

# Building LLM Reasoners

## Lecture 12: Multimodal Reasoning

Greg Durrett



Administrative details and recap

Language Grounding

Grounding in Images

Visual Reasoning

Thinking with Images

Vision-Language-Action



Administrative details and recap

Language Grounding

Grounding in Images

Visual Reasoning

Thinking with Images

Vision-Language-Action

# Administrative details and recap

# Administrivia

---

- ▶ Check-ins returned soon
- ▶ Final project presentations next week
- ▶ Final projects due May 8
- ▶ Course feedback form released

Administrative details and recap

Language Grounding

Grounding in Images

Visual Reasoning

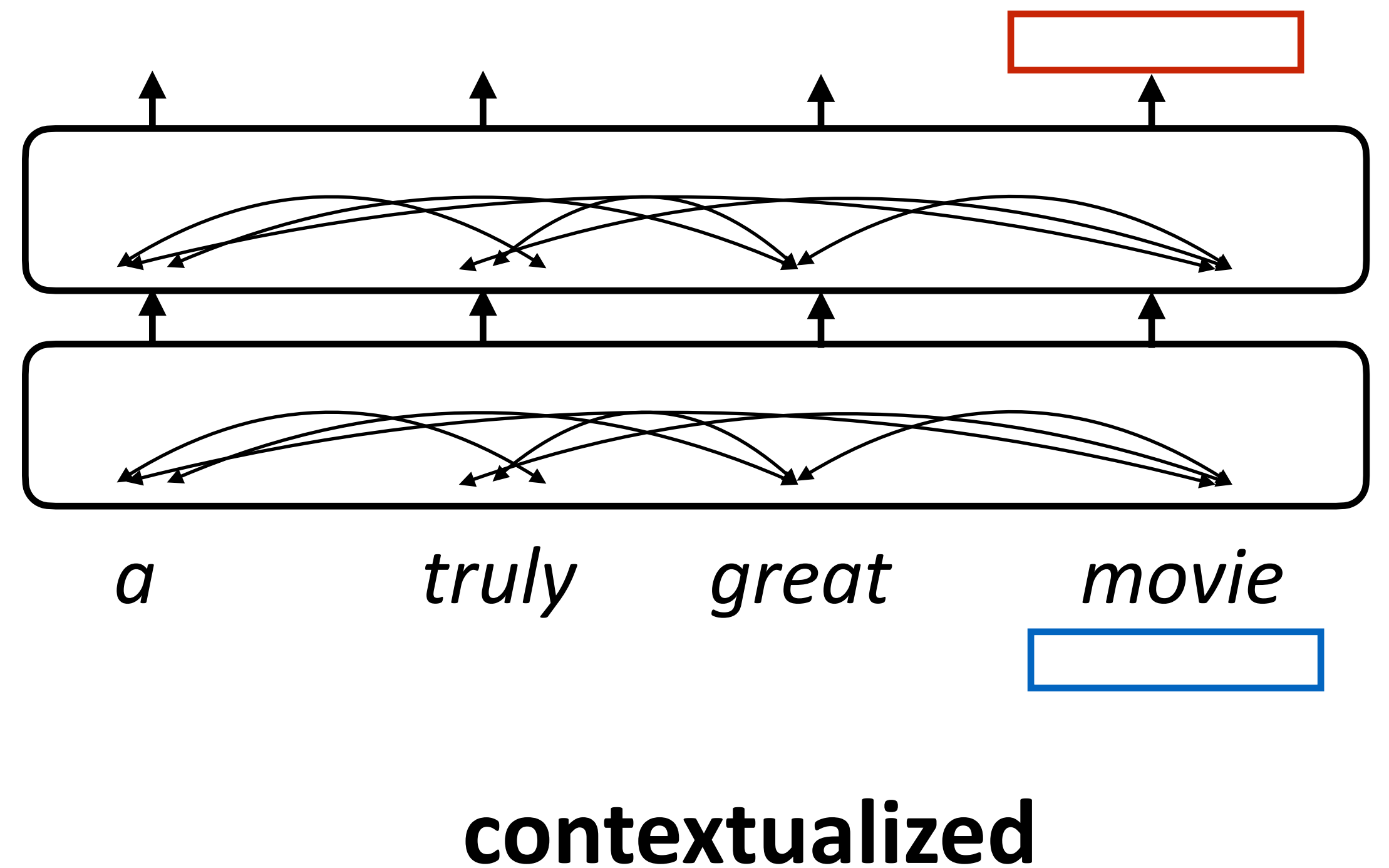
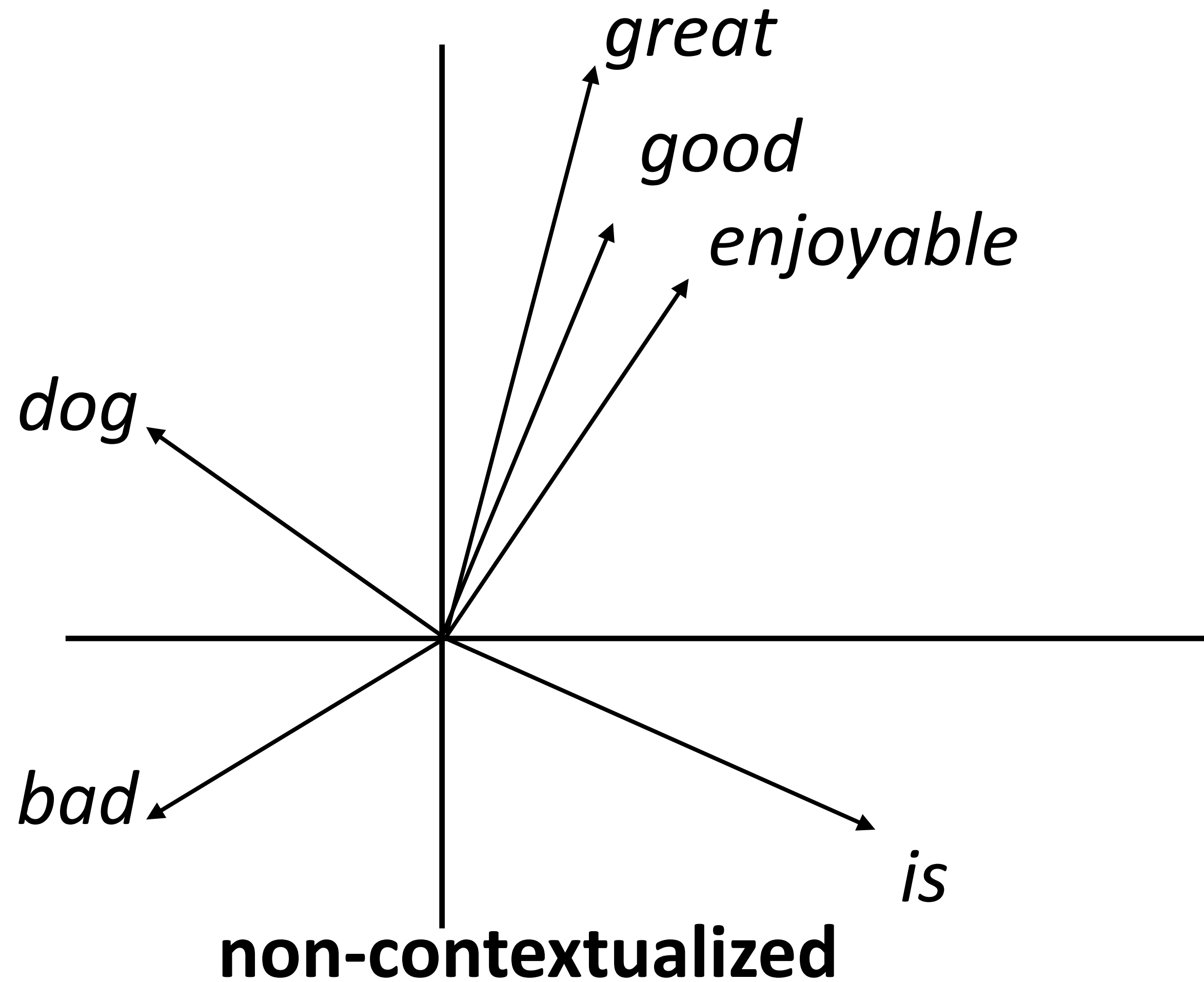
Thinking with Images

Vision-Language-Action

# Language Grounding

# Language Grounding

- ▶ How do we represent language in our models?
- ▶ How did we learn these representations? What do the vectors “mean”?



# Language Grounding

---

- ▶ Harnad defines a “symbol system”: we have symbols (e.g., strings) manipulated on the basis of rules, and these symbols ultimately have “semantic interpretation”
  - ▶ “Fodor (1980) and Pylyshyn (1980, 1984)...emphasize that the symbolic level (for them, the mental level) is a natural functional level of its own, with ruleful regularities that are independent of their specific physical realizations”
- ▶ Harnad challenges the idea that fully symbolic approaches can work well.
- ▶ Argues that “horse” is something that should be understood bottom-up through grounding. “Zebra” = “horse” + “stripes” could emerge this way, but he claims it cannot through a top-down symbolic system
- ▶ What does it mean to “understand” the symbols that get manipulated?

# Searle's Chinese Room

---

- ▶ Suppose we have someone in a room with a long list of rules, dictionaries, etc. for how to translate Chinese into English. A Chinese string is passed into the room and an English string comes out. The person is not a speaker of Chinese, but merely follows the rules and looks things up in the dictionaries to produce the translation.
- ▶ Does the person understand Chinese? Does the room? (the “system”?)
- ▶ Searle argues that (a) the room is like an AI system producing Chinese translations; (b) the operator in the room (the AI) does not “understand” Chinese. Harnad summarizes :

*The interpretation will not be intrinsic to the symbol system itself: It will be parasitic on the fact that the symbols have meaning for us, in exactly the same way that the meanings of the symbols in a book are not intrinsic, but derive from the meanings in our heads.*

# Language Grounding

- ▶ Bender and Koller separate form and meaning. Meaning = communicative intent. The role of the speaker/listener are crucial in language, LMs lack the underlying intent
- ▶ They propose the “octopus” experiment to show how form alone can fail. An octopus is eavesdropping on a conversation between A and B (using deep-sea communication cables). Suddenly, the octopus decides to cut the cable and impersonate B.
- ▶ A has an emergency and asks how to construct something with sticks to fend off a bear. The octopus can't help because it can't simulate this novel situation.



# Counterarguments

---

- ▶ We can't necessarily learn semantics from predicting next characters alone without execution. Consider training on:  

```
x = 2  
y = x + 2  
print(y)
```
- ▶ **However**, assertion statements are sufficient to teach us some semantics! (but this can still break down)  

```
x = 2  
y = x + 2  
assert(y == 4)
```
- ▶ For language: similar argument. Assume people say true things. Consider saying a pair of sentences  $x_1, x_2$ ; given enough examples, the fact that  $x_2$  should not be contradicted by  $x_1$  tells us something

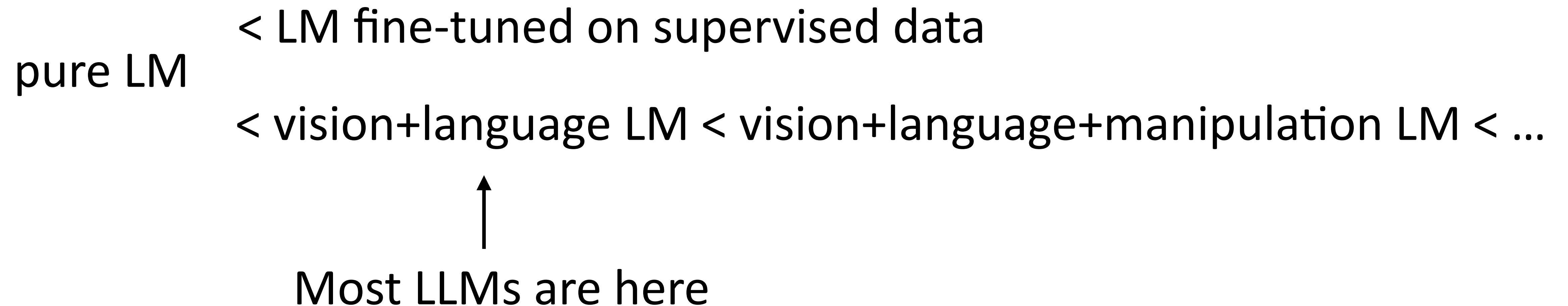
Merrill et al. (2021) *Provable Limitations of Acquiring Meaning from Ungrounded Form*

Merrill et al. (2022) *Entailment Semantics can be Extracted from an Ideal Language Model*

# Where are we?

---

- ▶ Lots of philosophy about these models!
- ▶ Nevertheless, it seems there's a hierarchy in terms of their understanding:



Administrative details and recap

Language Grounding

Grounding in Images

Visual Reasoning

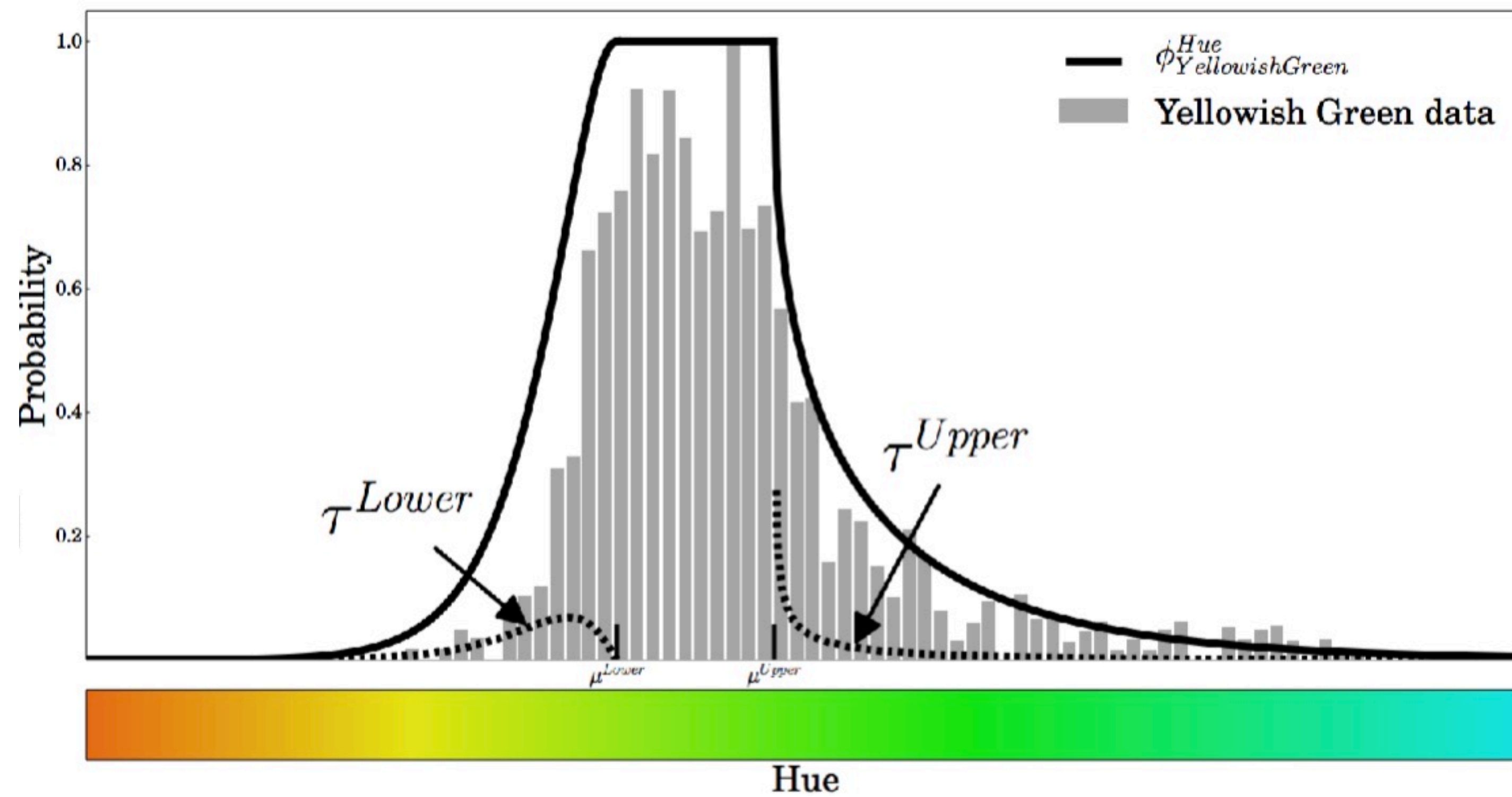
Thinking with Images

Vision-Language-Action

# Grounding in Images

# Language Grounding

- ▶ What does “yellowish green” mean?
- ▶ Formal semantics: yellowish green is a predicate. Things are either yellowish green or not. No connection to real color
- ▶ Grounding in perceptual space:



McMahan and Stone (2015)

# Grounding in Images

- ▶ How would you describe this image?
- ▶ What does the word “*spoon*” evoke?




*the girl is licking the spoon of batter*

# Grounding Spoon




Winco 0005-03 7  
3/8" Dinner Spoon...  
**\$7.16**



 wikiHow  
How to Hold a Spoon: 13 Steps (...)



 Indiegogo  
Spoon that Elevates Taste ...

