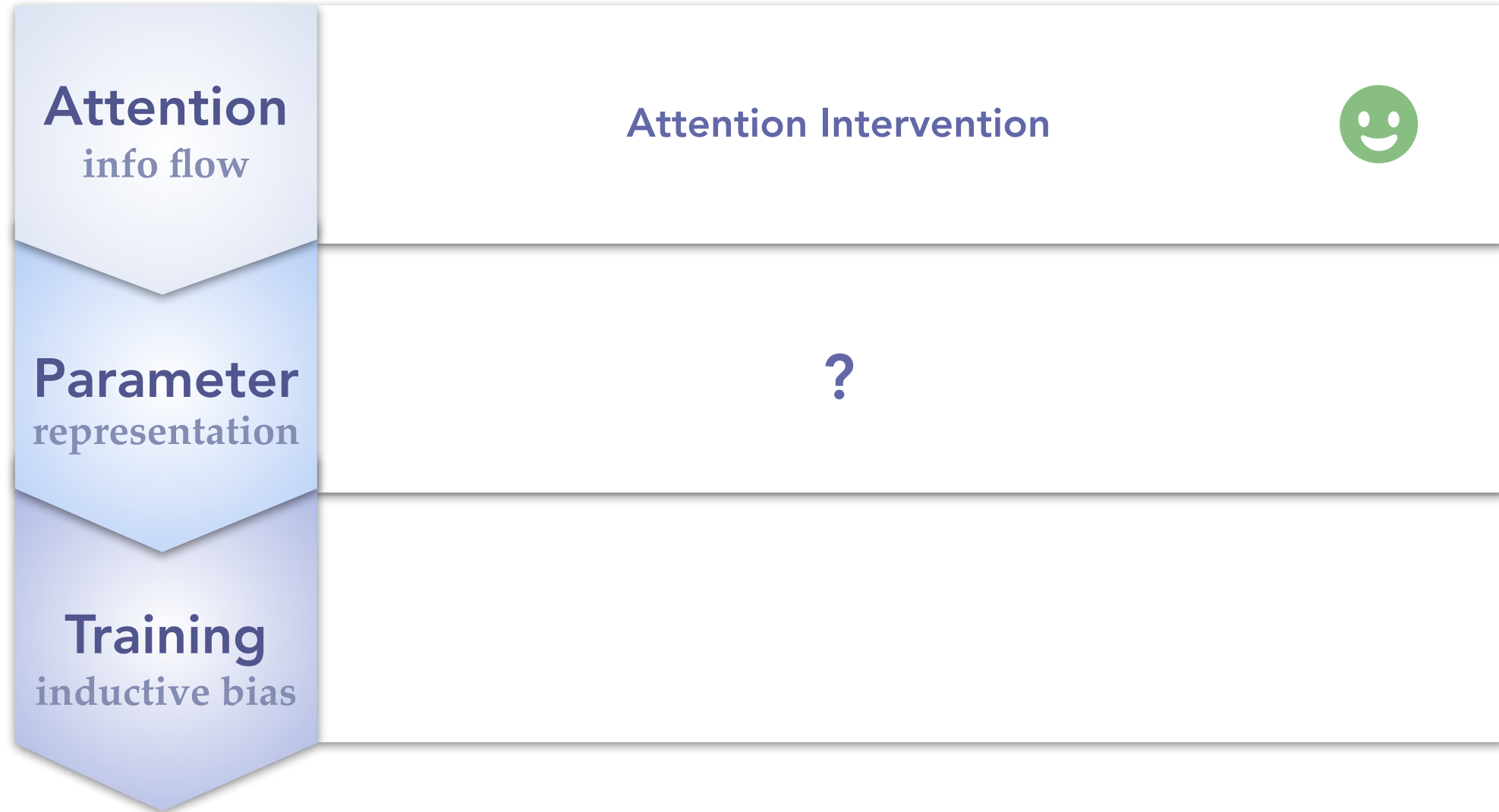
→ Intervention over the **overall** attention distribution



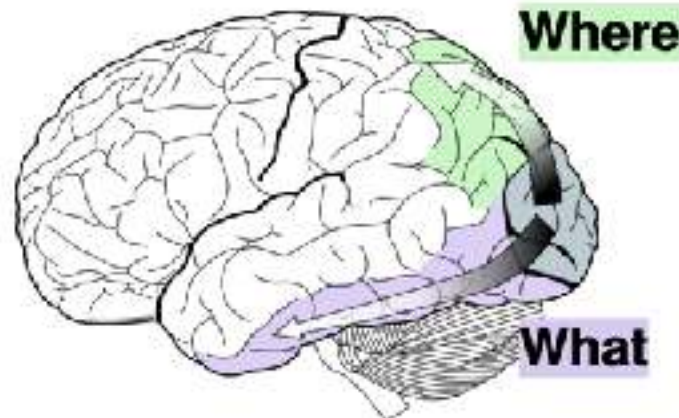
→ Intervention over the **geometry** of attention distribution



Part I: **Geometry Aware** Alignment



Inject Geometry Priors via **Parameter Merging**

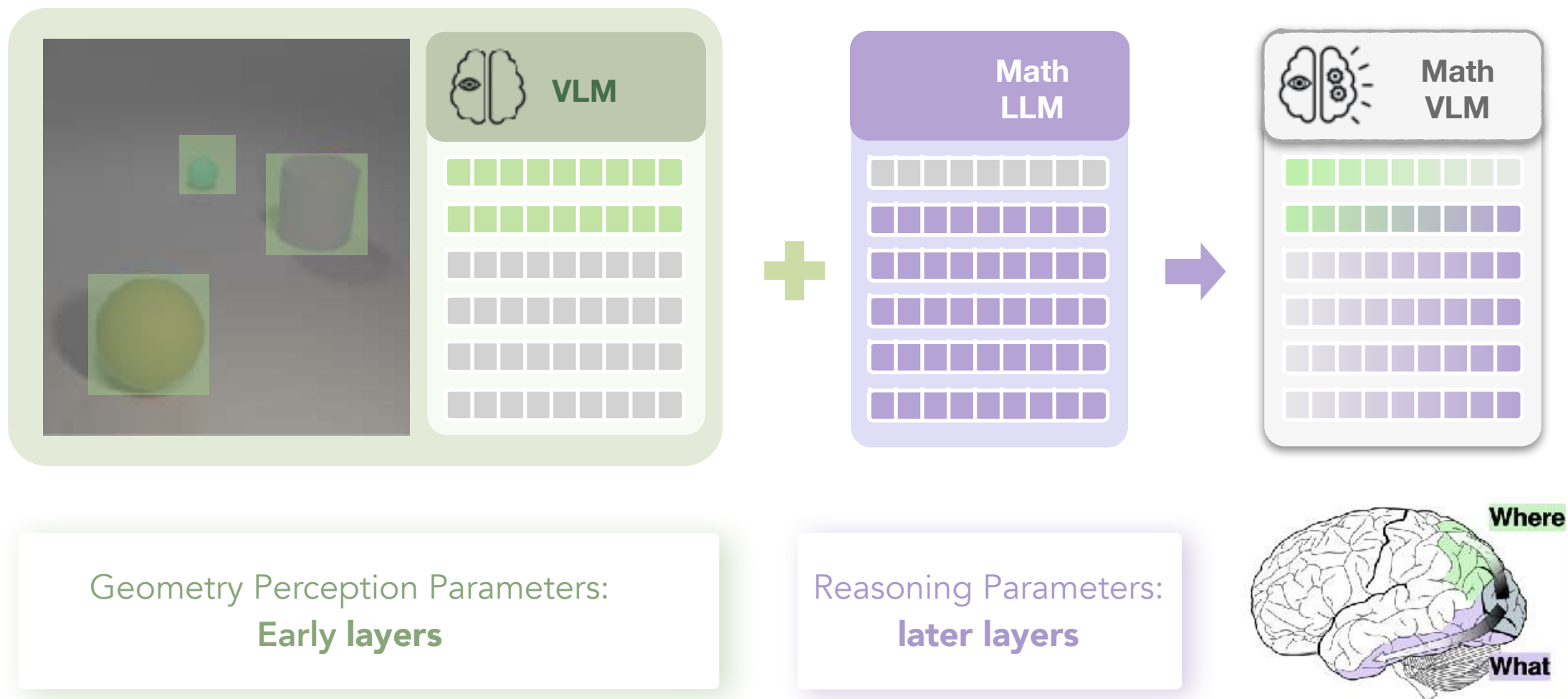


Two-streams Hypothesis an interesting human cognition analogy:

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

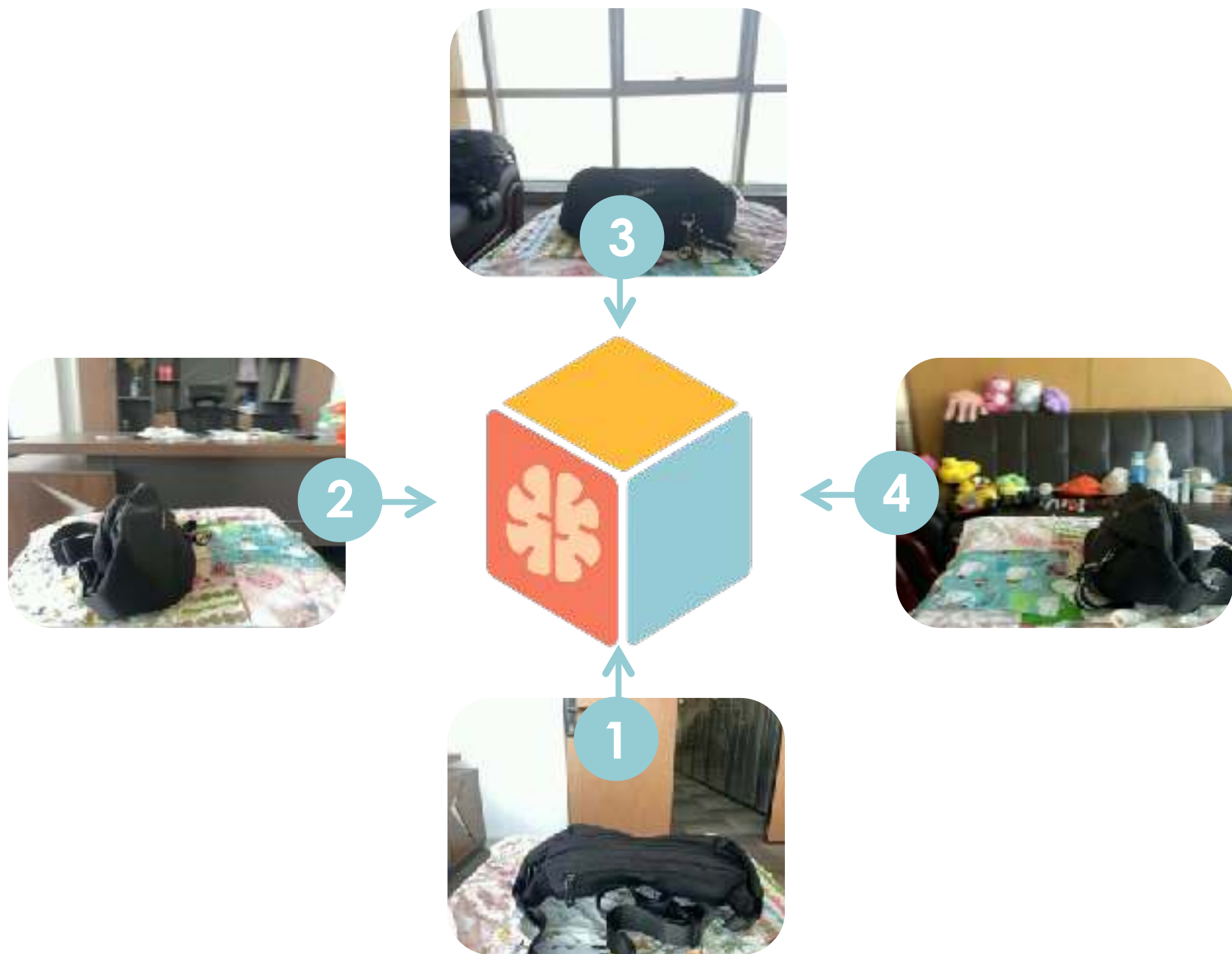
The dorsal stream (or, "where pathway") leads to the parietal lobe, which is involved with processing the object's spatial location relative to the viewer and with speech repetition.

Inject Geometry Priors via **Parameter Merging**



Part I: **Geometry Aware** Alignment





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

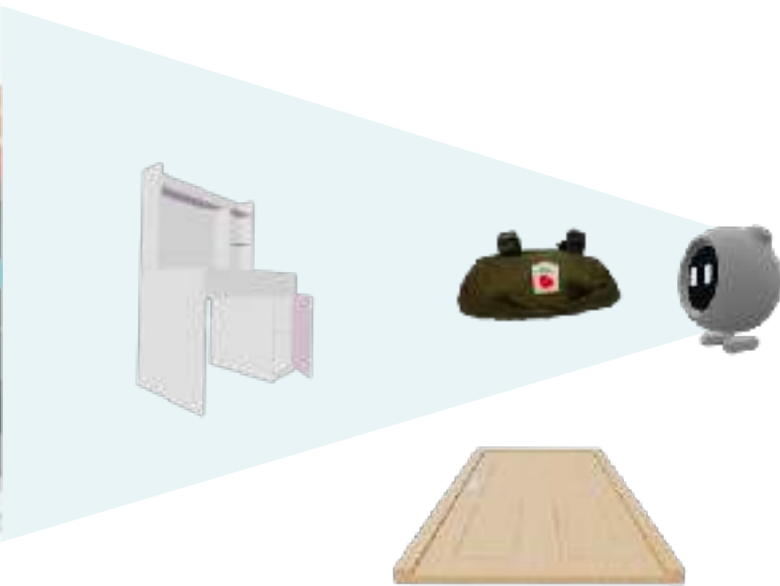
B. Door

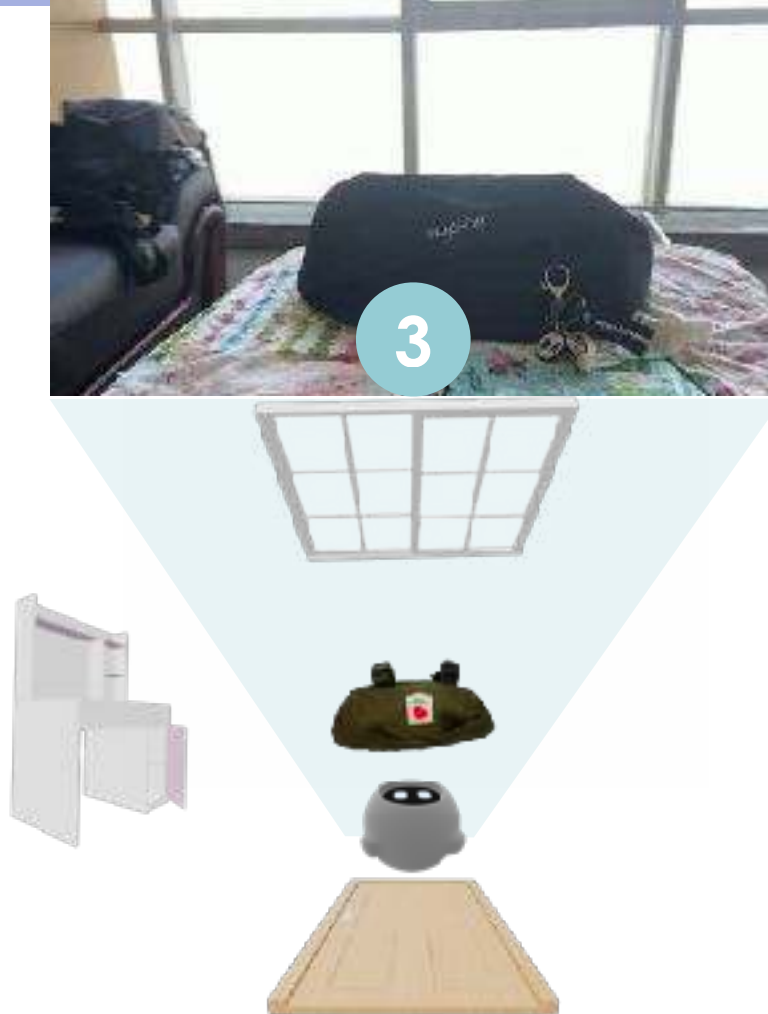
C. Desk

D. Sofa

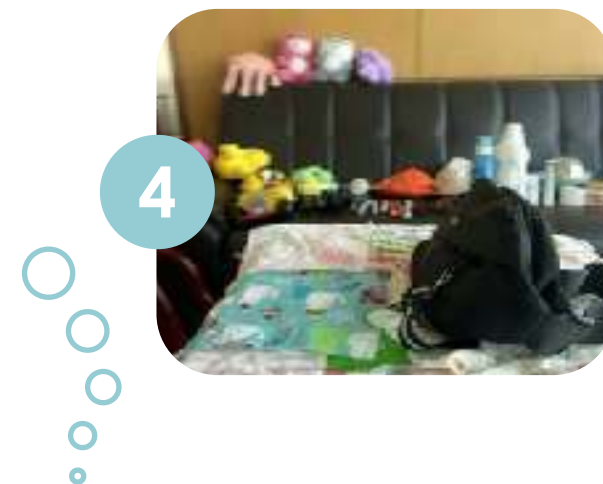












At **4** what is to the **left** of the **black bag** ?



