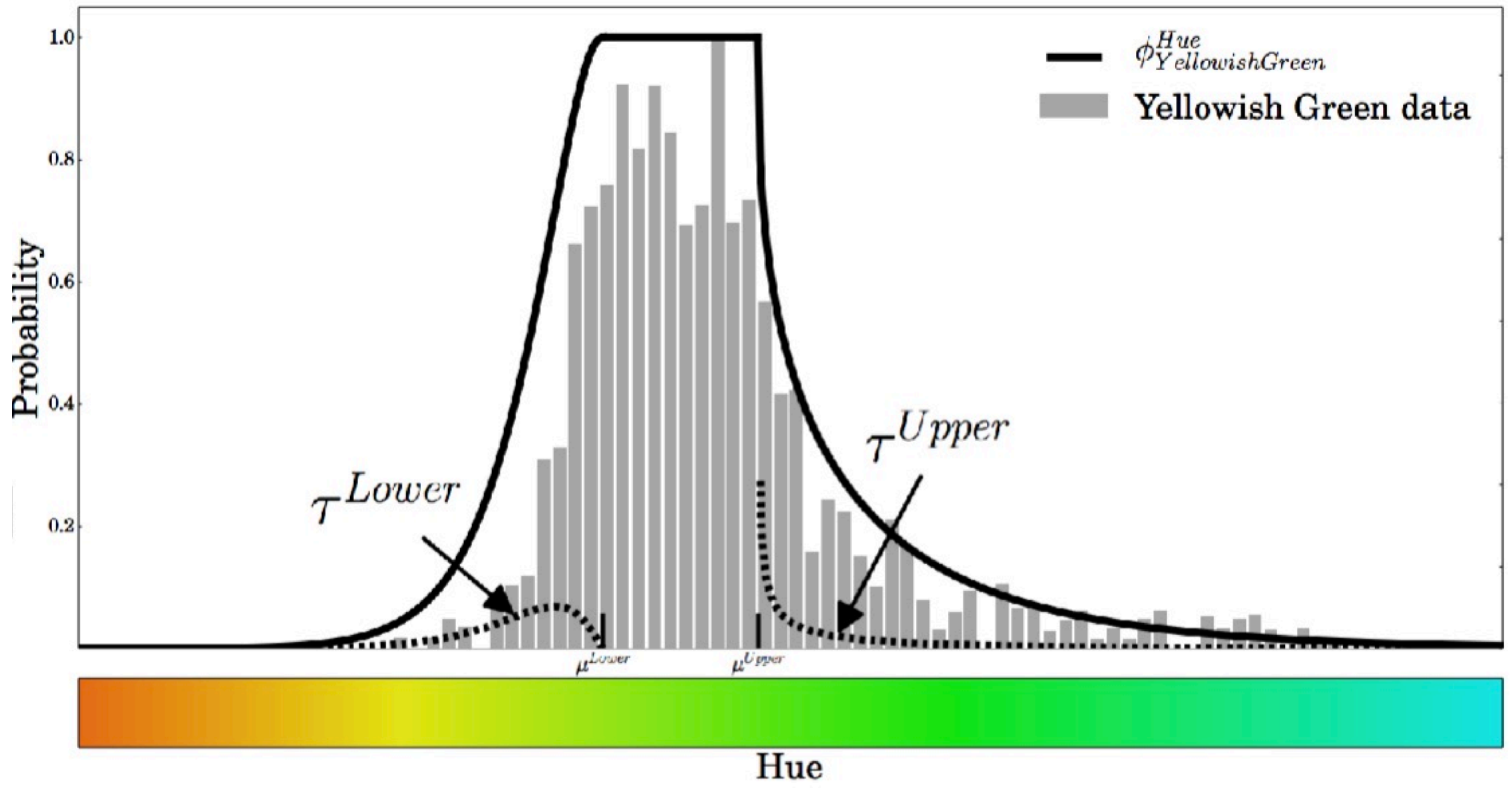
CS371N: Natural Language Processing

Lecture 24: Multimodality, Language Grounding

Greg Durrett

The University of Texas at Austin



McMahan and Stone (2015)