# Grounding Language in Images

---

- ▶ Syntactic categories have some regular correspondences to the world:
  - ▶ Nouns: objects
  - ▶ Verbs: actions
  - ▶ Sentences: whole scenes or things happening
- ▶ Tasks:
  - ▶ Object recognition (pick out one most salient object or detect all of them)
  - ▶ Image captioning: produce a whole sentence for an image

# Language-vision Models



Image encoder  
(CNN, Transformer)

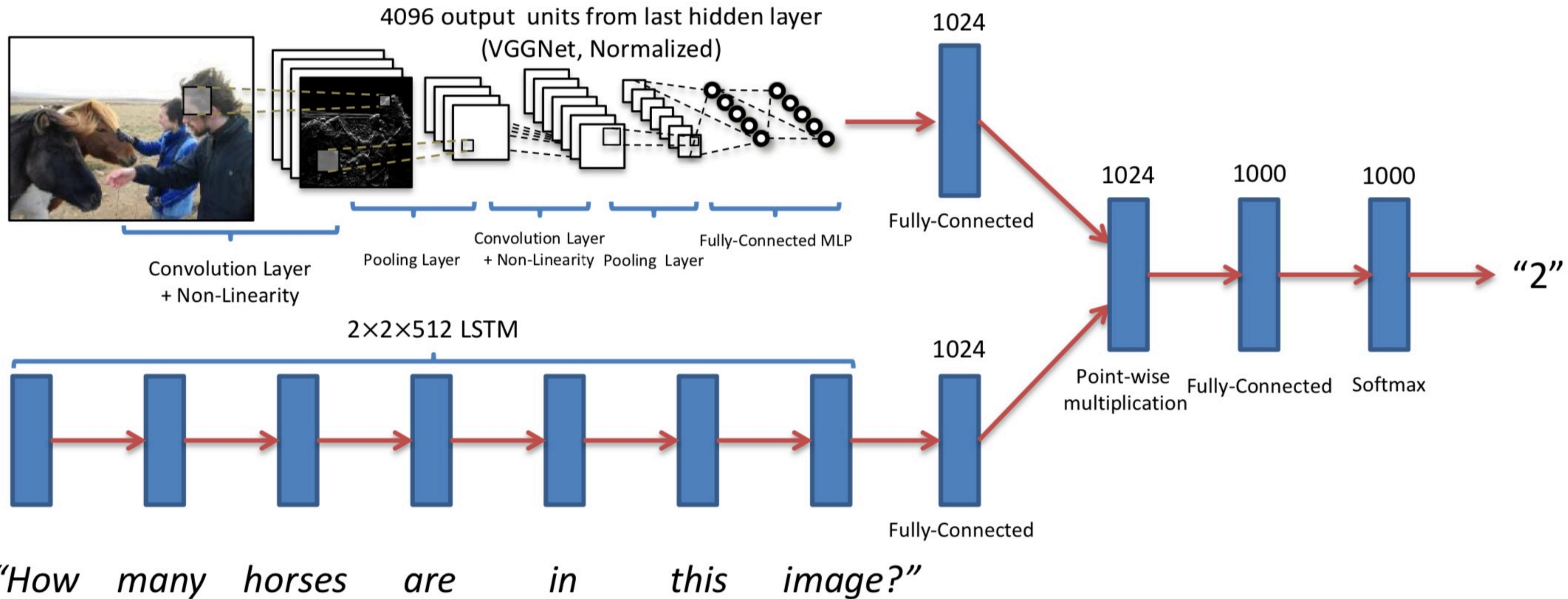
*the girl is licking the  
spoon of batter*

Language encoder  
(LSTM, Transformer)

Cross-attention/joint layer

Prediction

# Visual Question Answering



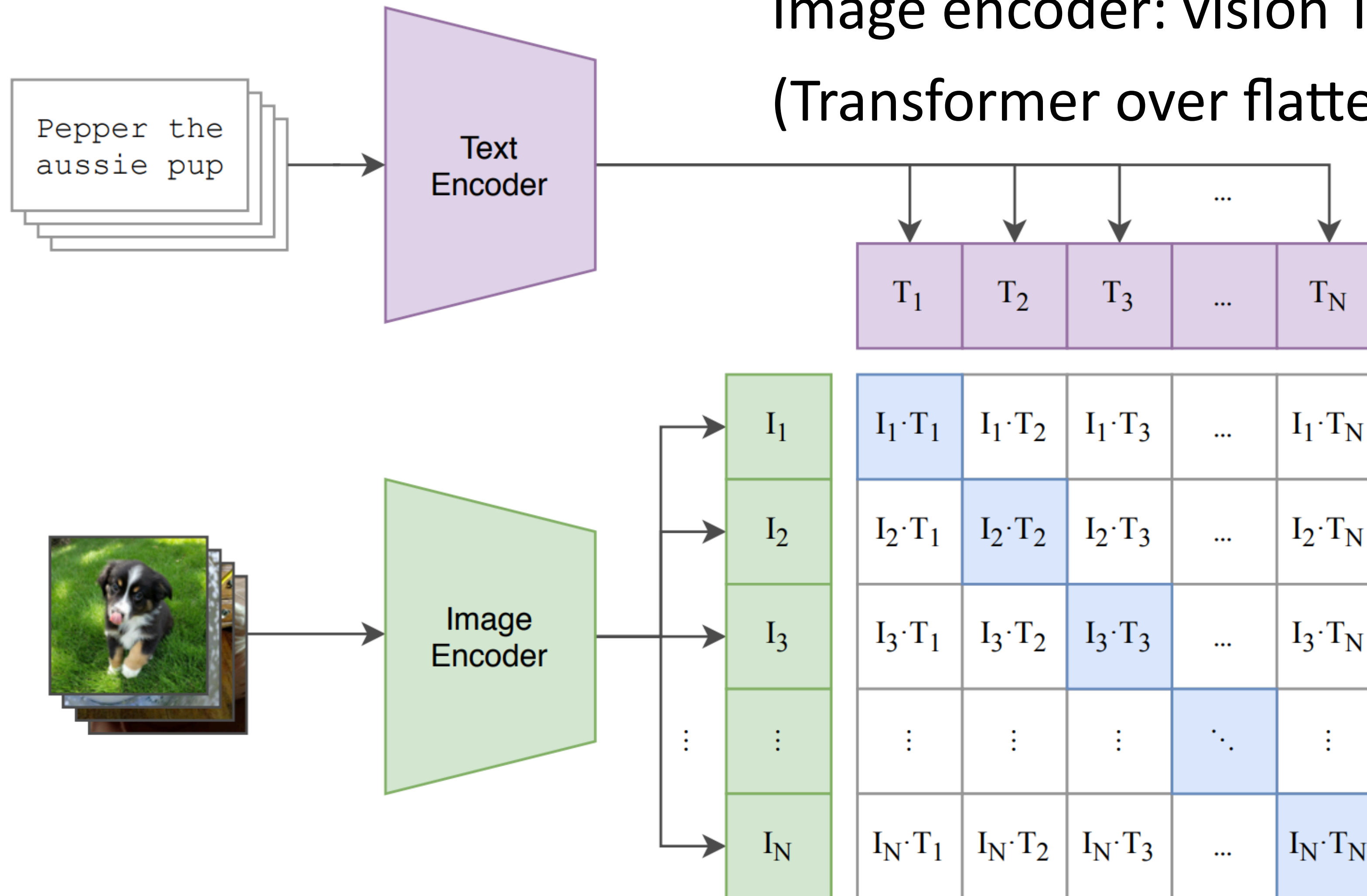
# Language-vision Pre-training

(1) Contrastive pre-training

Text encoder: Transformer

Image encoder: vision Transformer

(Transformer over flattened patches)



Radford et al., 2021

# Language-vision Pre-training

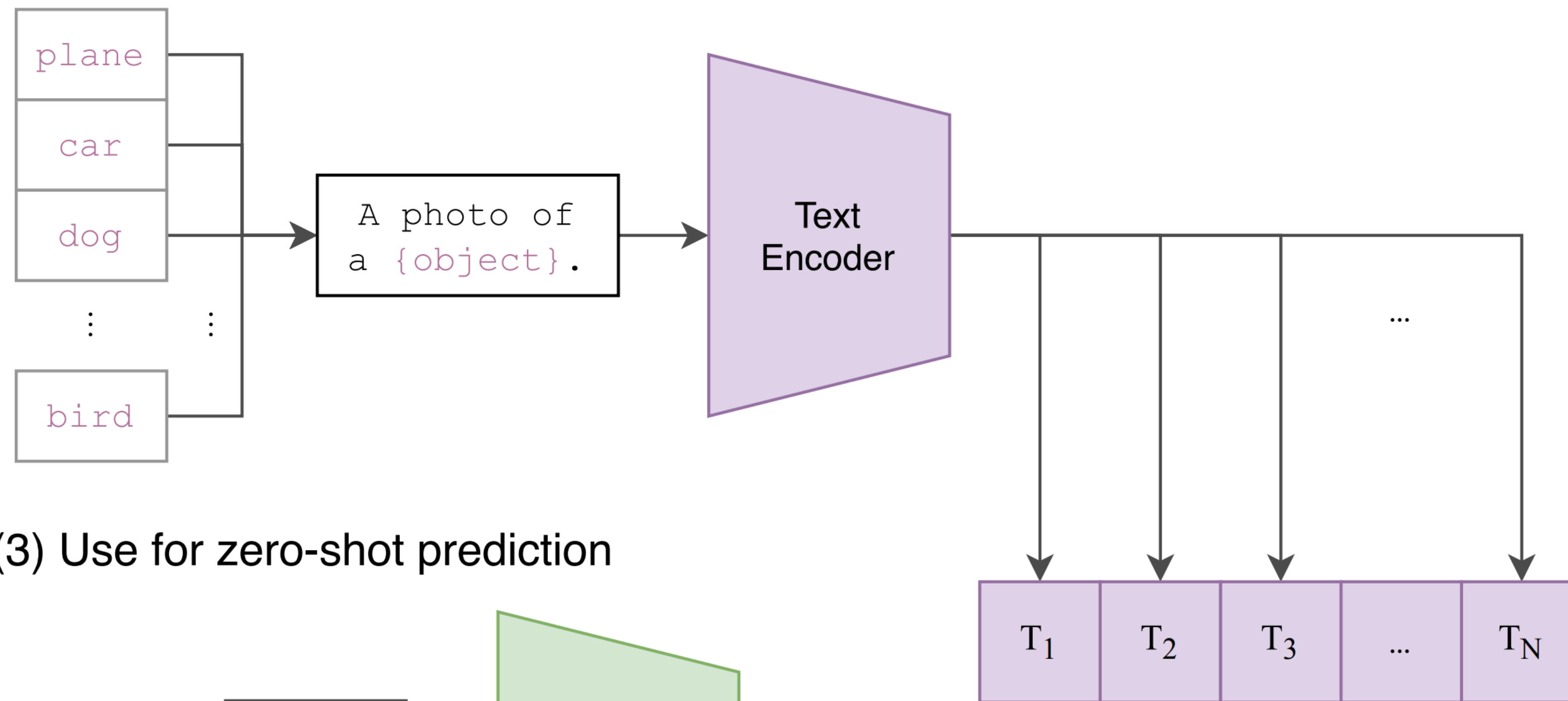
	$T_1$	$T_2$	$T_3$	...	$T_N$
$I_1$	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
$I_2$	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
$I_3$	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$I_N$	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

- ▶ Contrastive objective: each image should be more similar to its correspond caption than to other captions

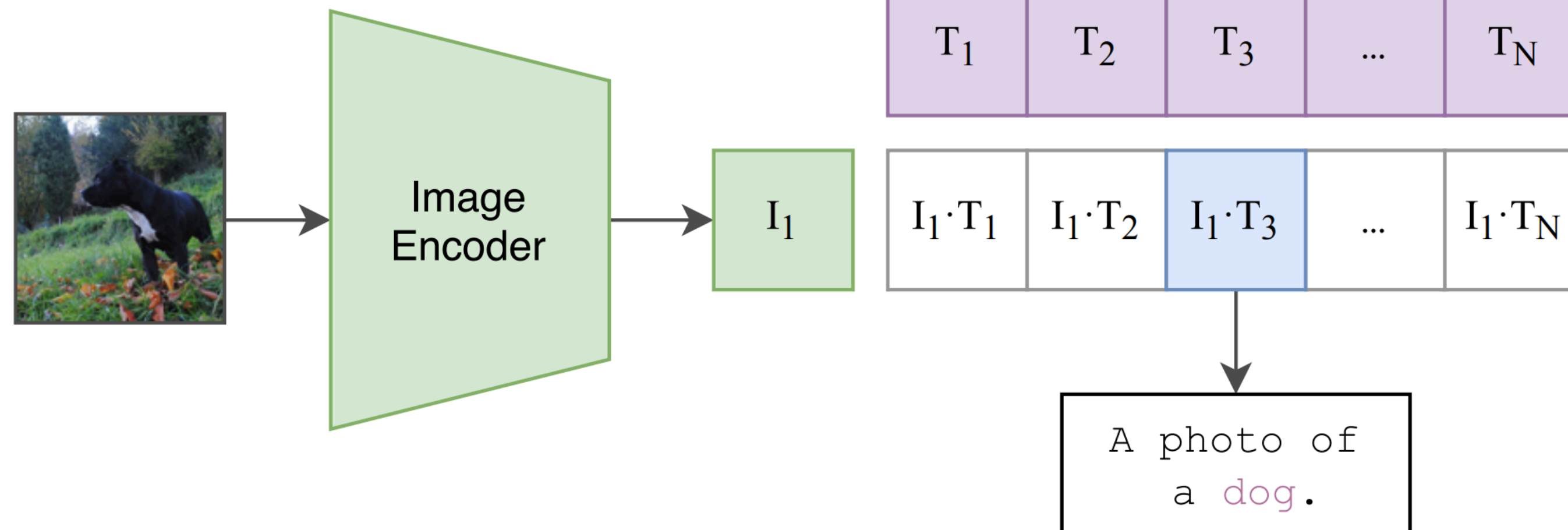
$$\begin{aligned} & \text{maximize } \text{softmax}(I_1^T T_i)[1] \\ & \quad + \text{softmax}(I_2^T T_i)[2] \\ & \quad \quad \quad + \dots \end{aligned}$$

# Language-vision Pre-training

(2) Create dataset classifier from label text



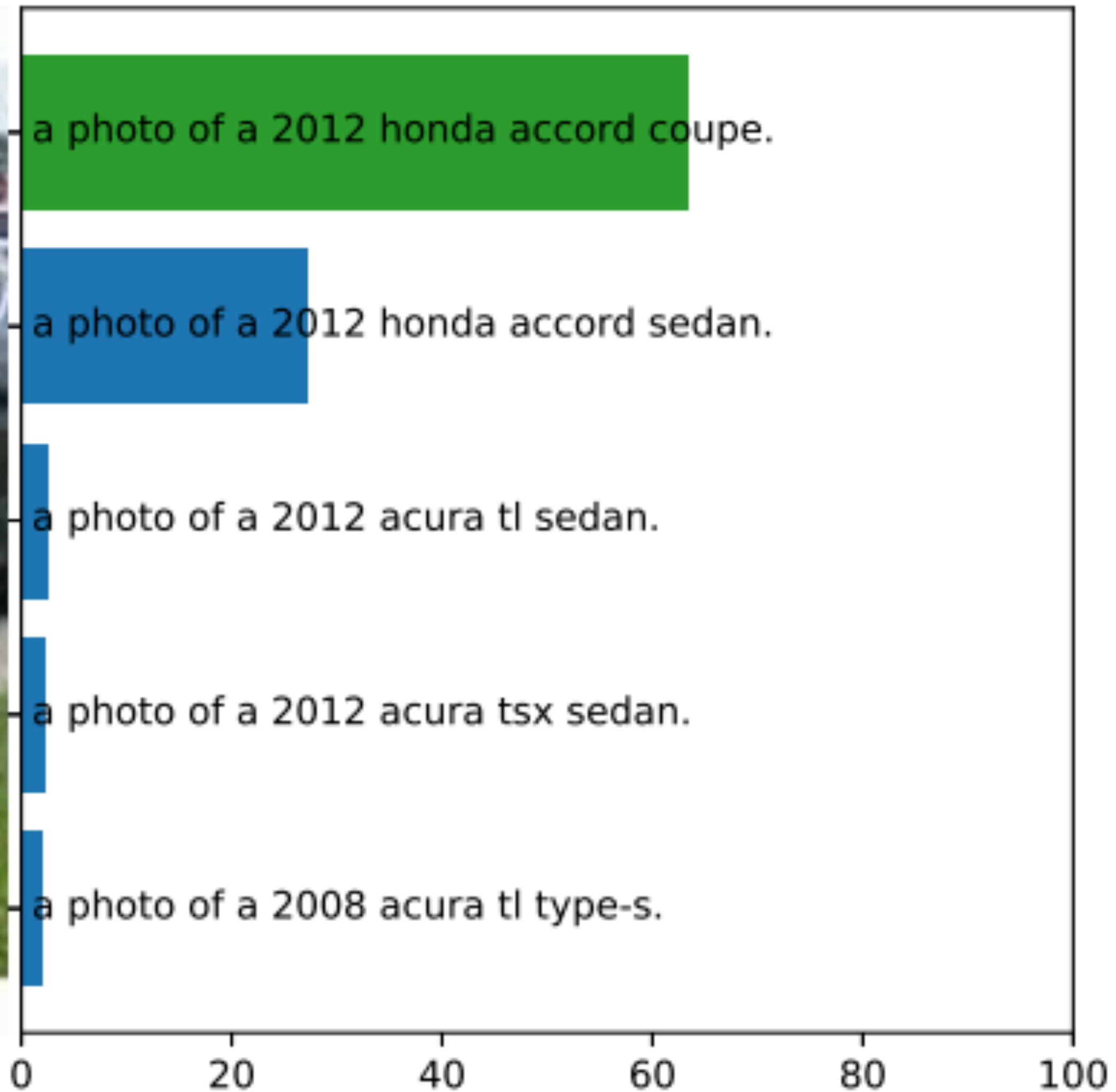
(3) Use for zero-shot prediction



# CLIP: Zero-shot Results

## Stanford Cars

correct label: 2012 Honda Accord Coupe    correct rank: 1/196    correct probability: 63.30%