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

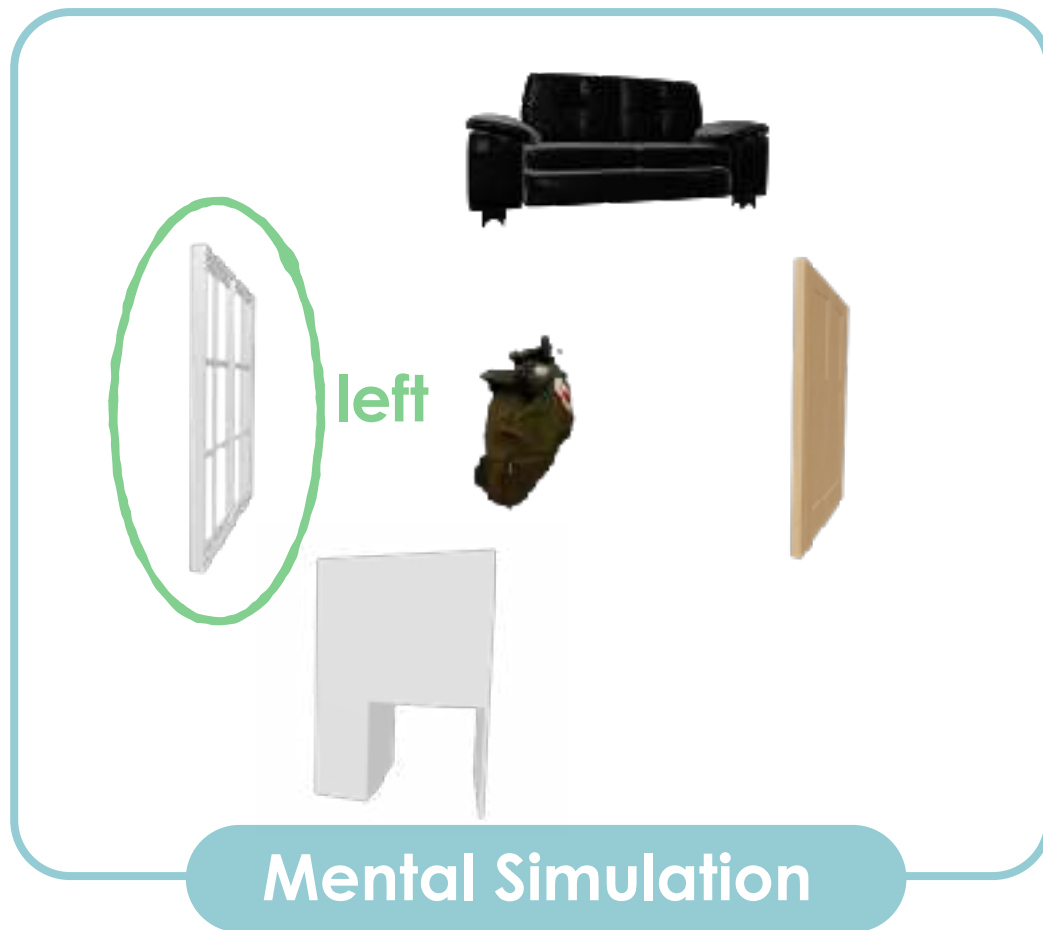
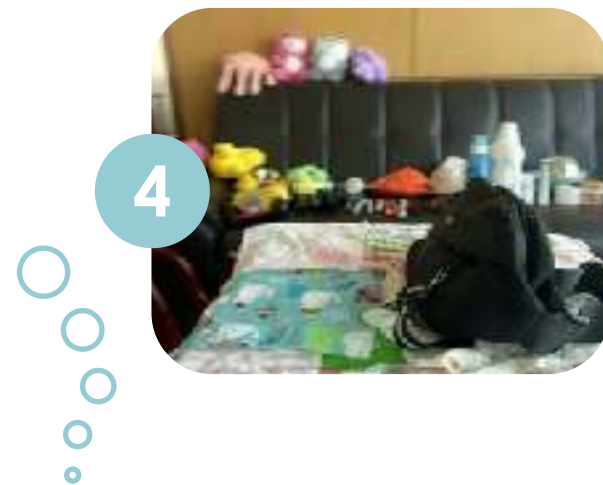
A. Window



B. Door

C. Desk

D. Sofa









MindCube

21,154 questions

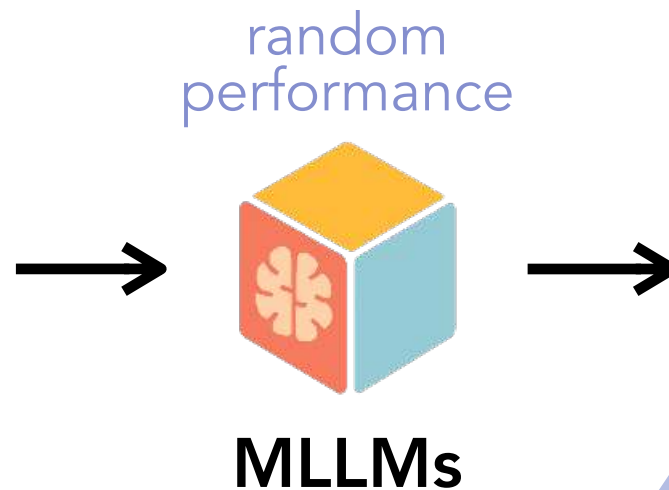
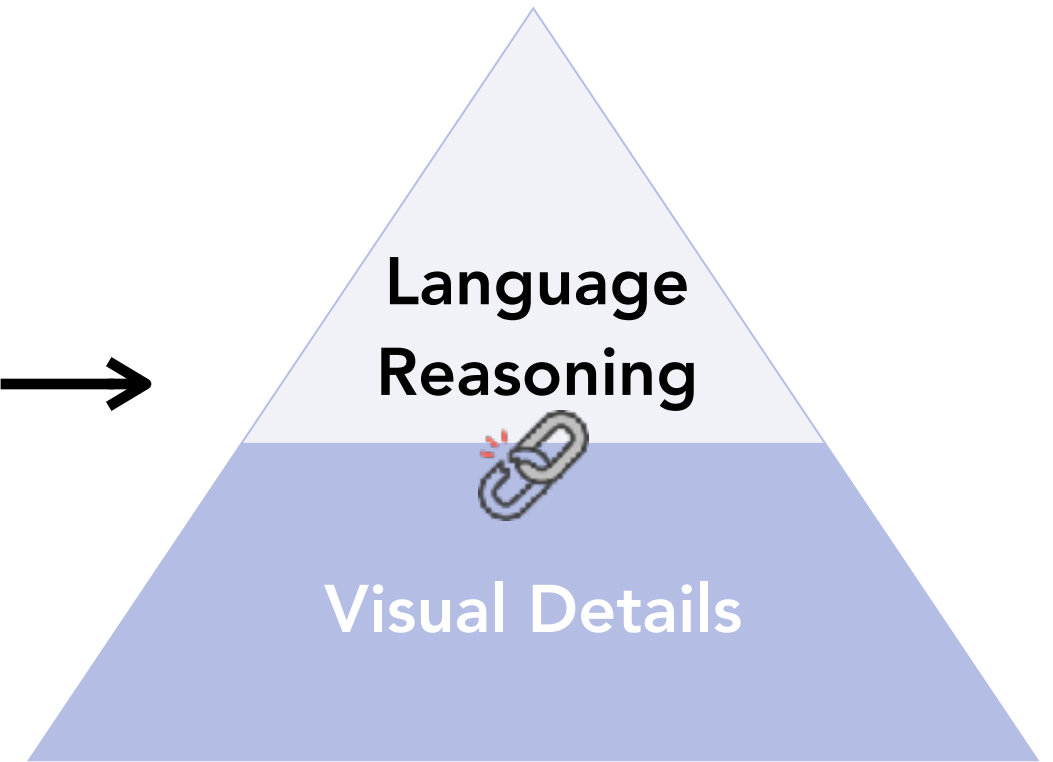
3,268 images