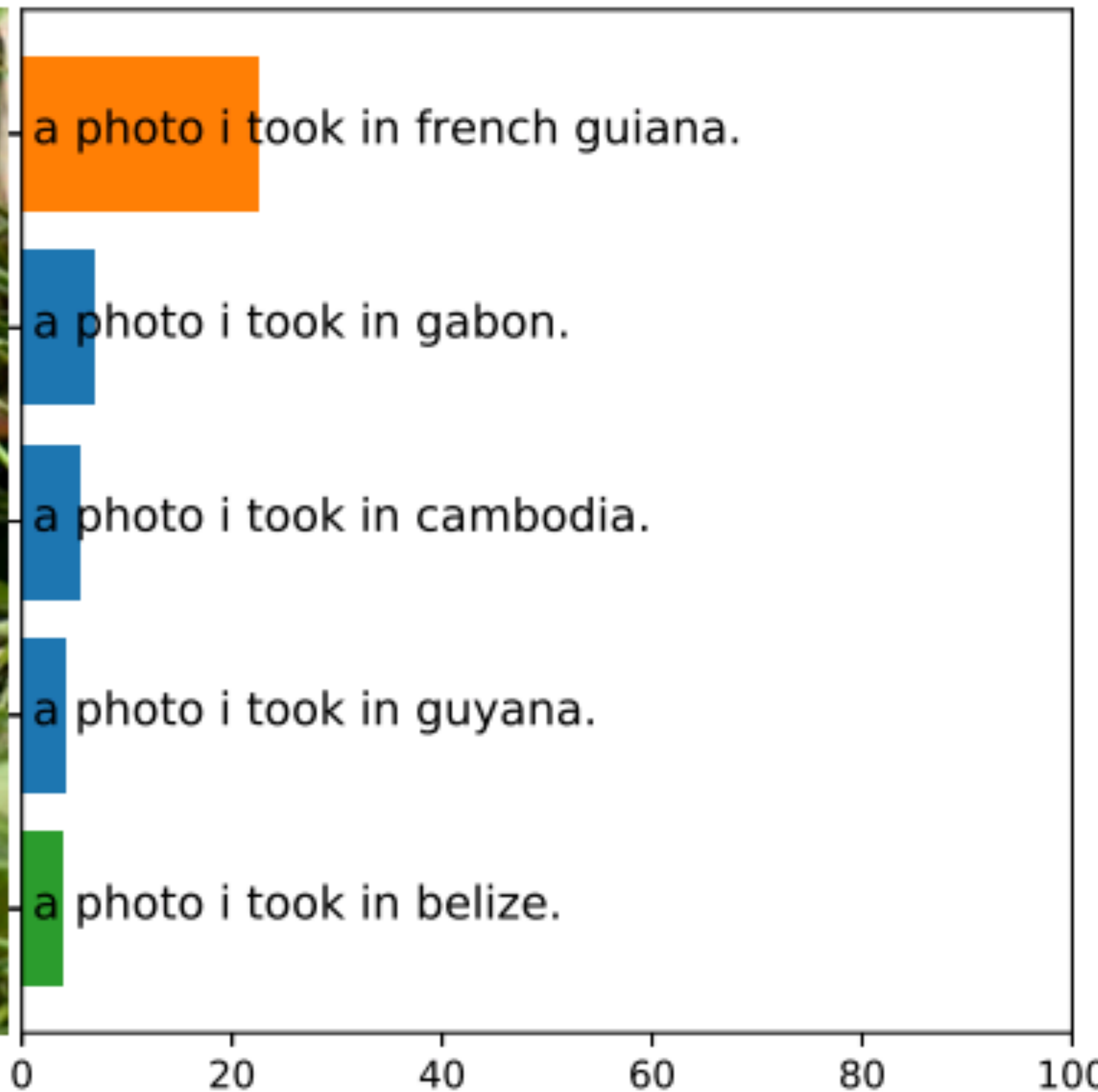
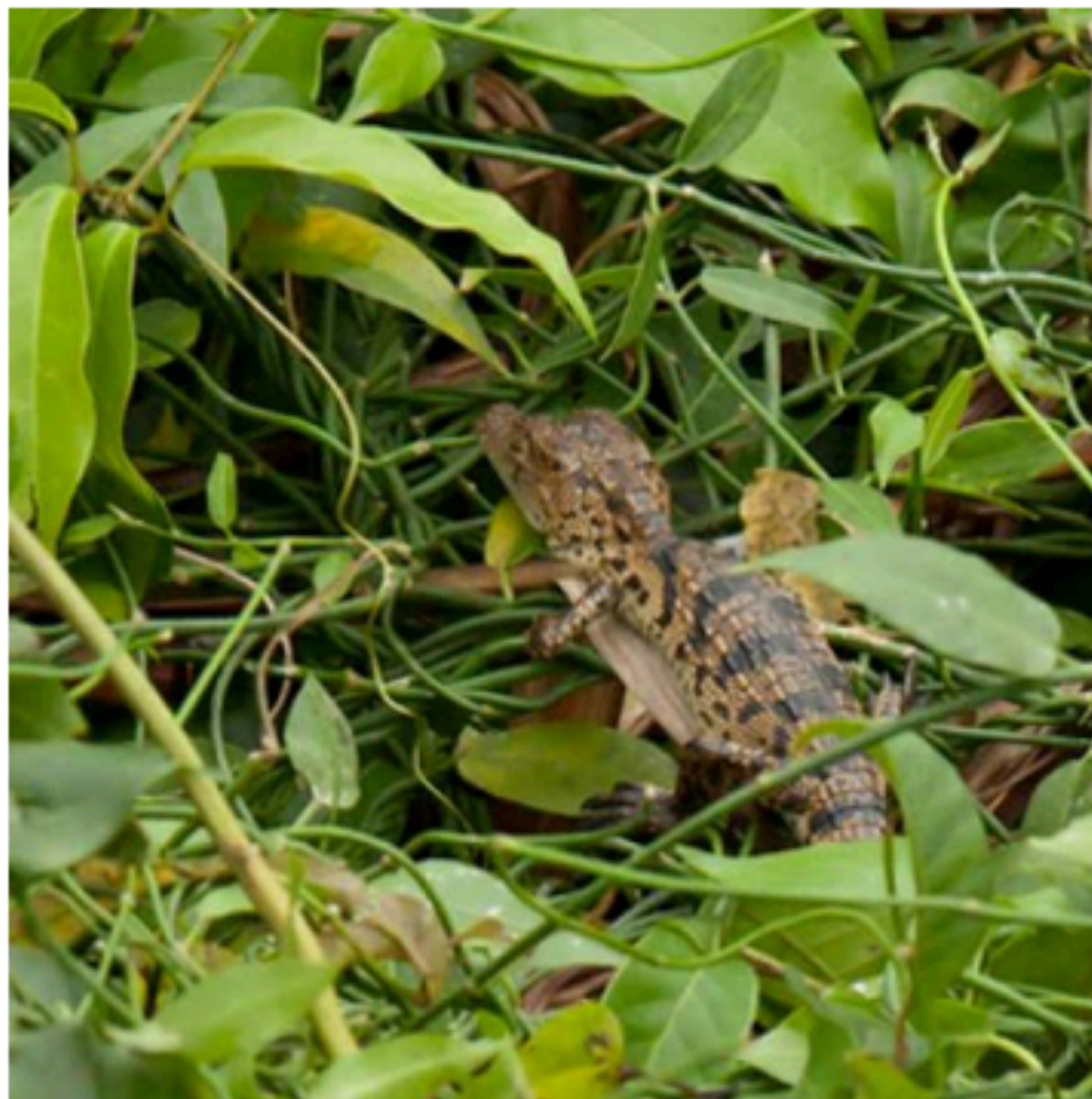
# CLIP: Zero-shot Results

## Country211

correct label: Belize

correct rank: 5/211

correct probability: 3.92%



Administrative details and recap

Language Grounding

Grounding in Images

Visual Reasoning

Thinking with Images

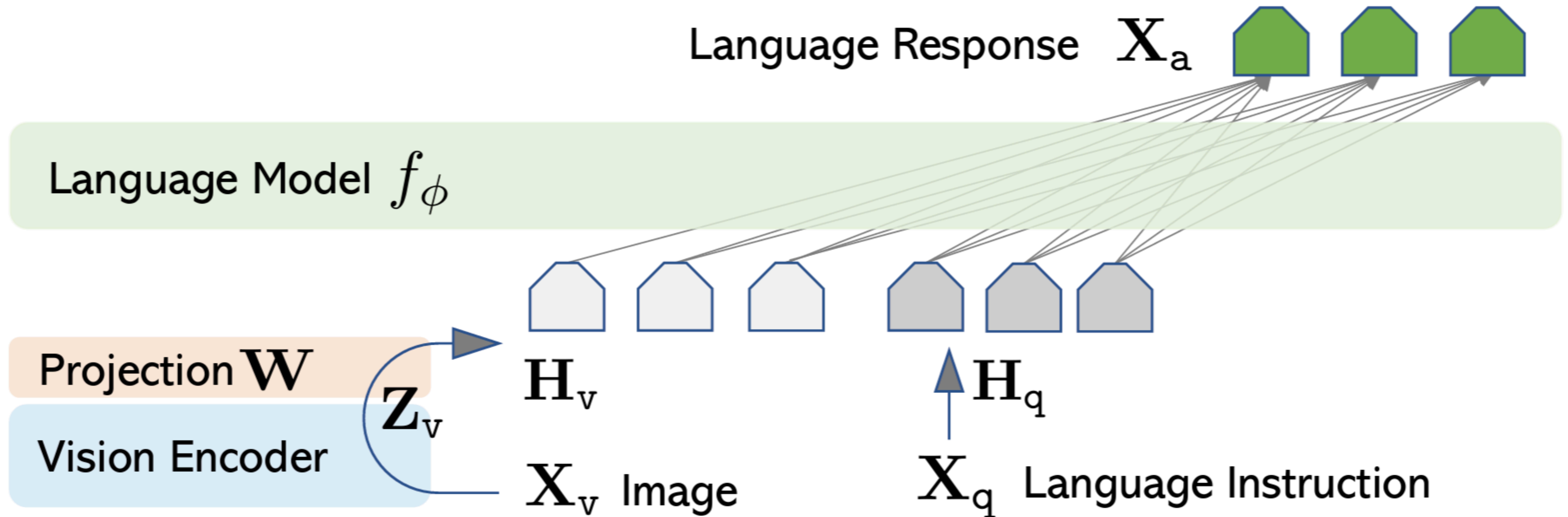
Vision-Language-Action

# Visual Reasoning

# CLIP lets us turn images into vectors

LLMs: words  $\rightarrow$  vectors, then predict the next word

If we can turn an image into a vector...can we use it in the same way as a word?



# Qwen-VL



```
1 #include <stdlib.h>
2
3 int solution(int *A, int A_length) {
4     int ans = 0;
5     for (int i = 1; i < A_length; i++) {
6         if (A[i] < ans) {
7             ans = A[i];
8         }
9     }
10    return ans;
11 }
12
```



The solution in the image is to find the minimum value in an array. And there is a bug in the function. Fix the bug.

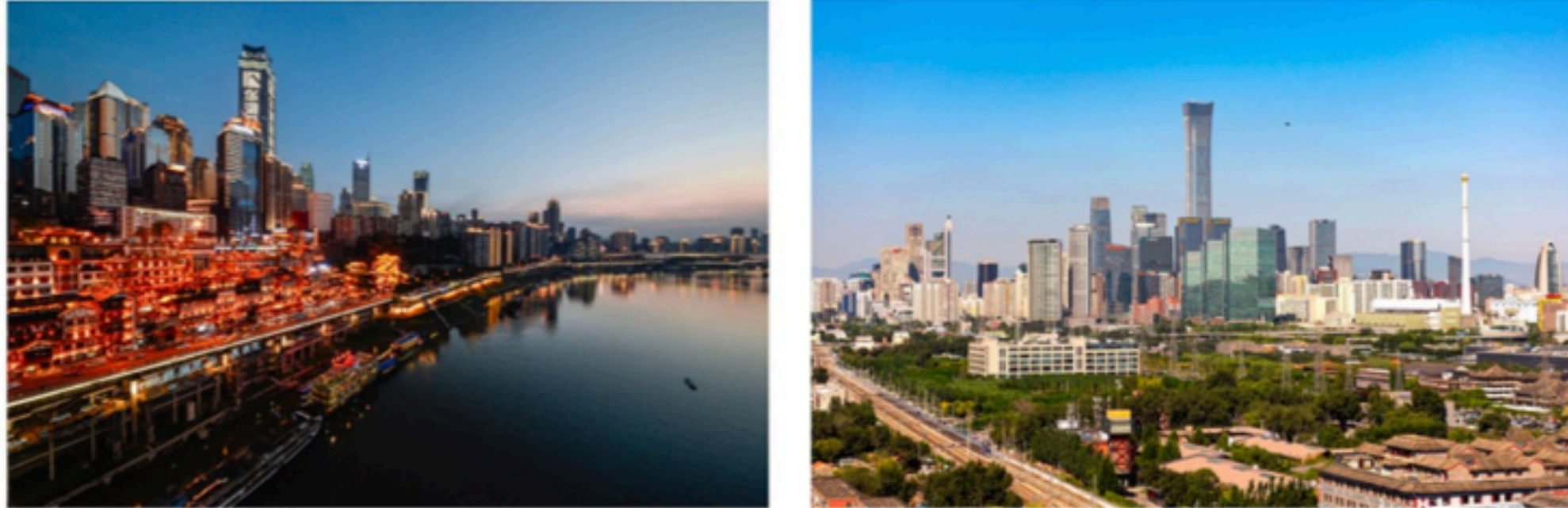
```
int solution(int A[], int n) {
    int ans = A[0];
    for (int i = 1; i < n; i++) {
        if (A[i] < ans)
            ans = a[i];
    }
    return ans;
}
```



Can you find spider man and Hulk?



# Qwen-VL



What are the two cities in the above pictures? Please compare them.

The first image is the city skyline of Chongqing, which reflects the hustle and bustle of a modern metropolis. The second image is the skyline of Beijing, symbolizing the modernization and internationalization of the Chinese capital. Both cities are important in China, with unique cultures and development histories.



## ABSTRACT



In this work, we introduce the Qwen-VL series, a set of large-scale vision-language models (LVLMs) designed to perceive and understand both texts and images. Starting from the Qwen-LM as a foundation, we assign it visual capacity by meticulously designed (i) visual receptor, (ii) input-output interface, (iii) 3-stage training pipeline, and (iv) multilingual multimodal cleaned corpus. Beyond the conventional description and question-answering, we inject the grounding ability into Qwen-VLs by importing fine-grained image-caption-box pairs. The resulting models, including Qwen-VL and Qwen-VL-Chat, set new records on a broad range of visual-centric benchmarks (e.g., image captioning, question answering, visual grounding) under different settings (e.g., zero-shot, few-shot). Moreover, on real-world dialog benchmarks, our instruction-tuned Qwen-VL-Chat also demonstrates conspicuous superiority compared to existing vision-language chatbots. All models will be made public to facilitate future research.



OCR this picture

## ABSTRACT

In this work, we introduce the Owen-VL series, a set of large-scale vision-language models (LVLMs) designed to perceive and understand both texts and images. Starting from the Owen-LM as a foundation, we assign it visual capacity by meticulously designed (i) visual receptor, (ii) input-output interface, (iii) 3-stage training pipeline, and (iv) multilingual multimodal cleaned corpus. Beyond the conventional description and question-answering, we inject the grounding ability into Owen-VLs by importing fine-grained image-caption-box pairs. The resulting models, including Owen-VL and Owen-VL-Chat, set new records on a broad range of visual-centric benchmarks (e.g., image captioning, question answering, visual grounding) under different settings (e.g., zero-shot, few-shot). Moreover, on real-world dialog benchmarks, our instruction-tuned Owen-VL-Chat also demonstrates conspicuous superiority compared to existing vision-language chatbots. All models will be made public to facilitate future research.



# Qwen-VL

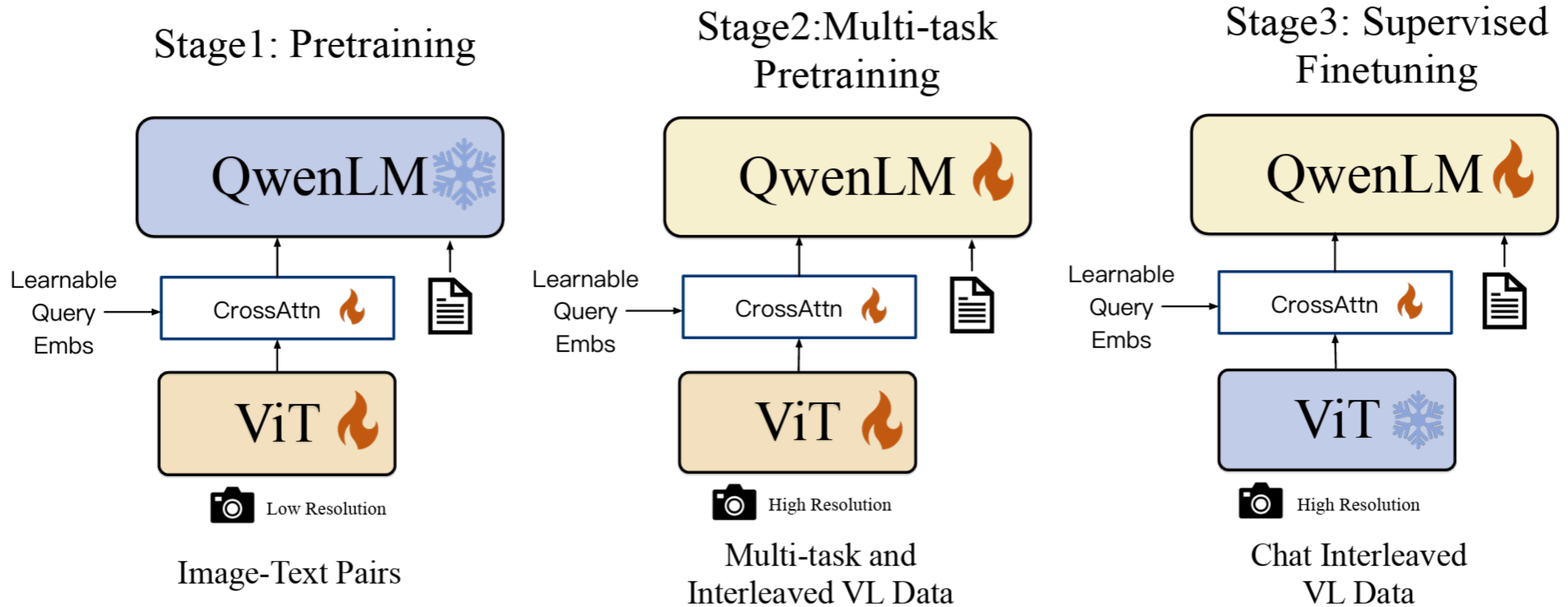


Table 1: Details of Qwen-VL model parameters.

Vision Encoder	VL Adapter	LLM	Total
1.9B	0.08B	7.7B	9.6B

# Qwen-VL

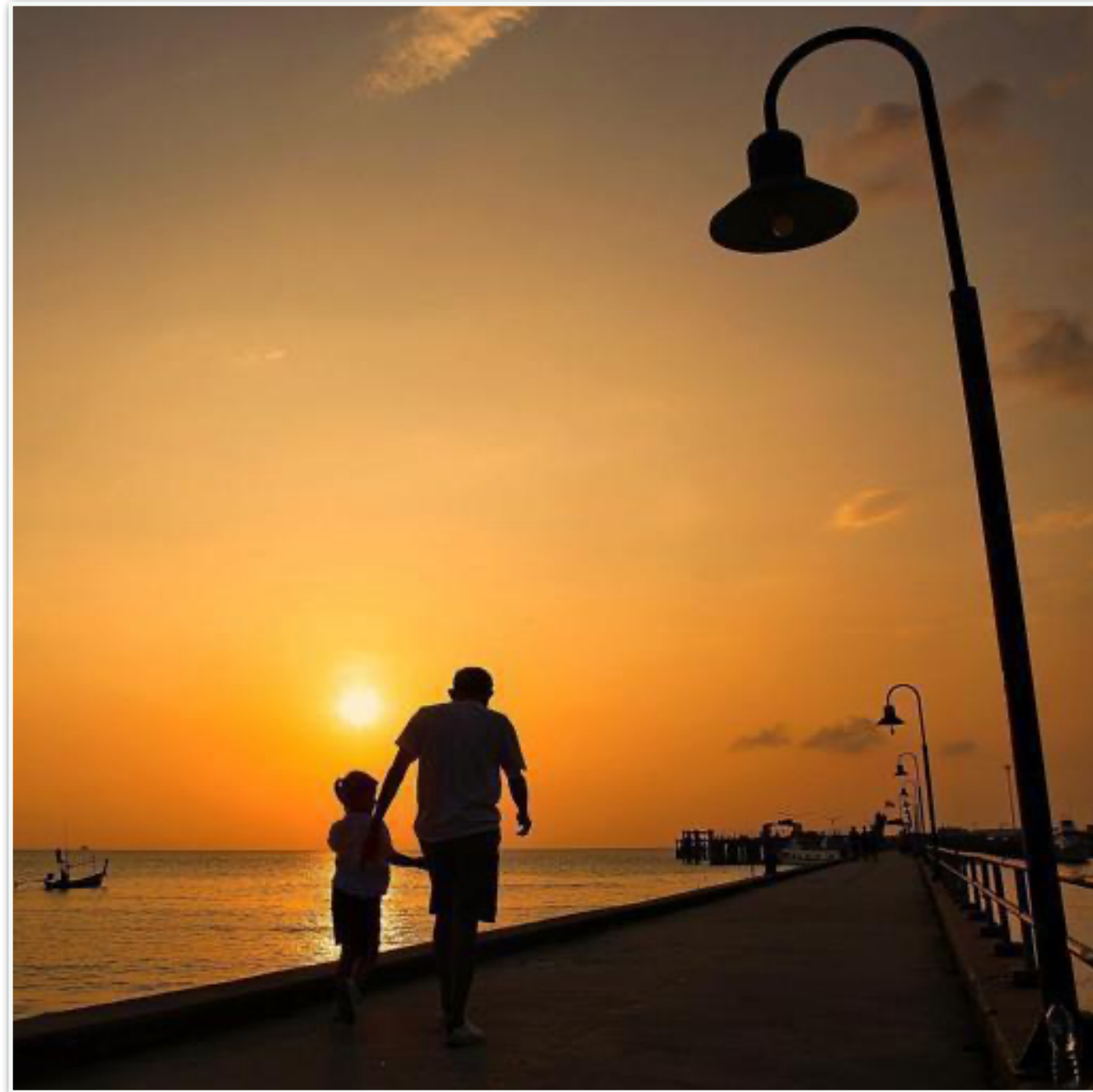
Table 3: Details of Qwen-VL multi-task pre-training data.

Task	# Samples	Dataset
Captioning	19.7M	LAION-en & zh, DataComp, Coyo, CC12M & 3M, SBU, COCO, In-house Data
VQA	3.6M	GQA, VGQA, VQAv2, DVQA, OCR-VQA, DocVQA, TextVQA, ChartQA, AI2D
Grounding <sup>2</sup>	3.5M	GRIT
Ref Grounding	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
Grounded Cap.	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
OCR	24.8M	SynthDoG-en & zh, Common Crawl pdf & HTML
Pure-text Autoregression	7.8M	In-house Data



# What can LVLMs do?

---



LVLM



“There are 2 people in the foreground of the image.”

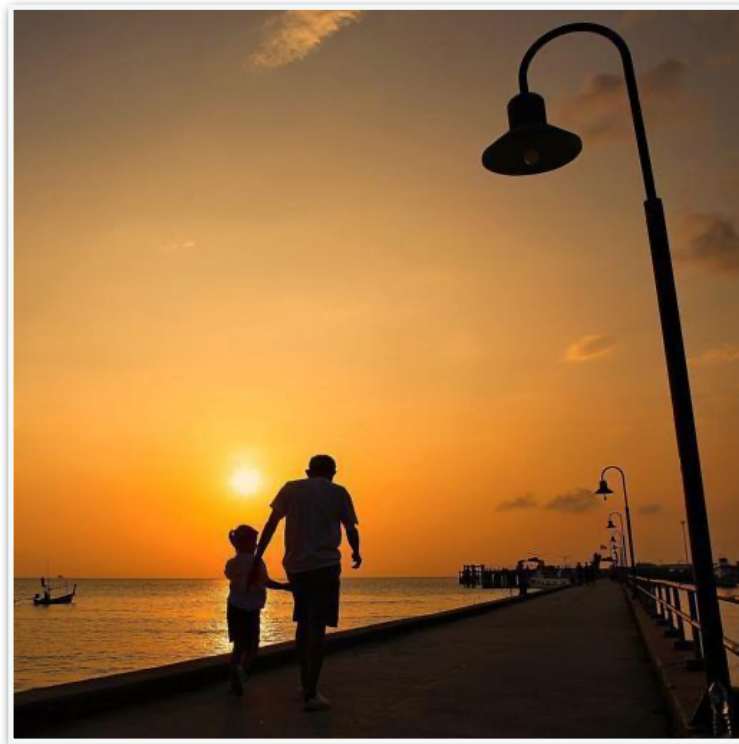
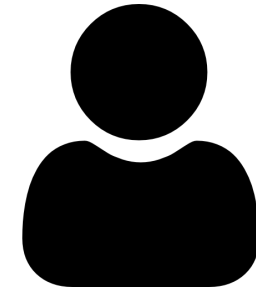
How many people are in the foreground of the image?



# LVLM Evaluations: Reasoning

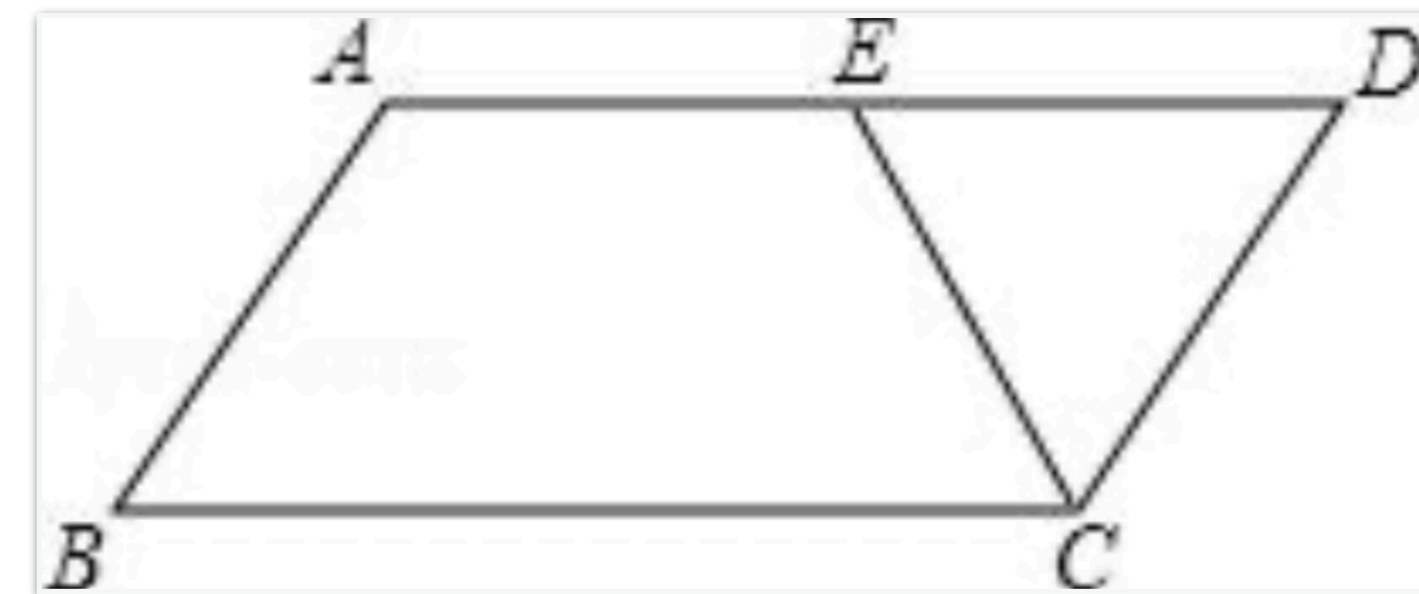
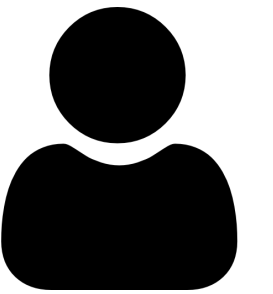
## Commonsense Reasoning

*Which mood does this image convey?  
A: Sad, B: Cozy, C: Happy, D: Angry*



## Math Reasoning

*As shown in the figure, in the parallelogram  $ABCD$ ,  $CE$  bisects  $\angle BCD$  and it intersects the  $AD$  edge at point  $E$ , and  $DE = 3.0$ , then the length of  $AB$  is ()*

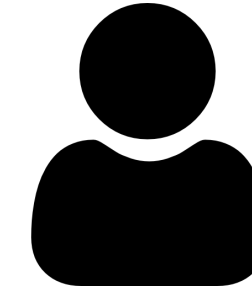


Example from Math-Vista



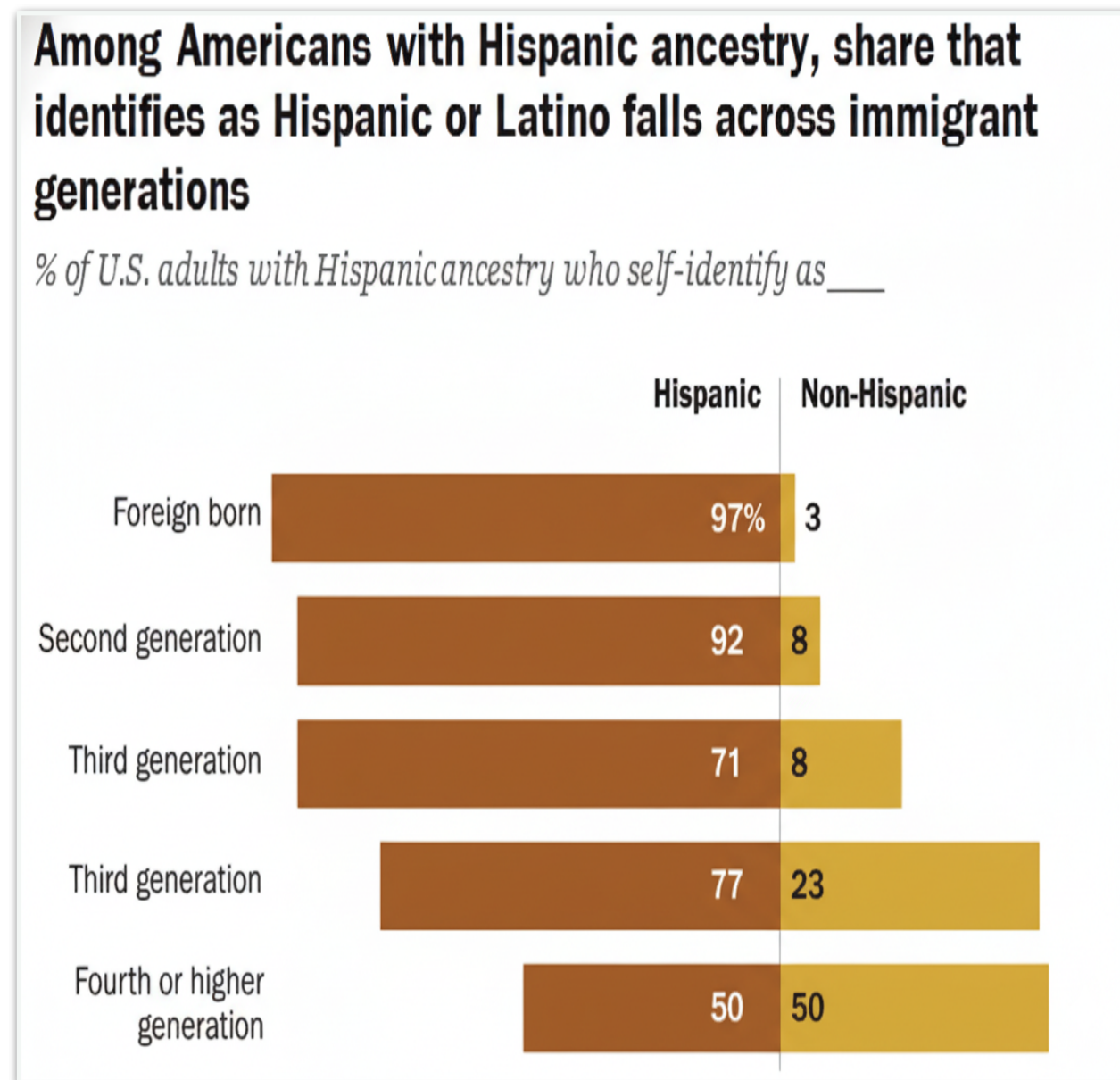
# Textual Reasoning vs Visual Reasoning

Q: What's the value of the smallest bar in the graph?



Many questions can be answered by a pipeline:

1. Extract text *explicitly* in the figure
2. Answer the question with a text LLM.



Example from ChartQA

LVLM



LVLM: "Display this information in text"



Extracted Text

Foreign born:

- \* Hispanic (97%)
- \* Non-Hispanic (3%)

Second generation:

- \* Hispanic (92%)
- \* Non-Hispanic (8%)

.....

LLM: "What's the value of the smallest bar in the graph?"



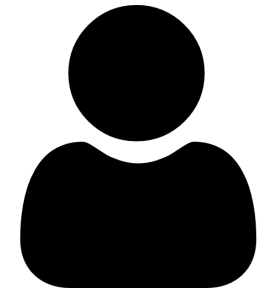
This example is very easy and doesn't really require *visual reasoning*.



# Prior Chart Benchmarks Over-Represent Textual Reasoning

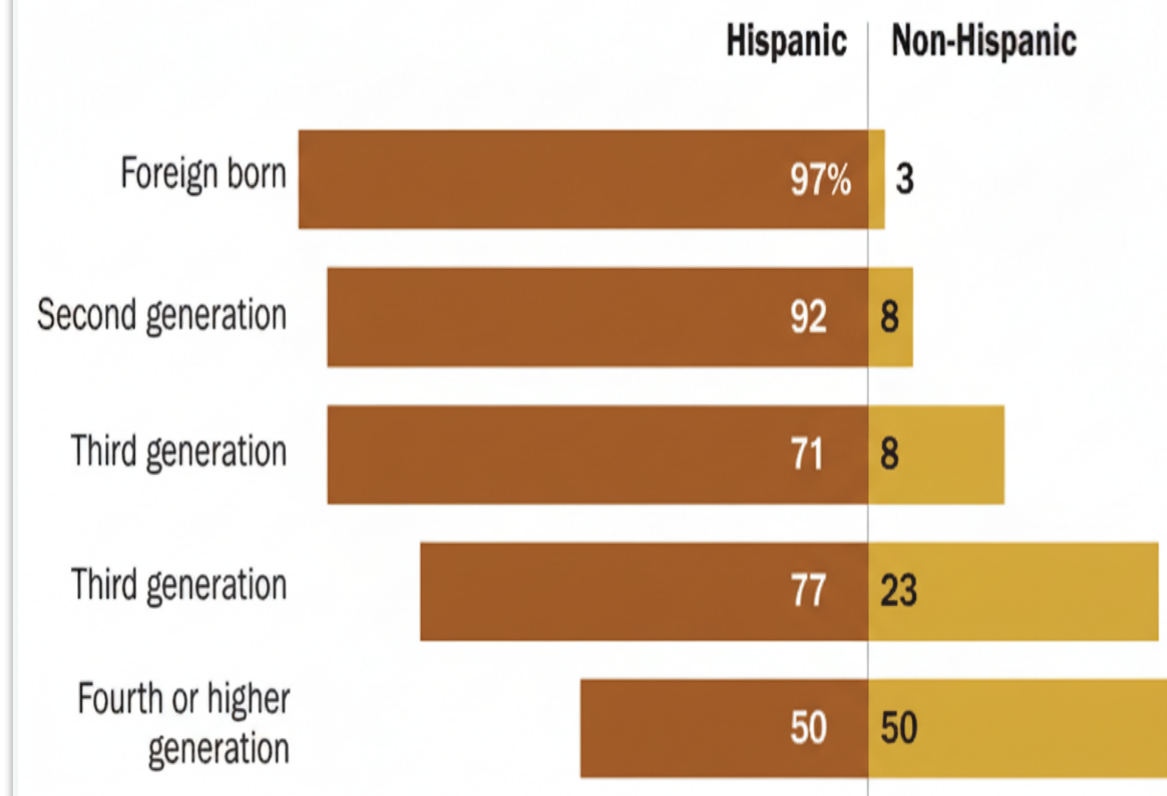
## Textual Reasoning

Q: What's the value of the smallest bar in the graph?