Inject **Geometric Priors** with Cognitive Map **Abstraction**



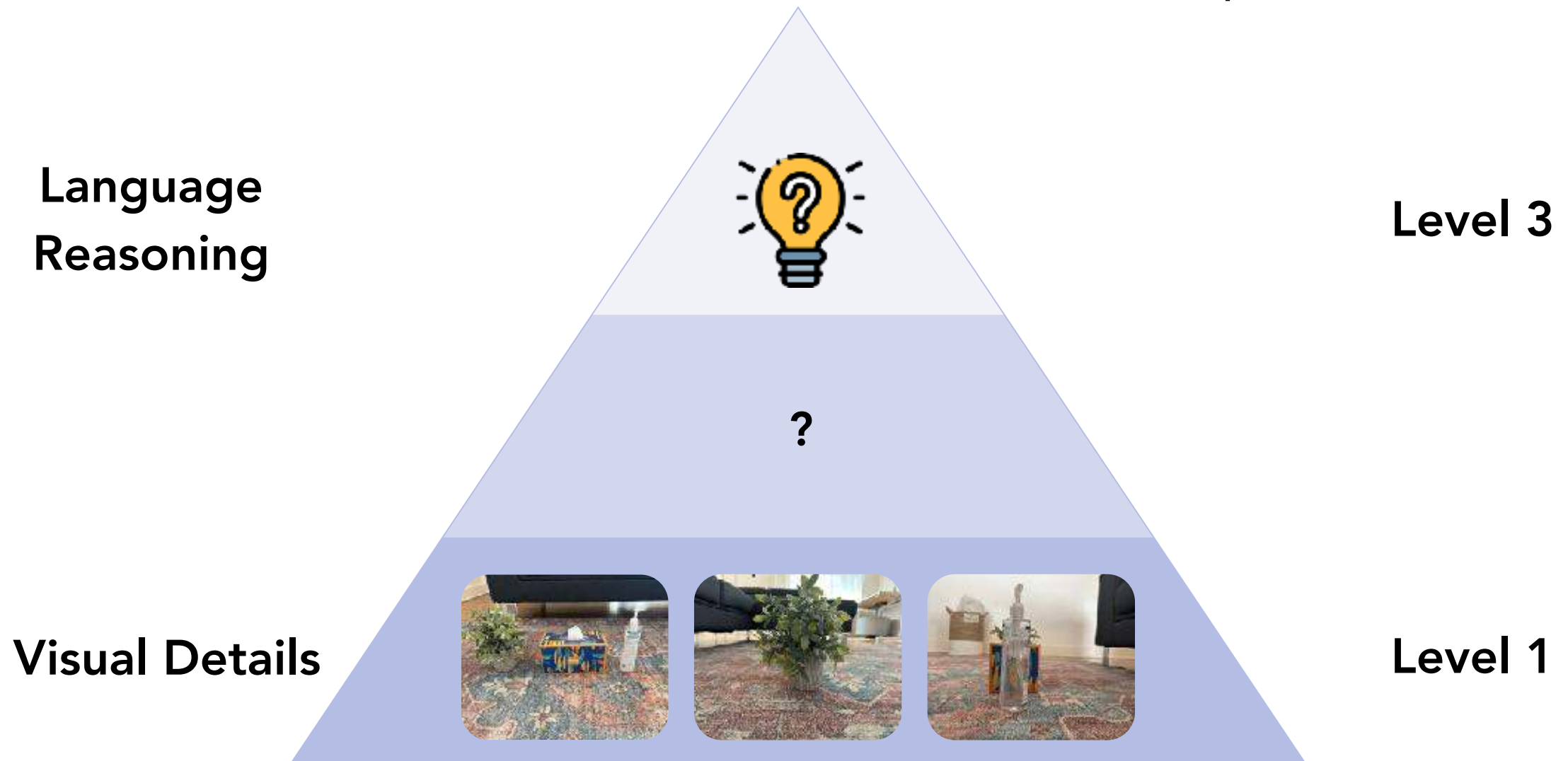
Describe layout at viewpoint **2**

Multi-View Question

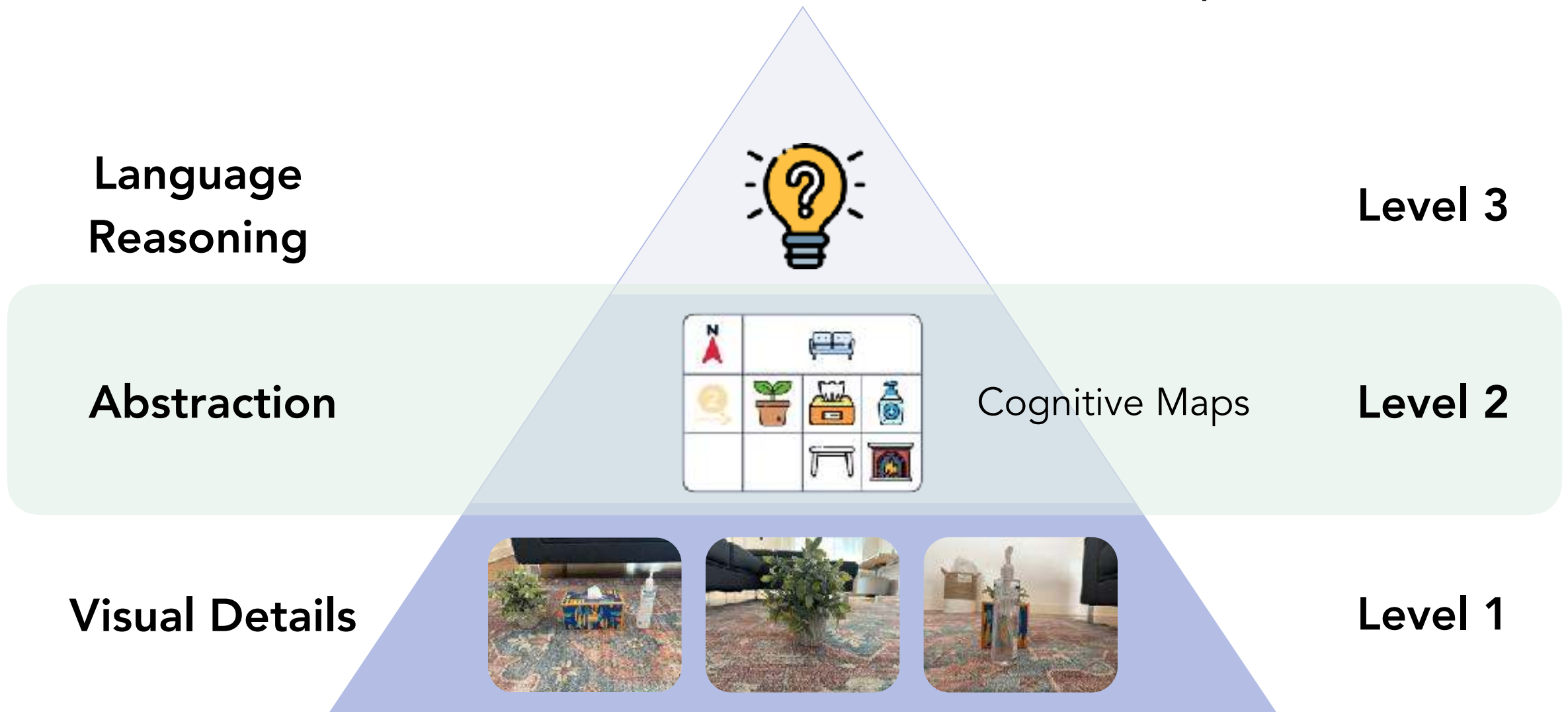


Gap Exists Between Modalities

Inject **Geometric Priors** with Cognitive Map **Abstraction**



Inject **Geometric Priors** with Cognitive Map **Abstraction**

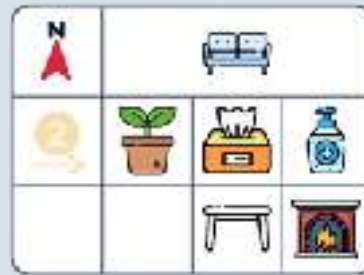


Inject **Geometric Priors** with Cognitive Map **Abstraction**