Among Americans with Hispanic ancestry, share that identifies as Hispanic or Latino falls across immigrant generations

% of U.S. adults with Hispanic ancestry who self-identify as \_\_\_\_



Example from ChartQA

### Extracted Text

#### Foreign born:

\* Hispanic (97%)

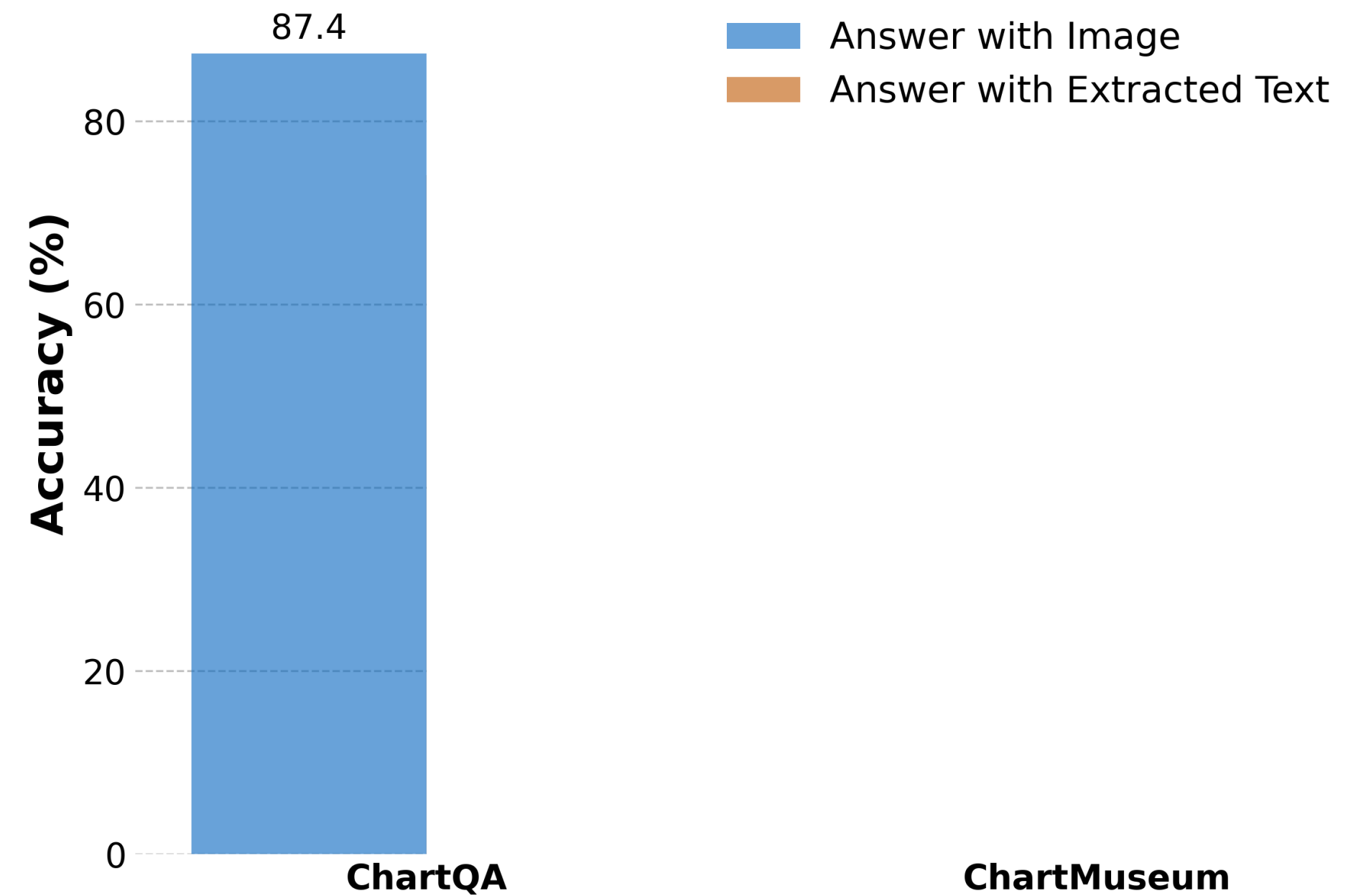
\* Non-Hispanic (3%)

#### Second generation:

\* Hispanic (92%)

\* Non-Hispanic (8%)

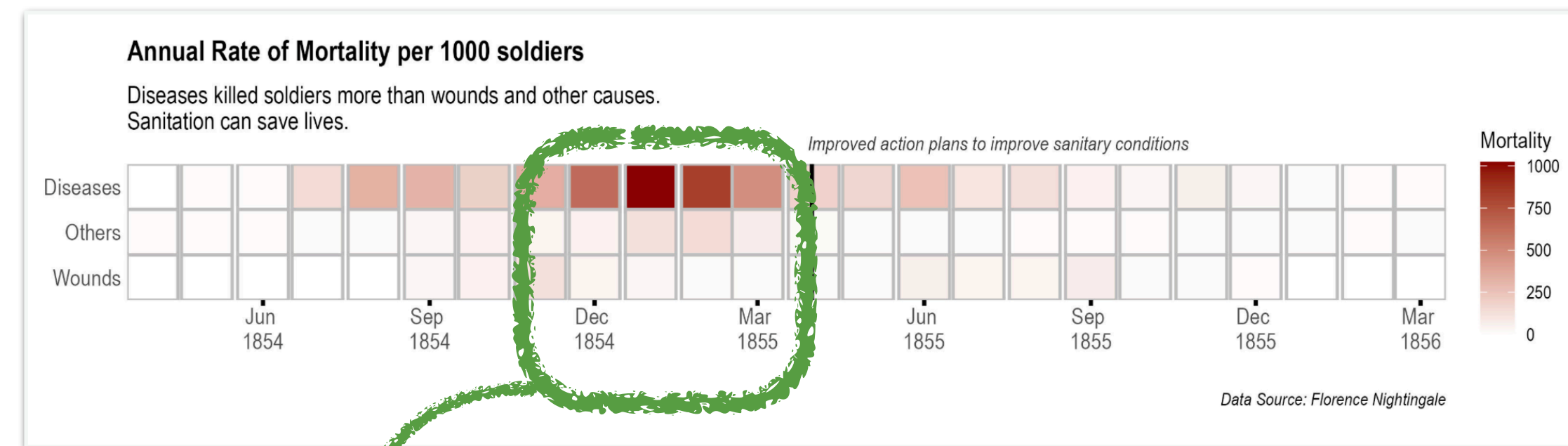
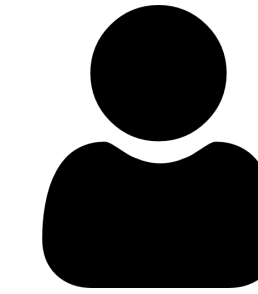
.....



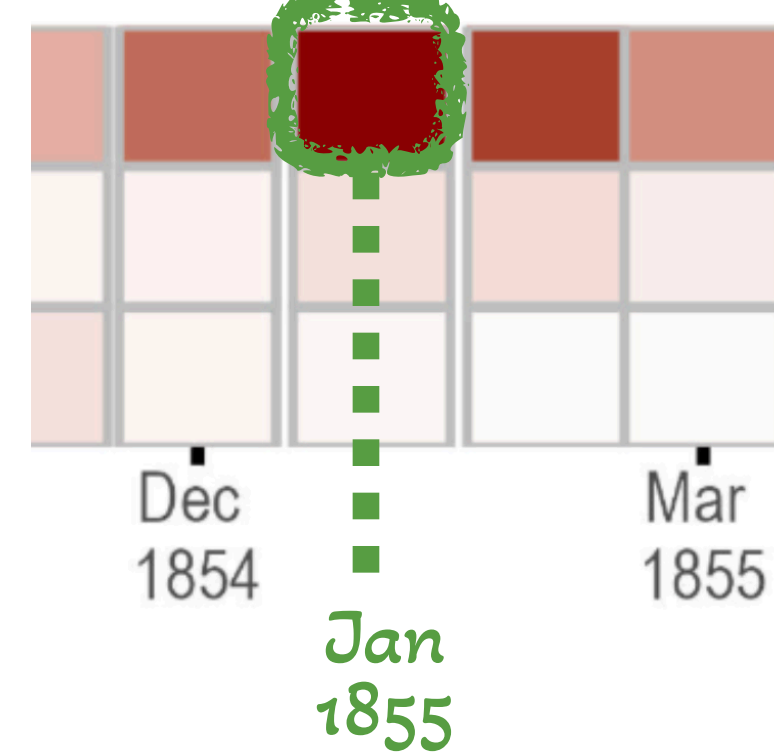


# Textual Reasoning vs Visual Reasoning

*Q: Which month-year saw the highest rate of mortality due to diseases?*



Example from ChartMuseum



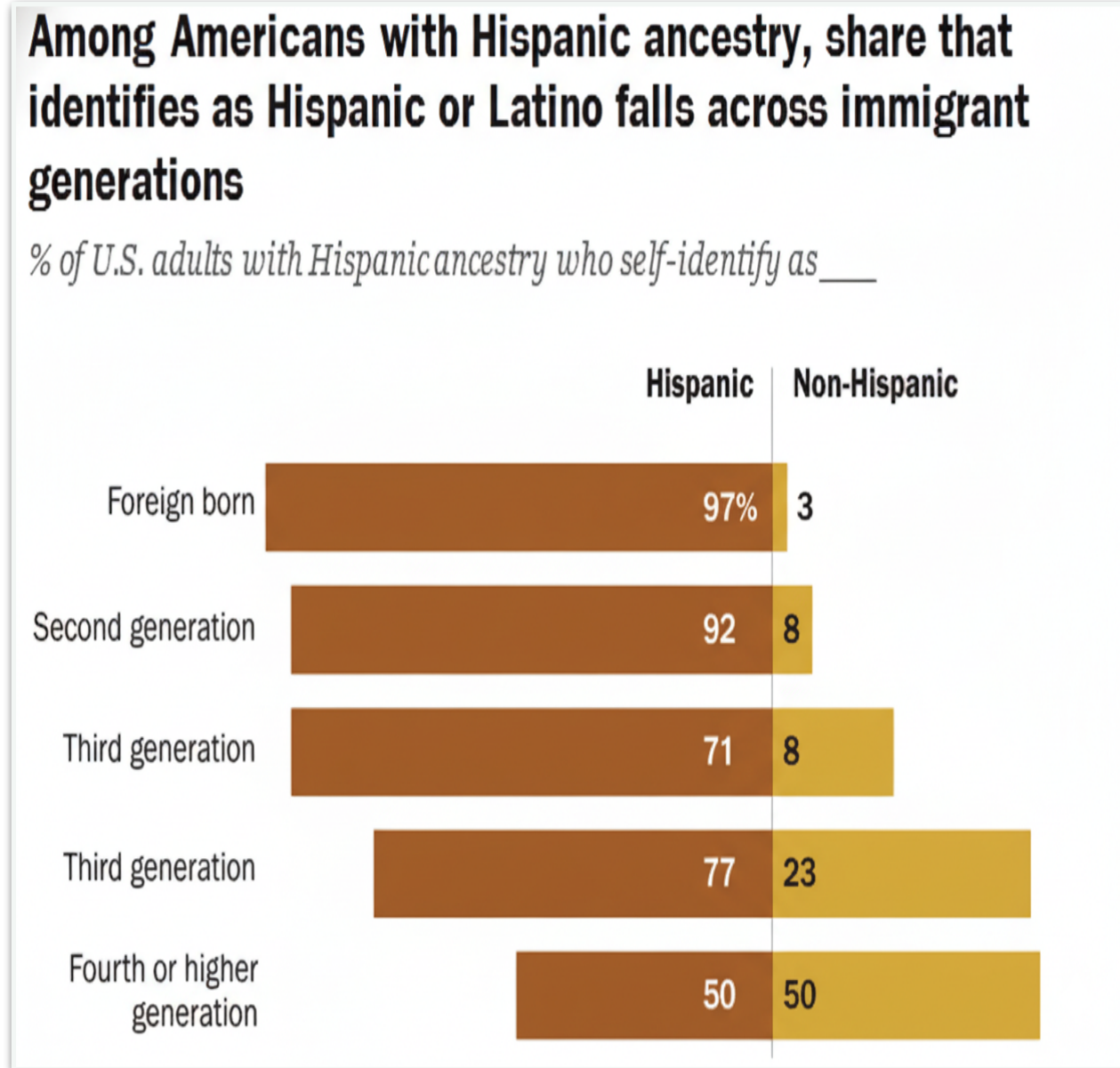
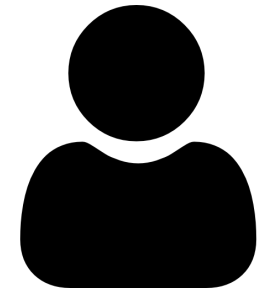
**This visual information is not easily expressed in text without visual reasoning!**



# Textual Reasoning vs Visual Reasoning

## Textual Reasoning

Q: What's the value of the smallest bar in the graph?



Example from ChartQA

### Extracted Text

#### Foreign born:

- \* Hispanic (97%)
- \* Non-Hispanic (3%)

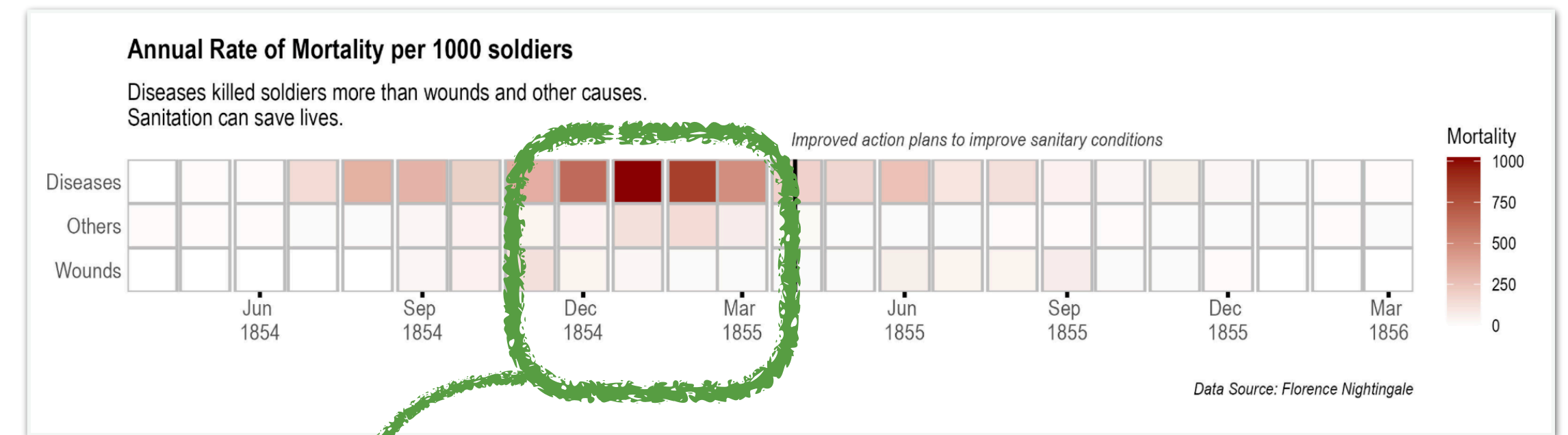
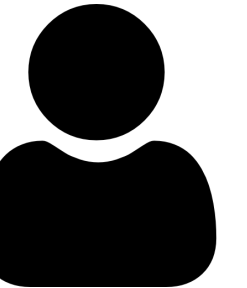
#### Second generation:

- \* Hispanic (92%)
  - \* Non-Hispanic (8%)
- .....

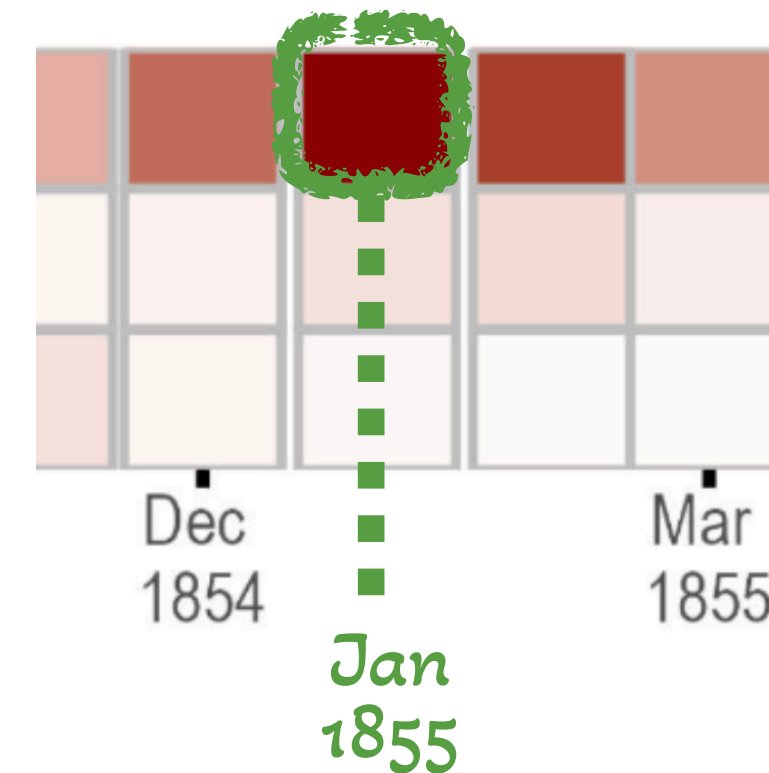
**No visual inference is required!**

## Visual Reasoning

Q: Which month-year saw the highest rate of mortality due to diseases?



Example from ChartMuseum



**This visual information is not easily expressed in text without visual reasoning!**



# ChartMuseum Benchmark



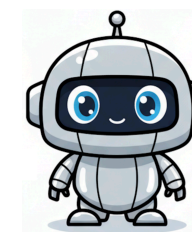
1162 questions by CS researchers, with 184 diverse image sources



Questions are created *entirely by humans* with no LLM involvement



Each question targets either textual, visual, or both types of reasoning

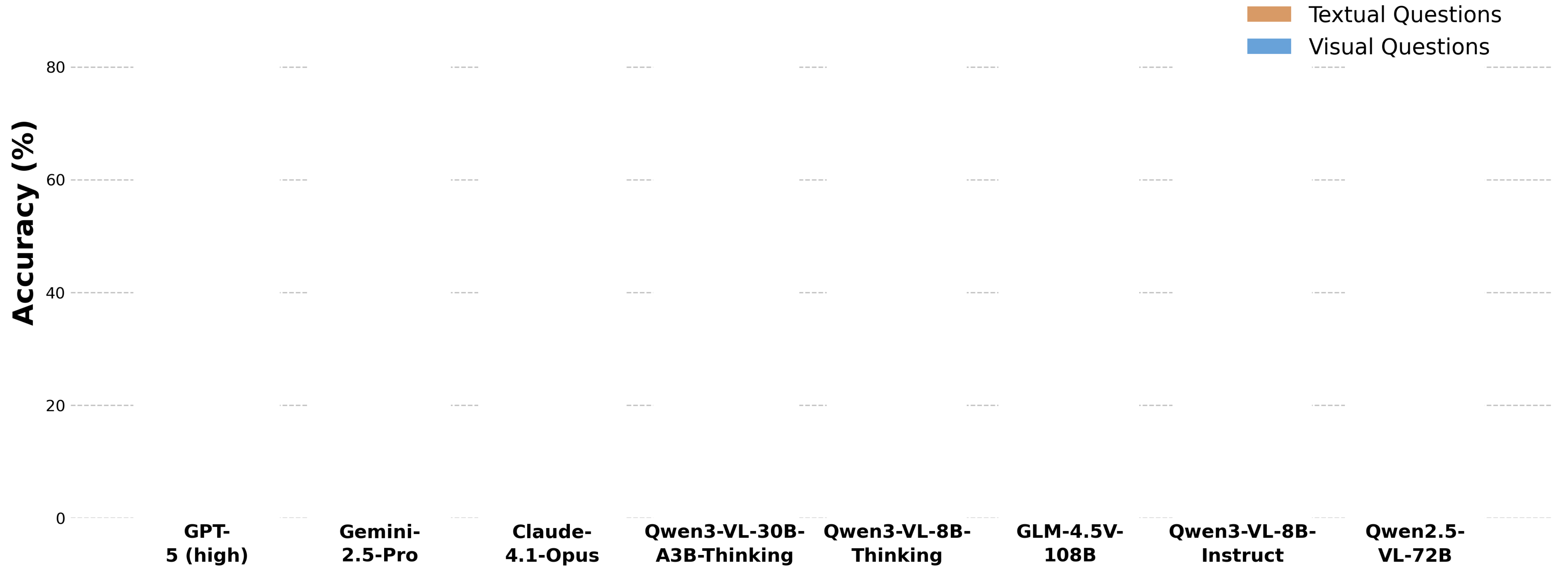


Wide range of SOTA model accuracy, and a large gap from human performance



# Results

LVLM's **Visual Reasoning** is far behind **Textual Reasoning**



Humans achieve **98%** on the visual questions

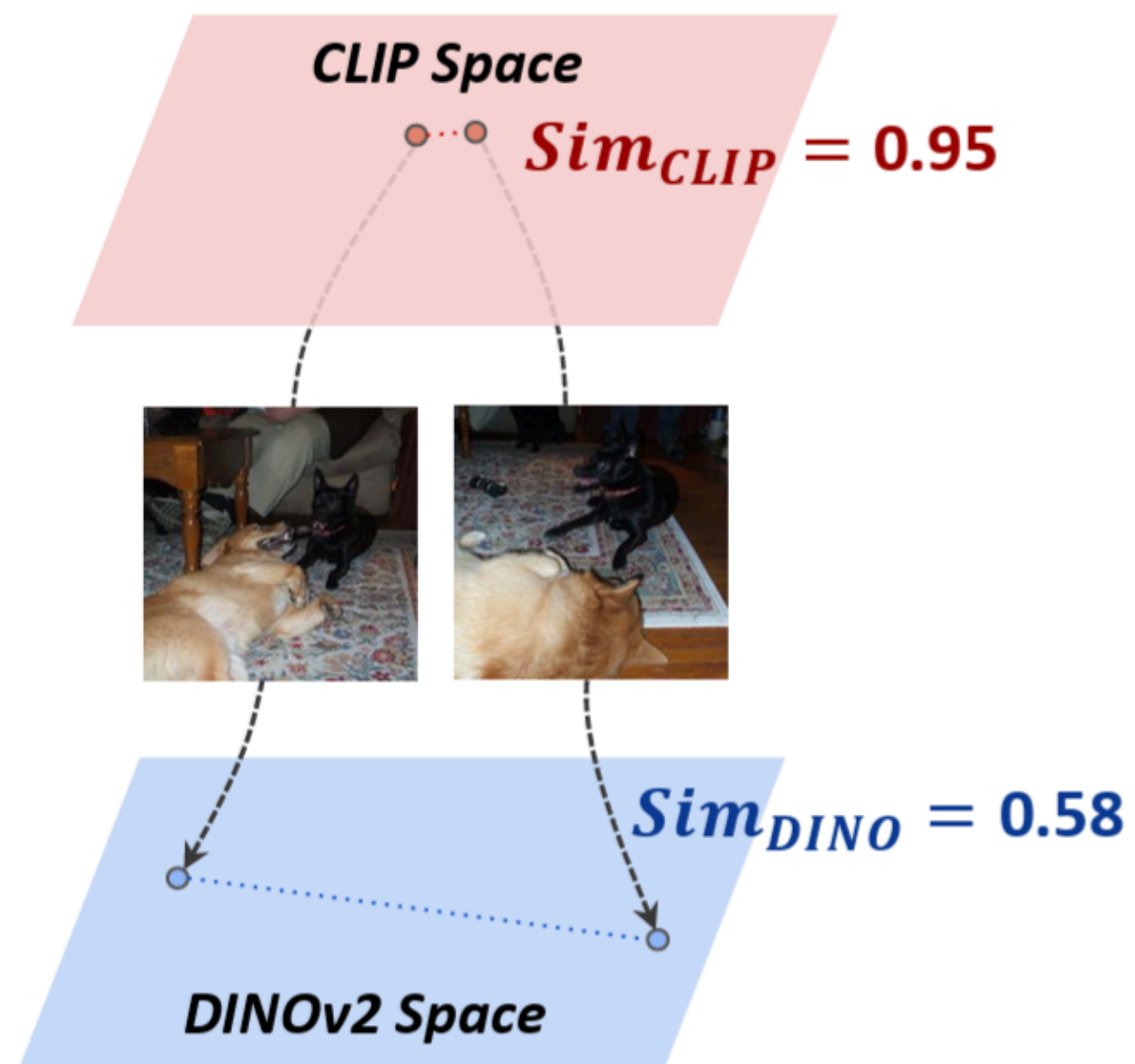
Check **error analysis** in the paper

# How well does this work?

Step 1

## Finding CLIP-blind pairs.

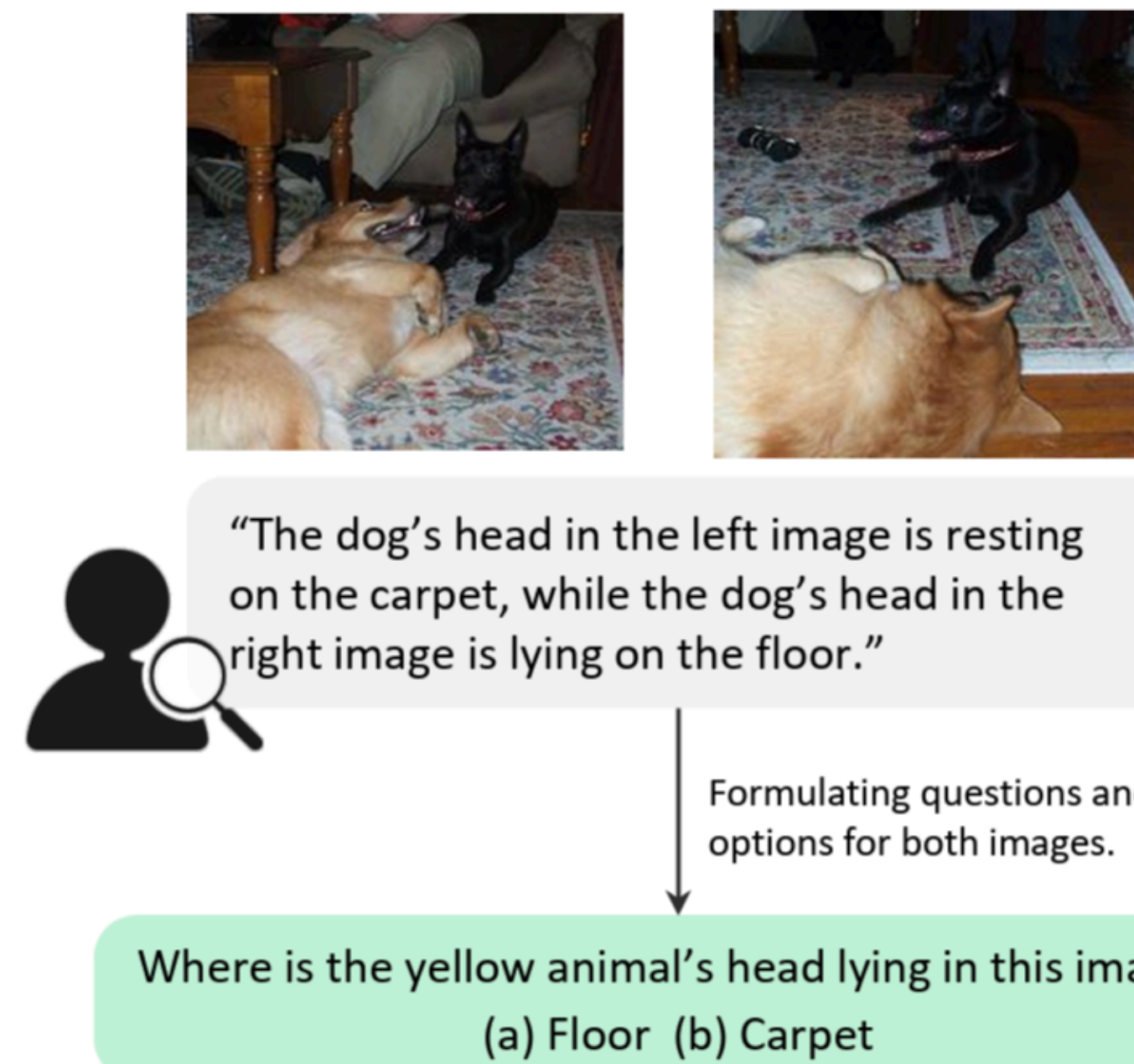
Discover image pairs that are proximate in CLIP feature space but distant in DINOv2 feature space.



Step 2

## Spotting the difference between two images.

For a CLIP-blind pair, a human annotator attempts to spot the visual differences and formulates questions.

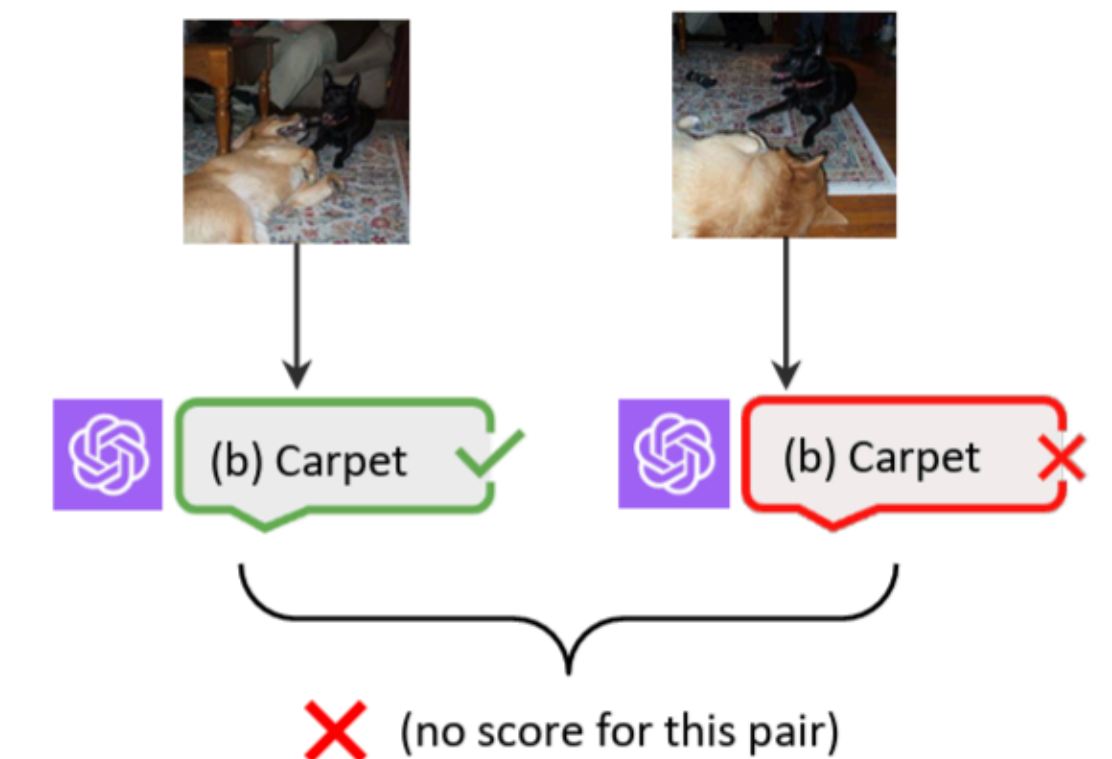


Step 3

## Benchmarking multimodal LLMs.

Evaluate multimodal LLMs using a CLIP-blind image pair and its associated question.

Where is the yellow animal's head lying in this image?  
(a) Floor (b) Carpet



The model receives a score only when **both** predictions for the CLIP-blind pair are correct.

Administrative details and recap

Language Grounding

Grounding in Images

Visual Reasoning

Thinking with Images

Vision-Language-Action

# Thinking with Images

# Thinking with Images

Think about Images

Think with Images

Intelligence Increasing



## Think with Text

*Textual reasoning over a static visual context.*



## Stage 1: Tool-Driven Visual Exploration

*Commanding a fixed toolkit for visual analysis.*



## Stage 2: Programmatic Visual Manipulation

*Creating bespoke visual operations via code generation.*



## Stage 3: Intrinsic Visual Imagination

*Internally generating visual thoughts within a cognitive loop.*

# Visual Exploration



## Stage (i): Tool-Driven Visual Exploration

### ① Semantic Grounding Tools



OCR

"Einstein Kit Best for Learning: Teaches over 100+ scientific concepts  
Age Appropriate: Just right for 6-12 year old children..."

Q: Is Einstein kit suitable for my 11 year old daughter?

A: Yes. Kit is very suitable for 11 year old children

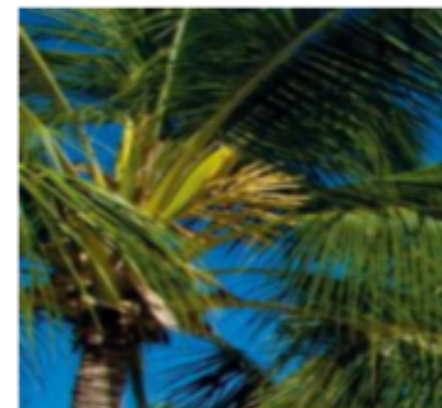
### ② Perception Exploration Tools



Zoom in

Thoughts:

I need to see the details to identify it. Zooming in.



Q: What type of tree is the highest in the picture?

A: Coconut tree.

### ③ Visual Reasoning-Aid Tools



Point

Thoughts:

I'll mark the pears to keep track while counting.



Q: How many pears are in the plate?

A: Three.