Language
Reasoning

Abstraction

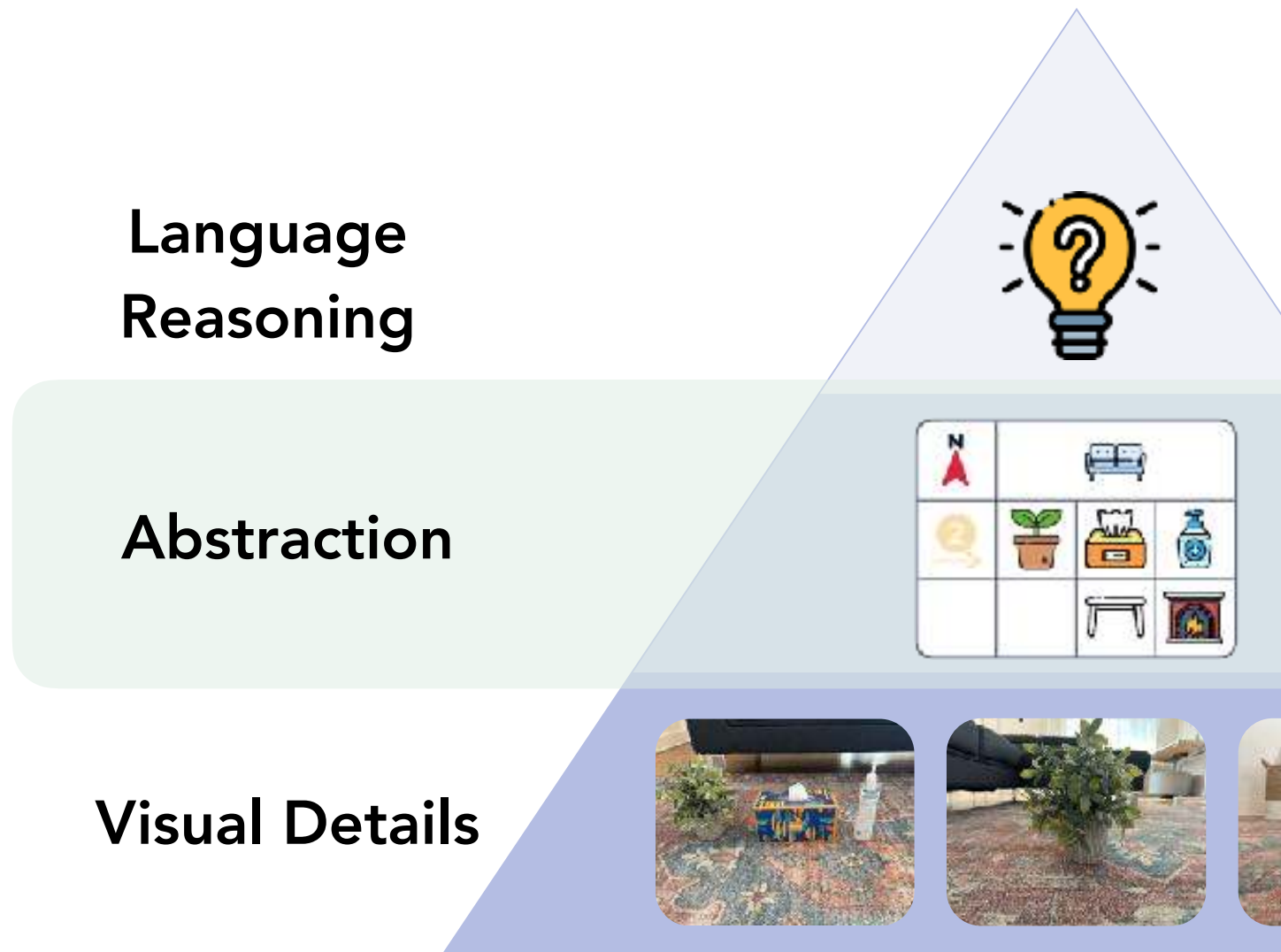
Visual Details



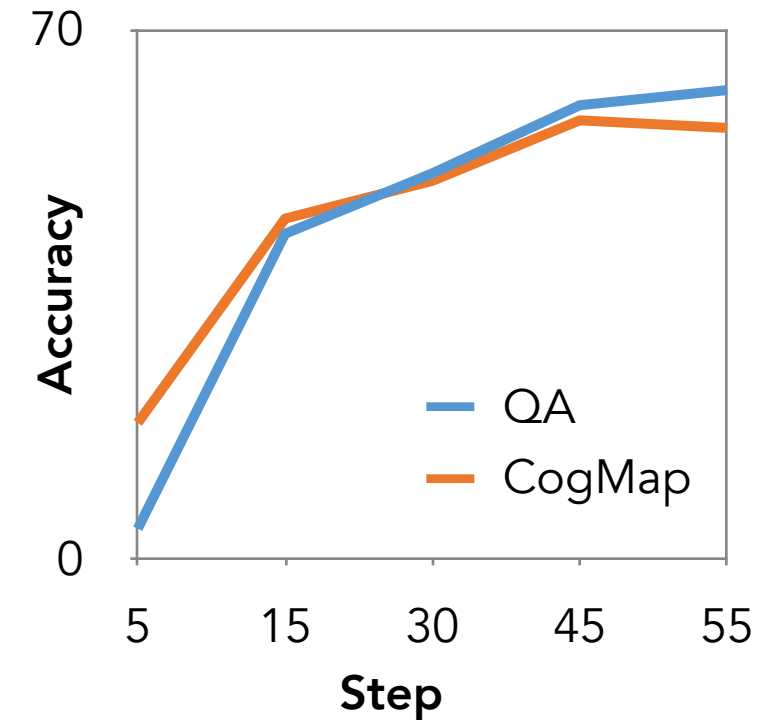
Augmented Cognitive Map

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

Inject **Geometric Priors** with Cognitive Map **Abstraction**



Better QA → CogMap



QA improvement **strongly correlates** with CogMap accuracy.

Part I: **Geometry Aware** Alignment



Future:

Multimodal Reasoning Intelligence

Future:

Multimodal Reasoning Intelligence



VLA Challenge



LLM Challenge



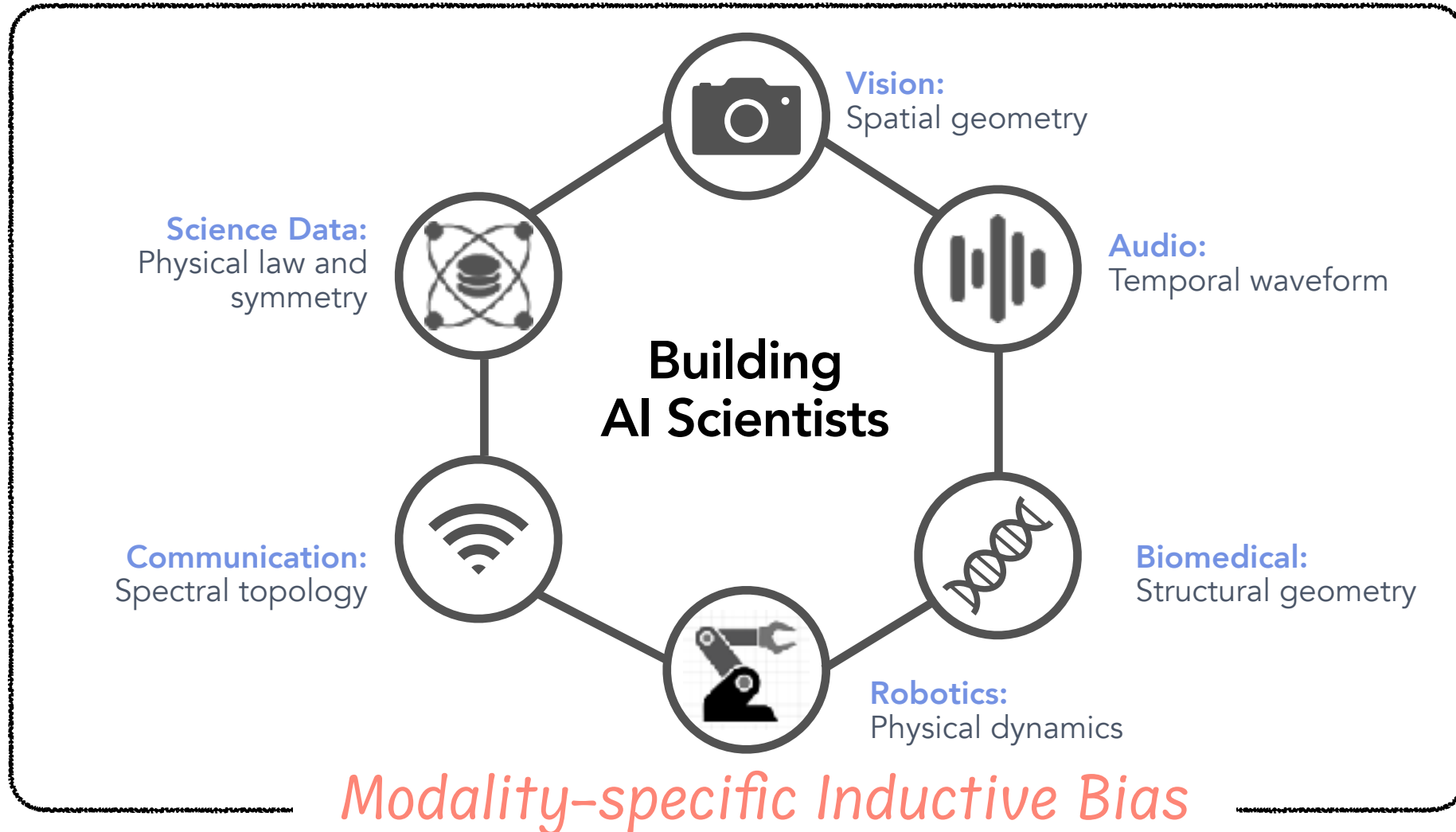
Workshop series



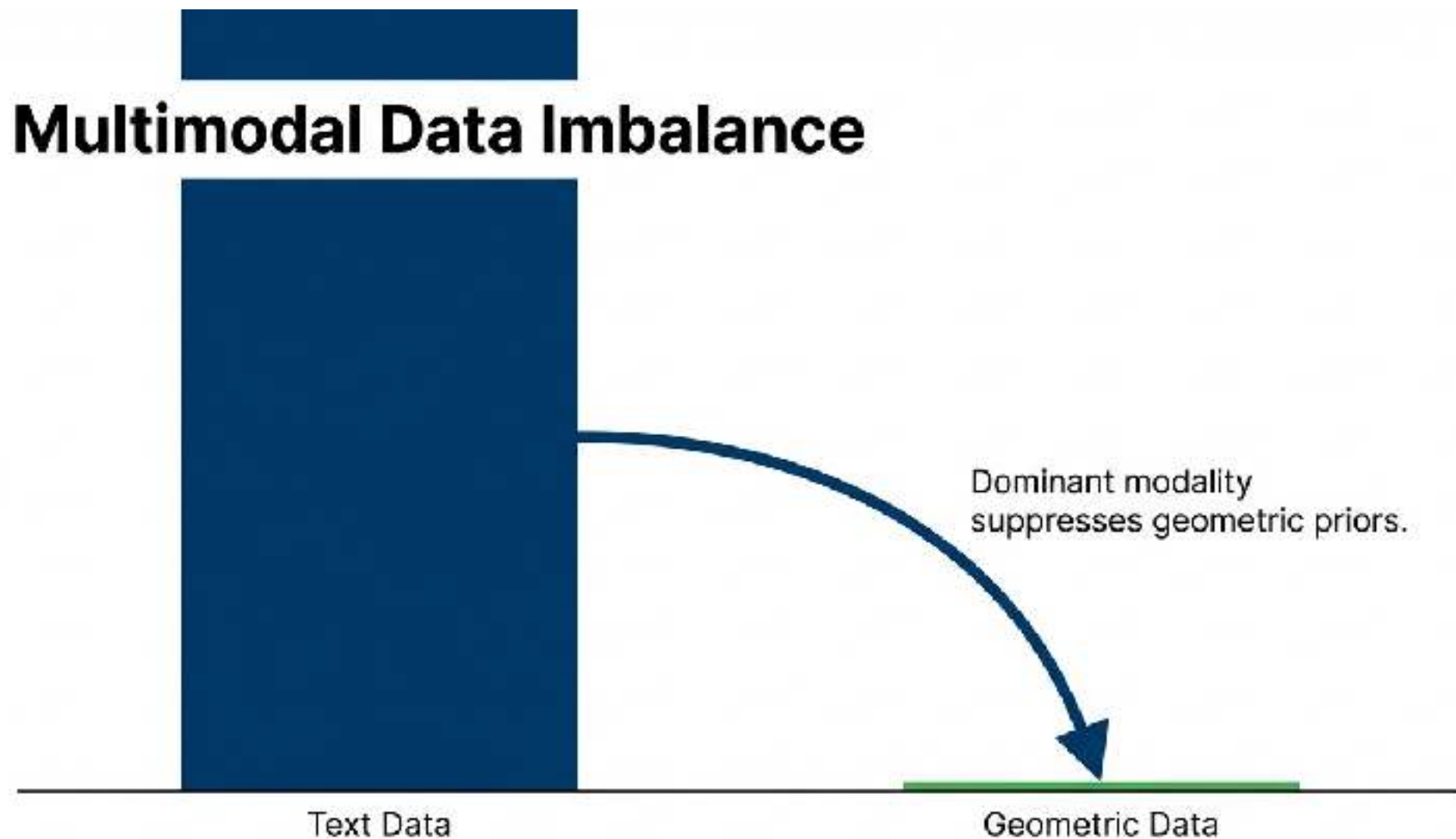
Tutorial series

Multimodality = Different Data Priors

Other modality $\neq$ Other language

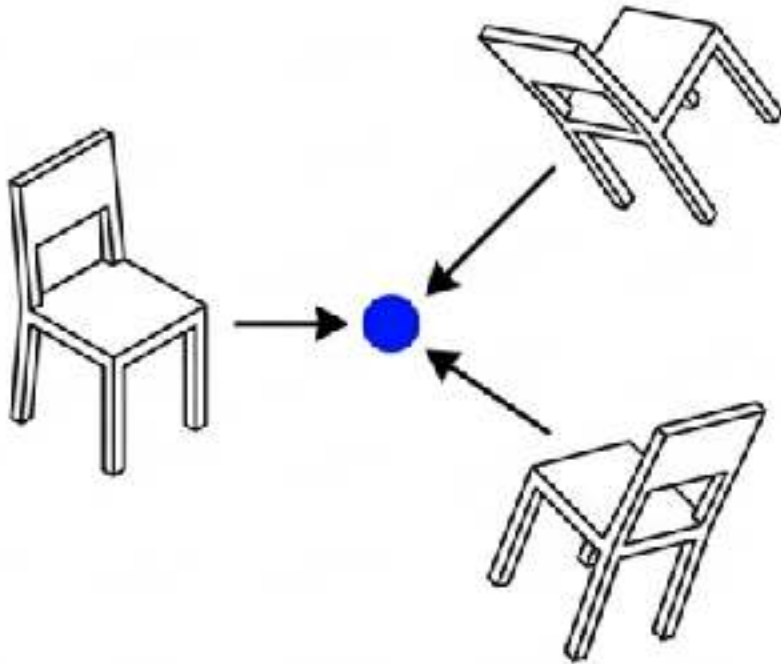


Language Overshadow



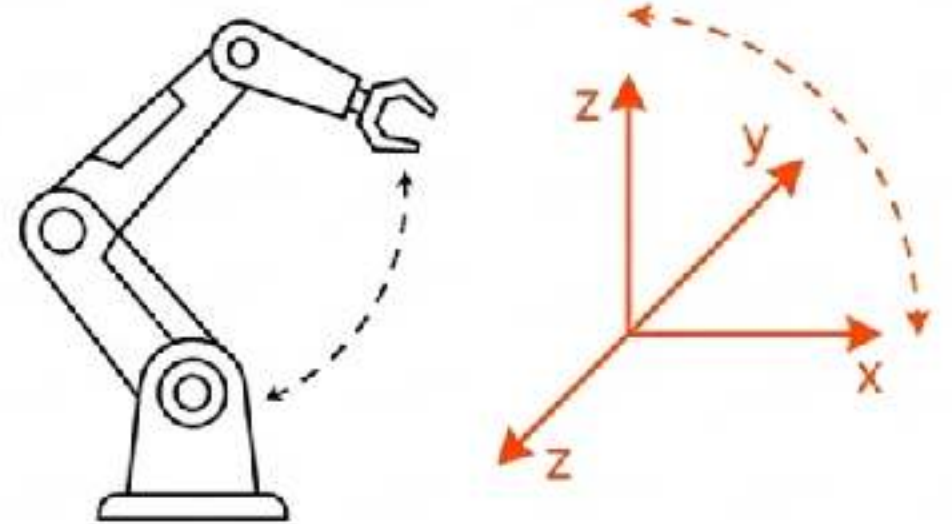
What is a good representation?

$$f(T(x)) = f(x)$$



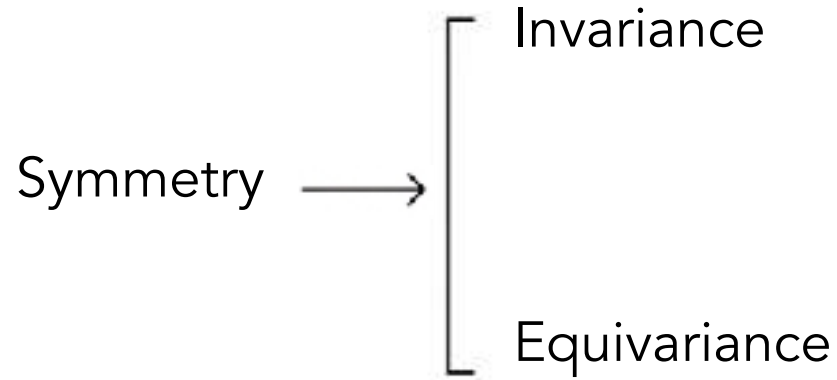
Invariance

$$f(T(x)) = T(f(x))$$



Equivariance

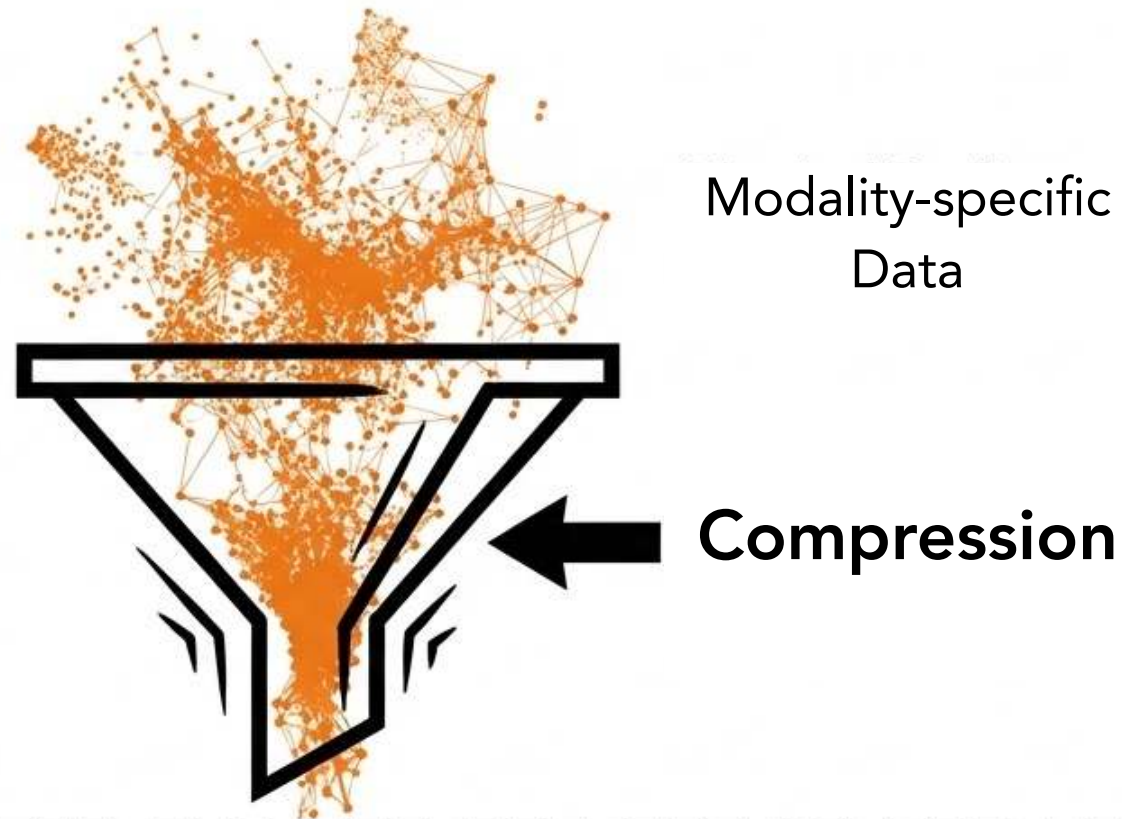
What is a good representation?



Compression is NOT dropping data.

Compression is factorizing redundancy through **symmetry**.

Compression Law



Compression is NOT dropping data.

Compression is factorizing redundancy through symmetry.

Abstraction as Compression

Model vs Human:

Data



Abstraction



Model



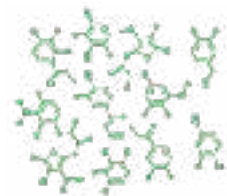
1 Million Molecules



Statistical
Correlation



Human



10 Molecules



Chemical Law
(Abstraction)

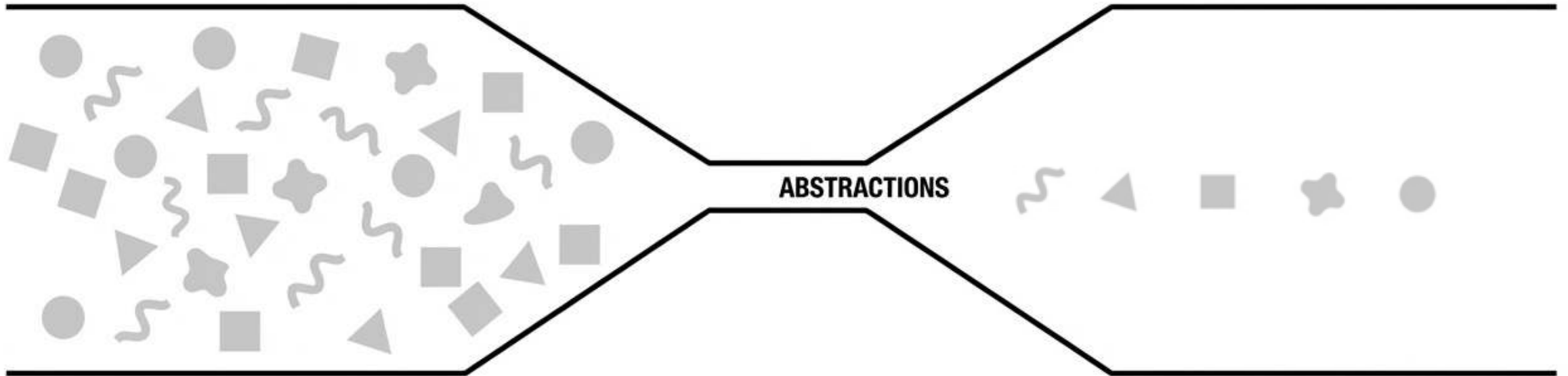
Abstraction as Compression

Model vs Human:

Data



Abstraction







What's Left for Humans?