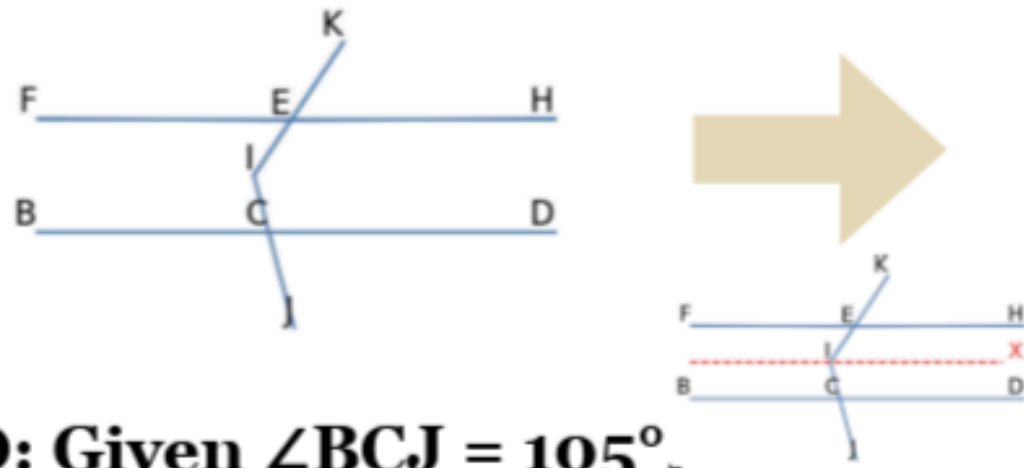
From Selection

# Visual Manipulation



## Stage (ii): Programmatic Visual Manipulation

### ① Visual Operations



**Q: Given  $\angle BCF = 105^\circ$ ,  $\angle KEF = 105^\circ$ , find  $\angle EIC$ .**

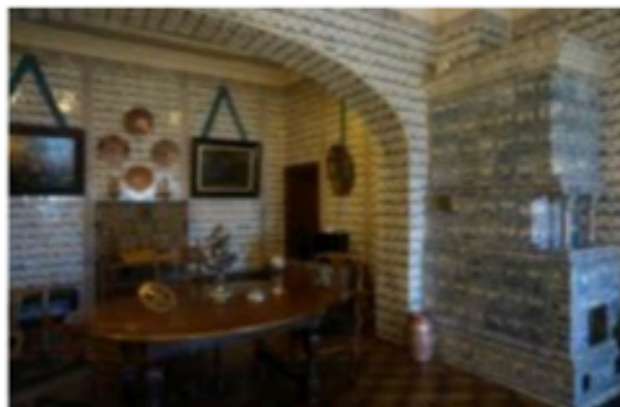
**Thoughts:**

Draw an auxiliary line through point will help solve it

```
# Define auxiliary point X for line IX
x_I = points["I"][0]
y_I = points["I"][1]
x_X = x_I + 100
y_X = y_I + slope_bd * (x_X - x_I)
ax.plot([x_I, x_X], [y_I, y_X], 'red')
```

**A:  $150^\circ$**

### ② Logical Operations



**Q: Where is the bookshelf relative to the chair to the left of the vase?**

**Thoughts:**

Building the spatial logic for 'chair left of vase'

```
for v_patch in vases:
    chairs_on_left = [
        c_patch for c_patch in chairs
        if c_patch.horizontal.center <
        v_patch.horizontal.center
    ]
```

**A: It is to the left.**

### ③ Composite Operations



**Q: Find red circles larger than the average.**



**Thoughts:**

I will Calculate average area, then filter the answer.

```
circles = find_all(image, "red & circle")
avg_area = average([c.area for c in circles])
large_red_circles = filter(circles,
    lambda c: c.area > avg_area)
mark_objects(image, large_red_circles)
```



**A: Two circles.**

to Creation

From Execution

# Visual Imagination



## Stage (iii): Intrinsic Visual Imagination

### ① Implicit Latent Reasoning



Q: Please redraw the second image using the style of the first image.

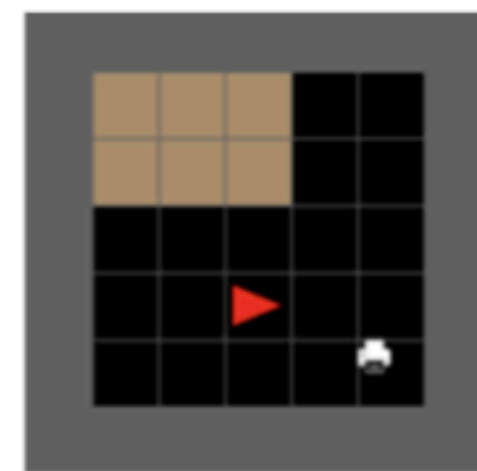
Thoughts:  
First, I need to capture the image's style and features.



From the captured feature, I will generate the final image:

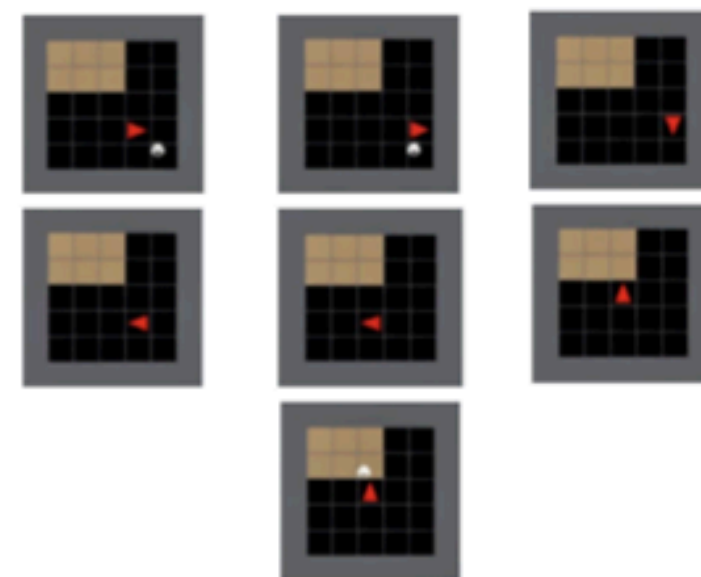


### ② Explicit Visual Reasoning



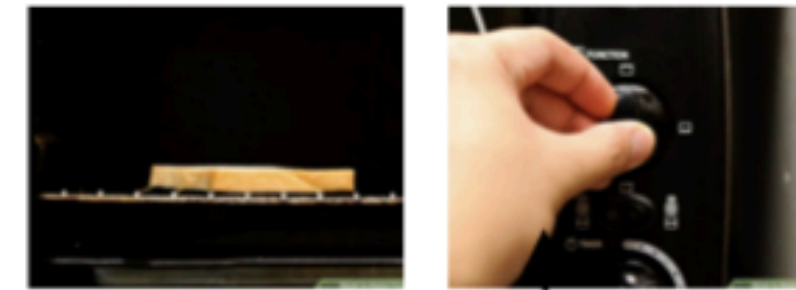
Q: Can I get the treasure?

Thoughts:



A: Action Success.

### ③ Interleaved Multimodal Reasoning



Step 1: Place bread on rack. Step 2: Turn on oven broiler.

Q: Predict the next steps to make toast.

Thoughts:

[Text] Now, the bread needs to be toasted until golden brown. [Text] This shows baking. A better next step is to flip the bread.

[Visual]

[Visual]



to Imagination

# General Paradigm

---

Administrative details and recap

Language Grounding

Grounding in Images

Visual Reasoning

Thinking with Images

Vision-Language-Action

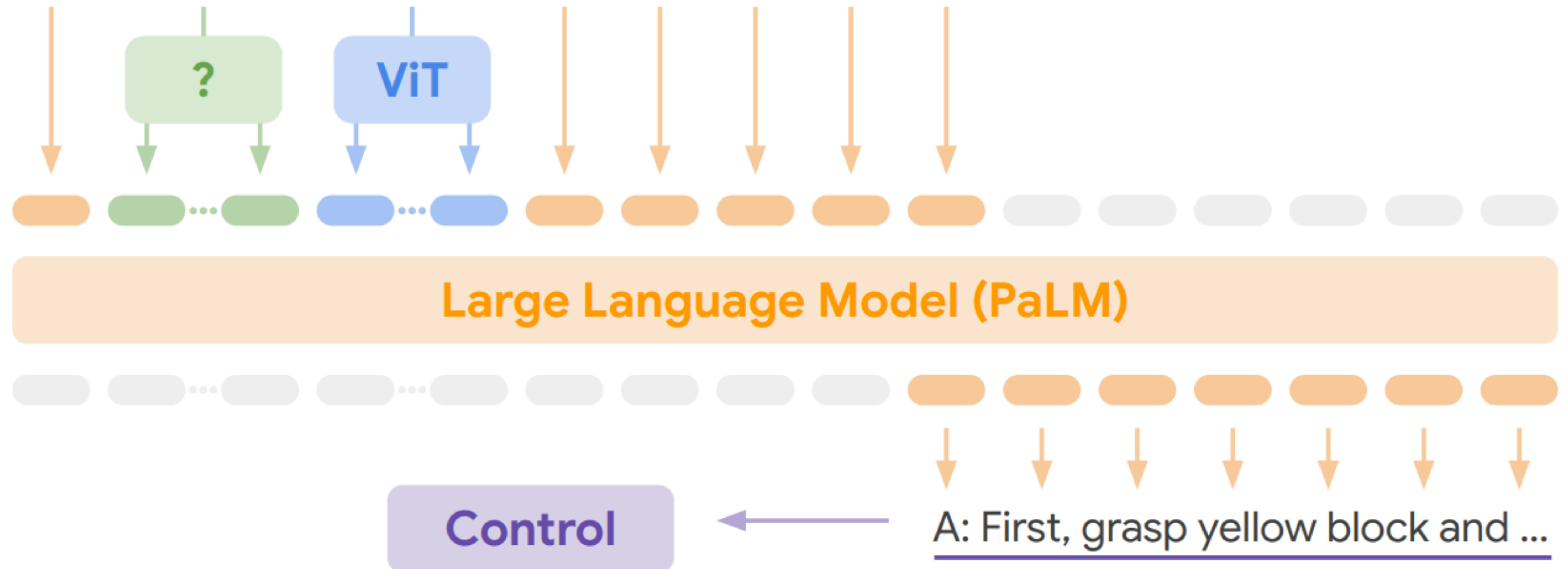
# Vision-Language-Action

# PaLM-E

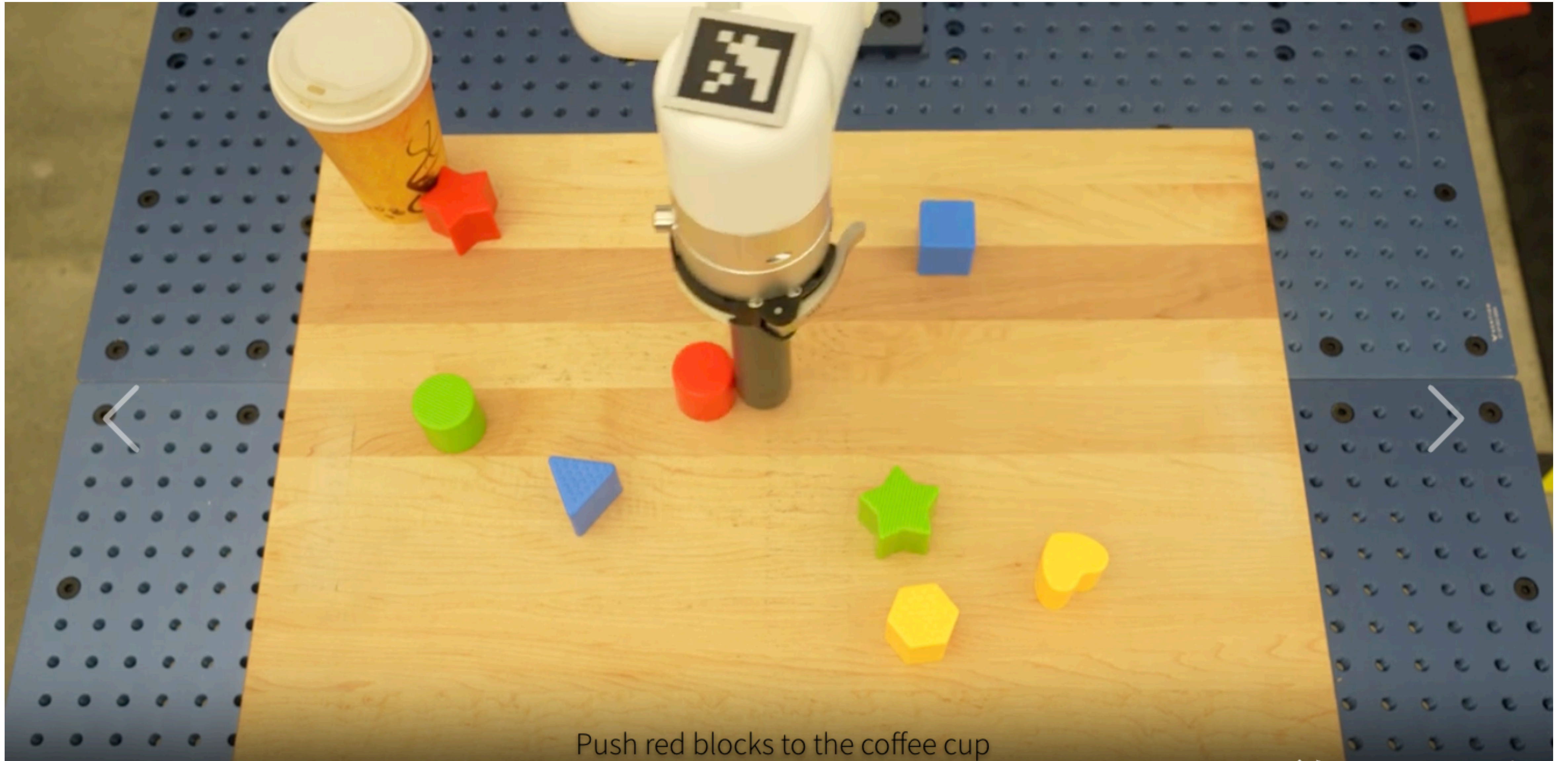
- ▶ Most models like CLIP are just vision+language

## PaLM-E: An Embodied **Multimodal Language** Model

Given **<emb>** ... **<img>** Q: How to grasp blue block? A: First, grasp yellow block



# PaLM-E



Push red blocks to the coffee cup

# OpenVLA

---

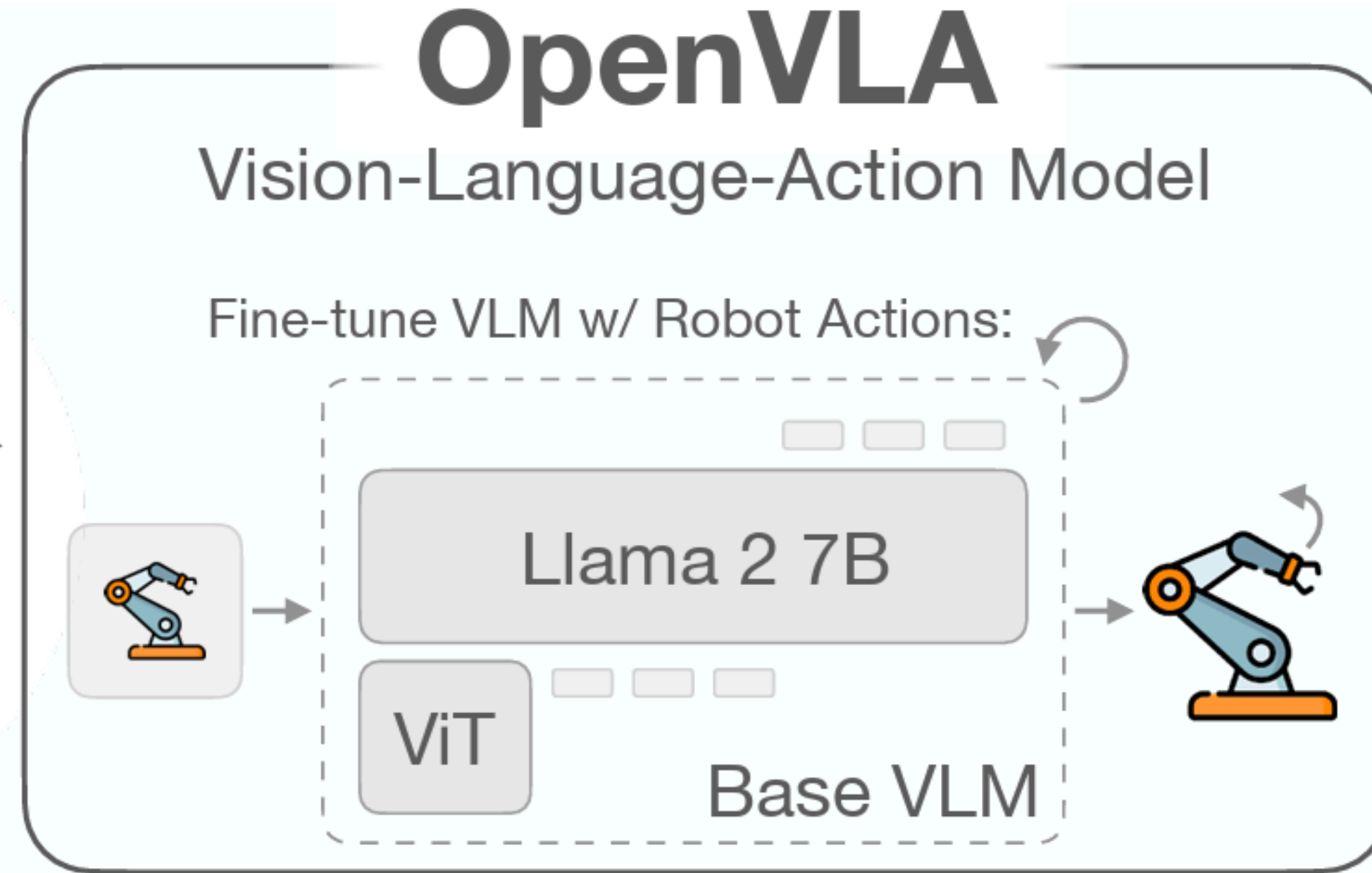
## OpenVLA: An Open-Source Vision-Language-Action Model

**Moo Jin Kim<sup>\*,1</sup> Karl Pertsch<sup>\*,1,2</sup> Siddharth Karamcheti<sup>\*,1,3</sup>**  
**Ted Xiao<sup>4</sup> Ashwin Balakrishna<sup>3</sup> Suraj Nair<sup>3</sup> Rafael Rafailov<sup>1</sup> Ethan Foster<sup>1</sup> Grace Lam**  
**Pannag Sanketi<sup>4</sup> Quan Vuong<sup>5,†</sup> Thomas Kollar<sup>3</sup> Benjamin Burchfiel<sup>3</sup> Russ Tedrake<sup>3,6</sup> Dorsa Sadigh<sup>1</sup>**  
**Sergey Levine<sup>2</sup> Percy Liang<sup>1</sup> Chelsea Finn<sup>1</sup>**

<https://openvla.github.io>

# OpenVLA

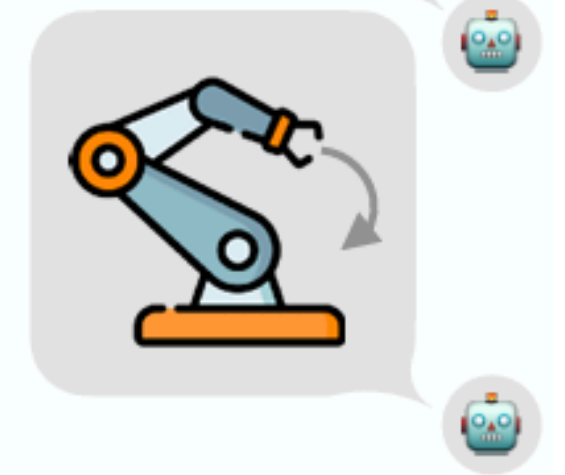
Large-Scale Robot Training Data



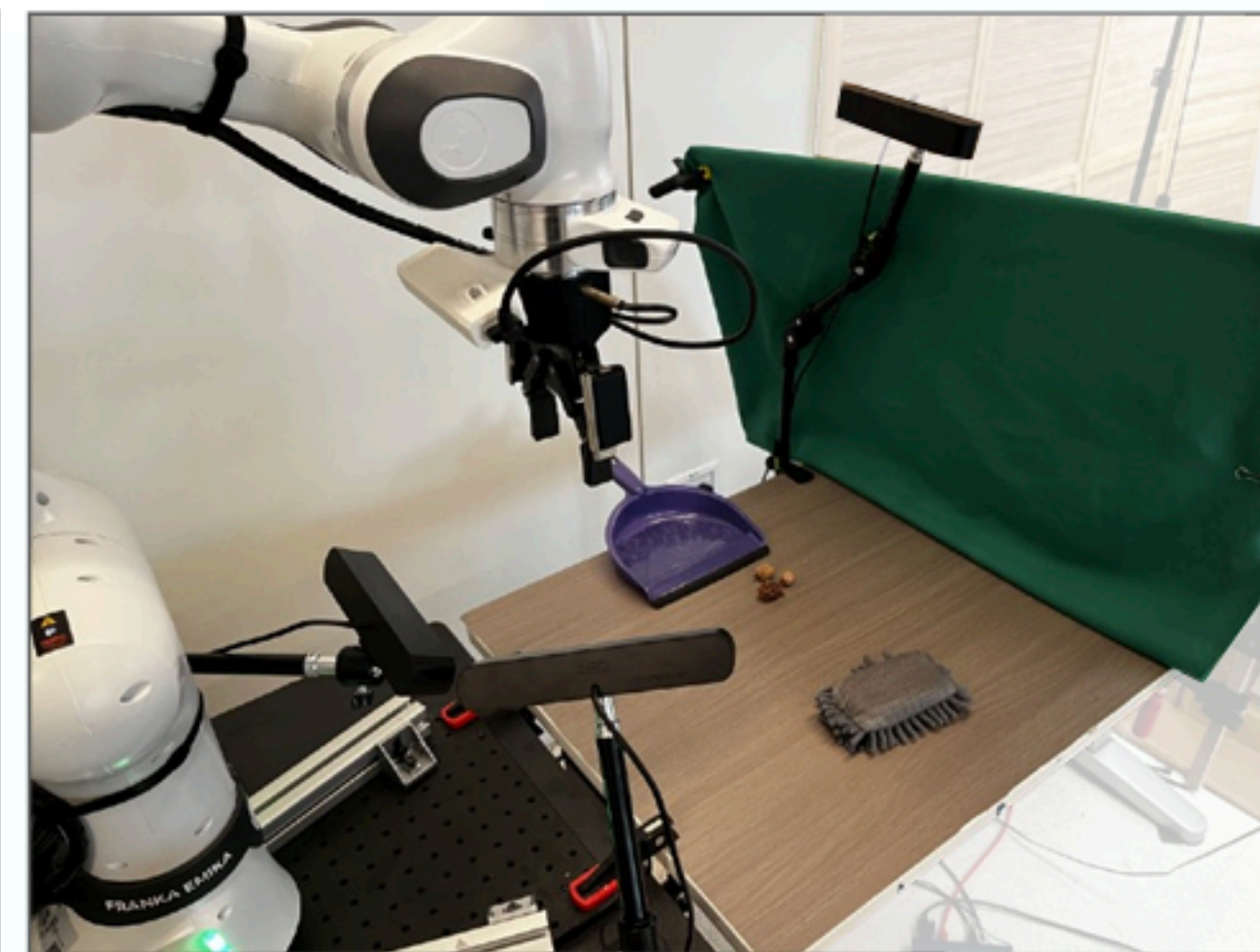
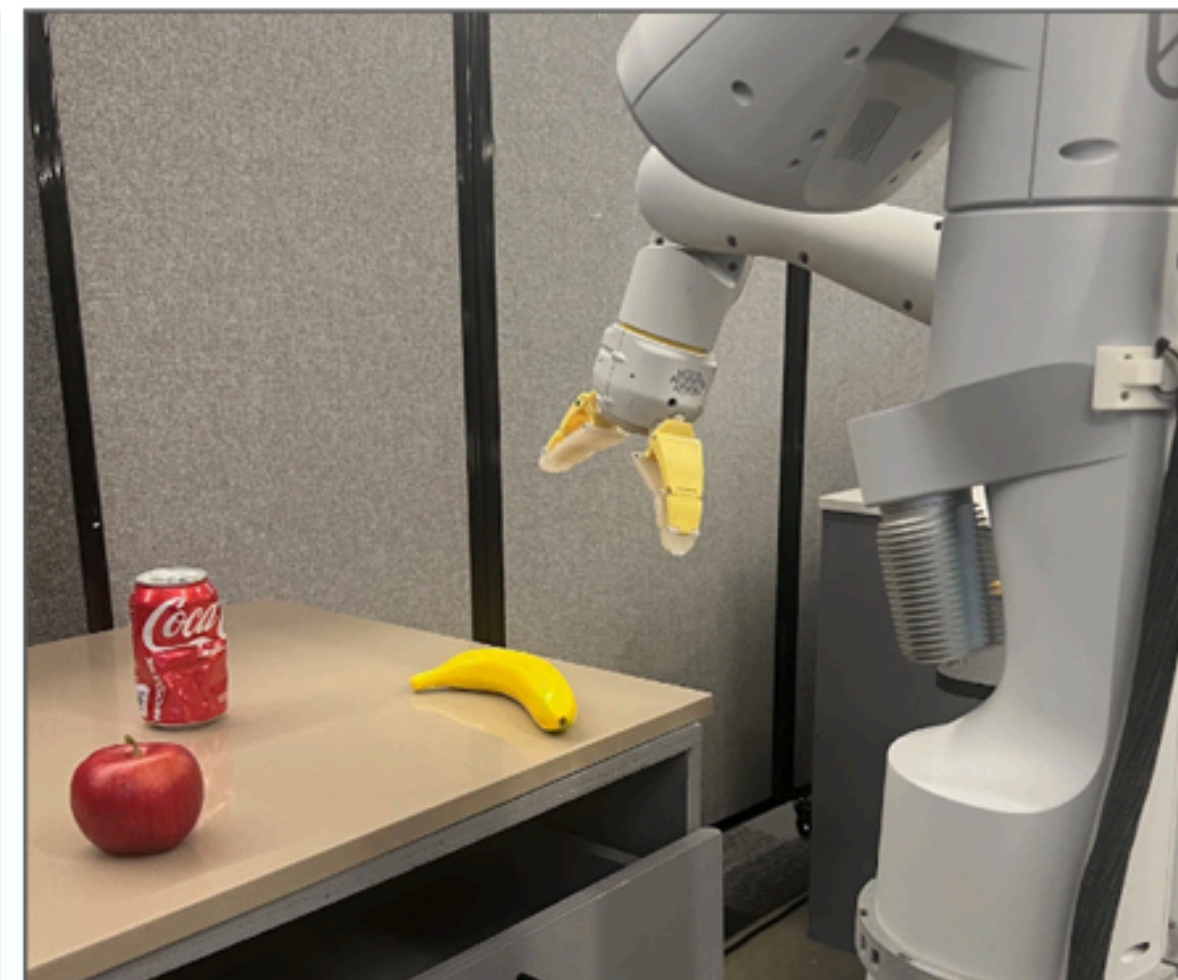
Closed-Loop Robot Control Policy

User: Wipe the table.




OpenVLA:  
 $[\Delta x, \Delta \theta, \Delta \text{Grip}] = \dots$



Multi-Robot Control & Efficient Fine-Tuning



Fully Open-Source

-  Data
-  Weights
-  Code

# OpenVLA

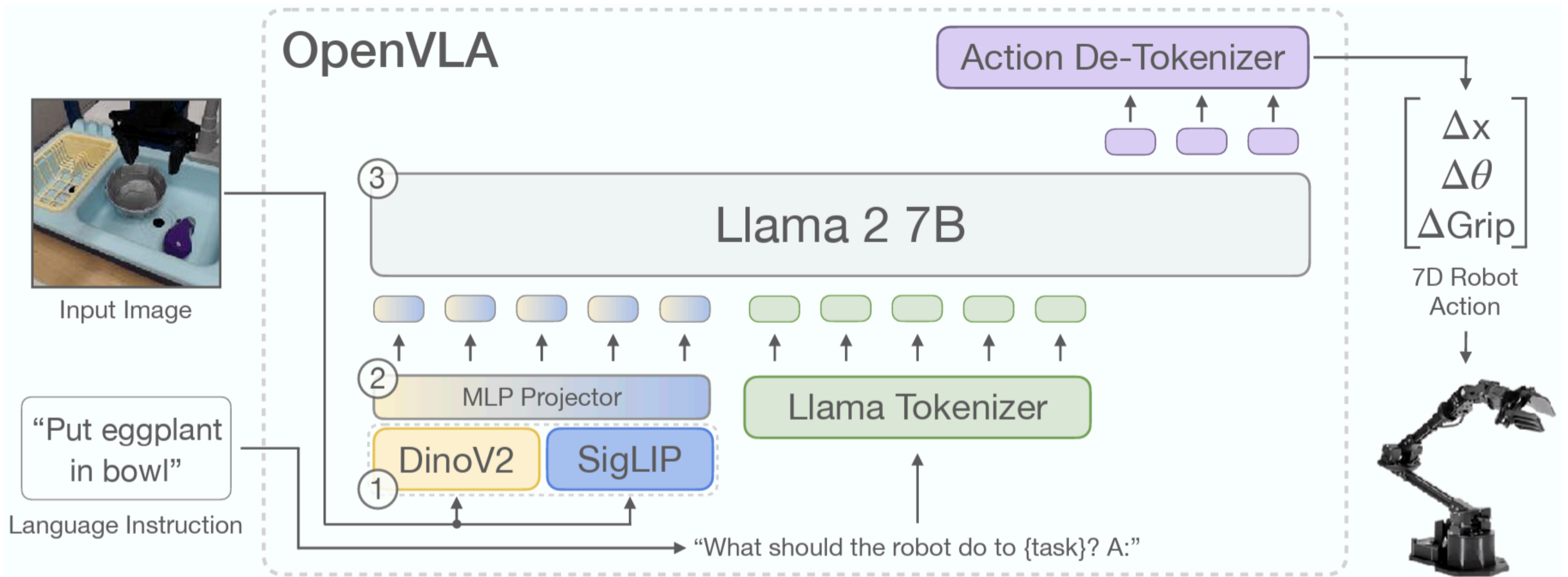


Figure 2: **OpenVLA model architecture.** Given an image observation and a language instruction, the model predicts 7-dimensional robot control actions. The architecture consists of three key components: (1) a **vision encoder** that concatenates Dino V2 [25] and SigLIP [79] features, (2) a **projector** that maps visual features to the language embedding space, and (3) the **LLM backbone**, a Llama 2 7B-parameter large language model [10].

# OpenVLA

---



2025-3-28

---

## GR00T N1: An Open Foundation Model for Generalist Humanoid Robots

NVIDIA<sup>1</sup>

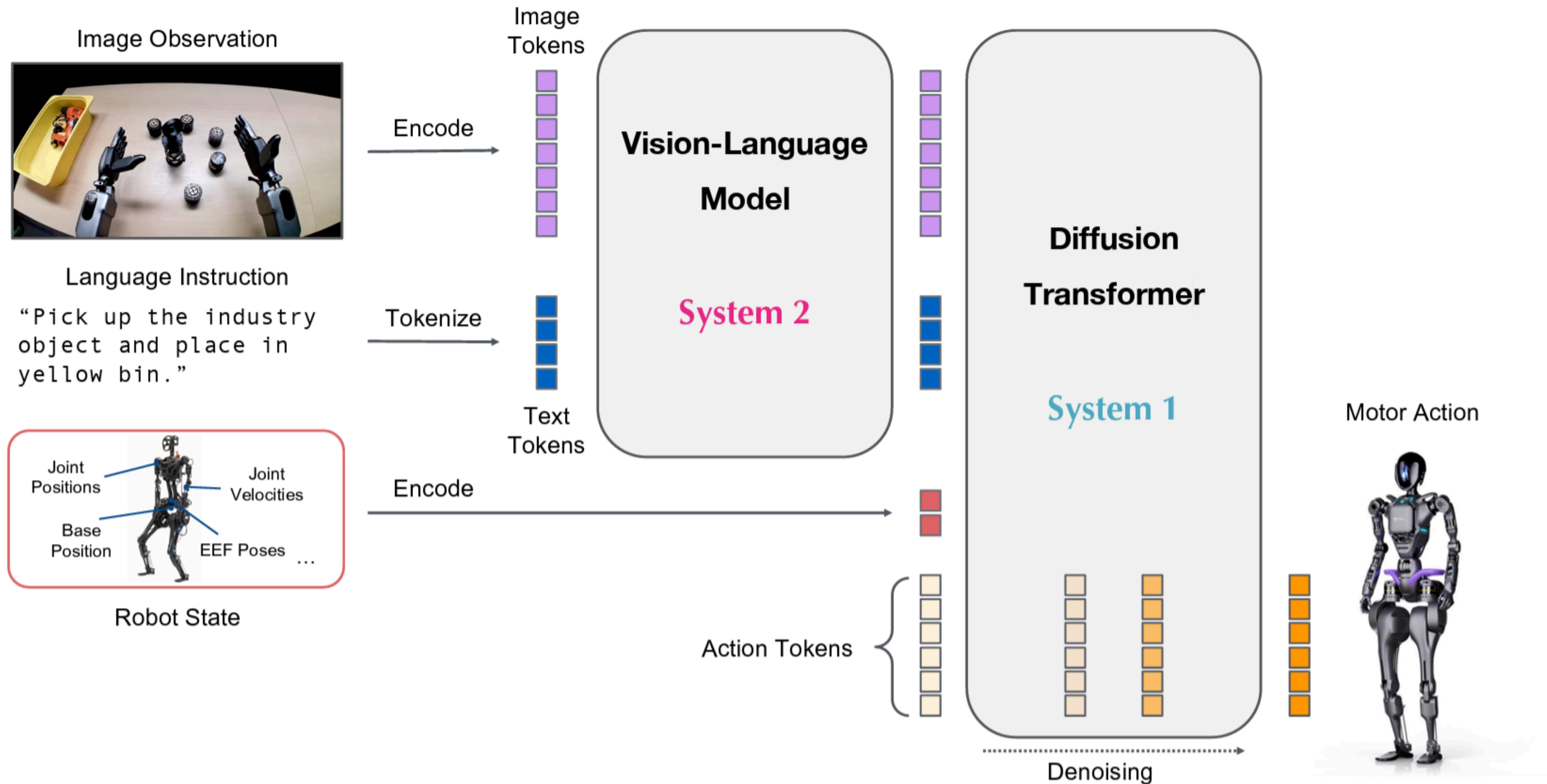


Figure 2: **GR00T N1 Model Overview.** Our model is a Vision-Language-Action (VLA) model that adopts a dual-system design. We convert the image observation and language instruction into a sequence of tokens to be processed by the Vision-Language Model (VLM) backbone. The VLM outputs, together with robot state and action encodings, are passed to the Diffusion Transformer module to generate motor actions.

# Takeaways

---

- ▶ VLA models are a bit less universal than VLMs:
  - ▶ Require actions customized to a particular robotic platform
  - ▶ Images are likely drawn from a particular domain of physical scenes
  - ▶ Text may be less broad
- ▶ “Moravec’s paradox” — seemingly simple physical tasks like picking things up are much harder than seemingly complex tasks like writing poetry
- ▶ How much will LLMs / LLM reasoners advance robotics?

Administrative details and recap

Language Grounding

Grounding in Images

Visual Reasoning

Thinking with Images

Vision-Language-Action