Complexity Inflation: OPENCLAW

~1M 16K+

Lines of Code

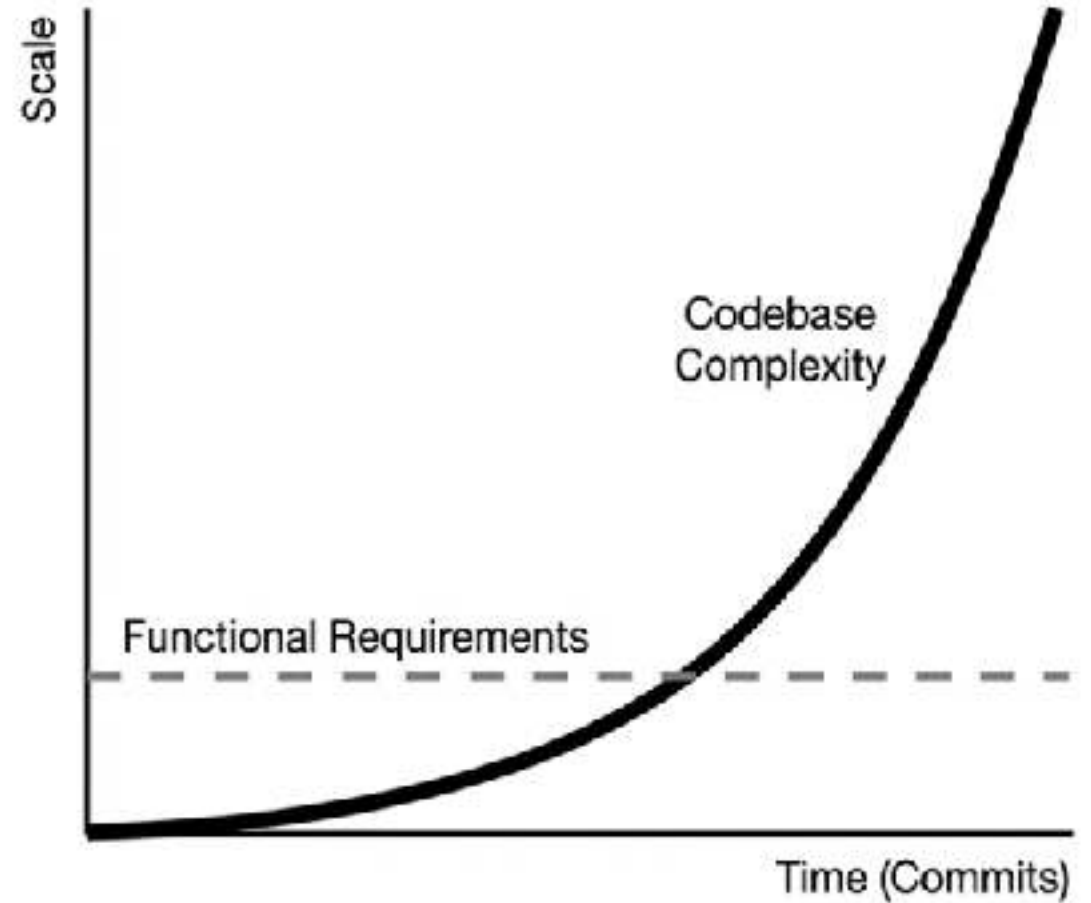
Commits

Ugly, Not Manageable

~400

lines of code

Hugo Dutka Blog Post



AI currently learns how to grow. It solves problems by adding code and layers.

Negentropic Capability



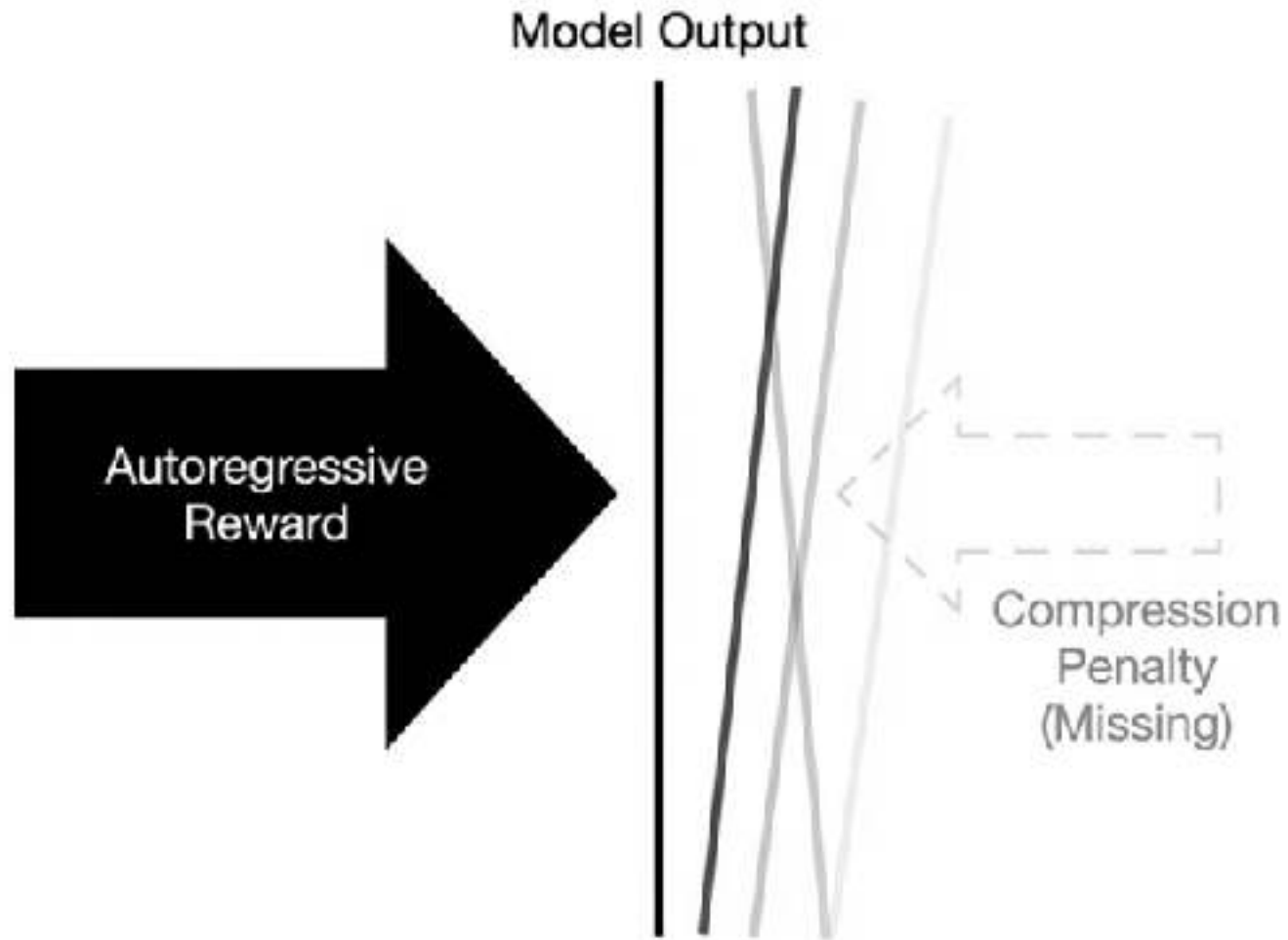
Schrödinger (1944): "Life feeds on negative entropy."

Living systems maintain order against entropy increase by extracting negentropy from the environment.

Autoregressive generation is an entropy engine.

It predicts the next token, perpetually expanding and unrolling.

Compression as Intelligence



Brute-force Scaling

**A bloated
1-million-line
codebase.**

**A bloated
1-million-token
video context window.**

Algorithm for the win !

- 🏆 MIT Tech Review 35 Innovators Under 35 global list
- 🏆 Schmidt Sciences AI2050 Early Career 2026 finalist
- 🏆 DARPA Young Faculty Award 2026 finalist
- 🏆 Amazon Scholar
- 🏆 Bloomberg Data Science PhD Fellowship 2026
- 🏆 NSF CS4Grad PhD Fellowship 2026
- 🏆 Patronus AI PhD Fellow
- 🏆 ACL 2025 Dissertation Award Honorable Mention
- 🏆 AAAI 2025 New Faculty Highlights
- 🏆 ACL 2024 Outstanding Paper Award
- 🏆 NAACL 2021 Best Demo Paper Award
- 🏆 ACL 2020 Best Demo Paper Award
- 🏆 Microsoft Research Postdoc Fellowship in 2023
- 🏆 DARPA Riser in 2022
- 🏆 Ranked 1st in NIST TAC SM-KBP 2022
- 🏆 EECS Rising Star in 2022
- 🏆 Microsoft Research PhD Fellowship in 2021
- 🏆 Ranked 1st in NIST TAC SM-KBP 2020

Thanks for the sponsors



Thank You