Announcements

- FP check-ins due Friday

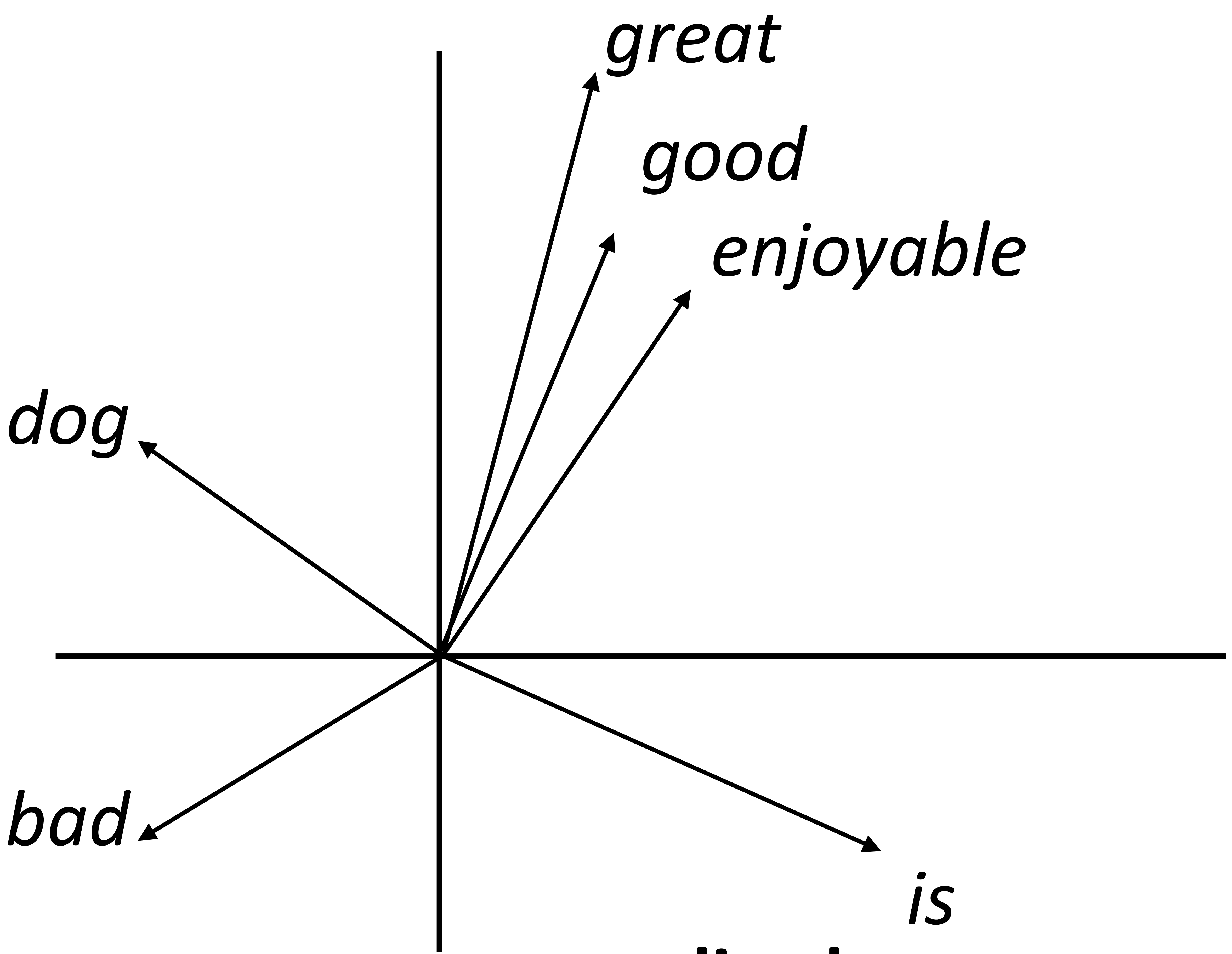
Today's Lecture

- Language grounding: how do we understand the meaning of language deeper than a system of abstract symbols?
- Multimodality
- Language and vision models
- Language and manipulation

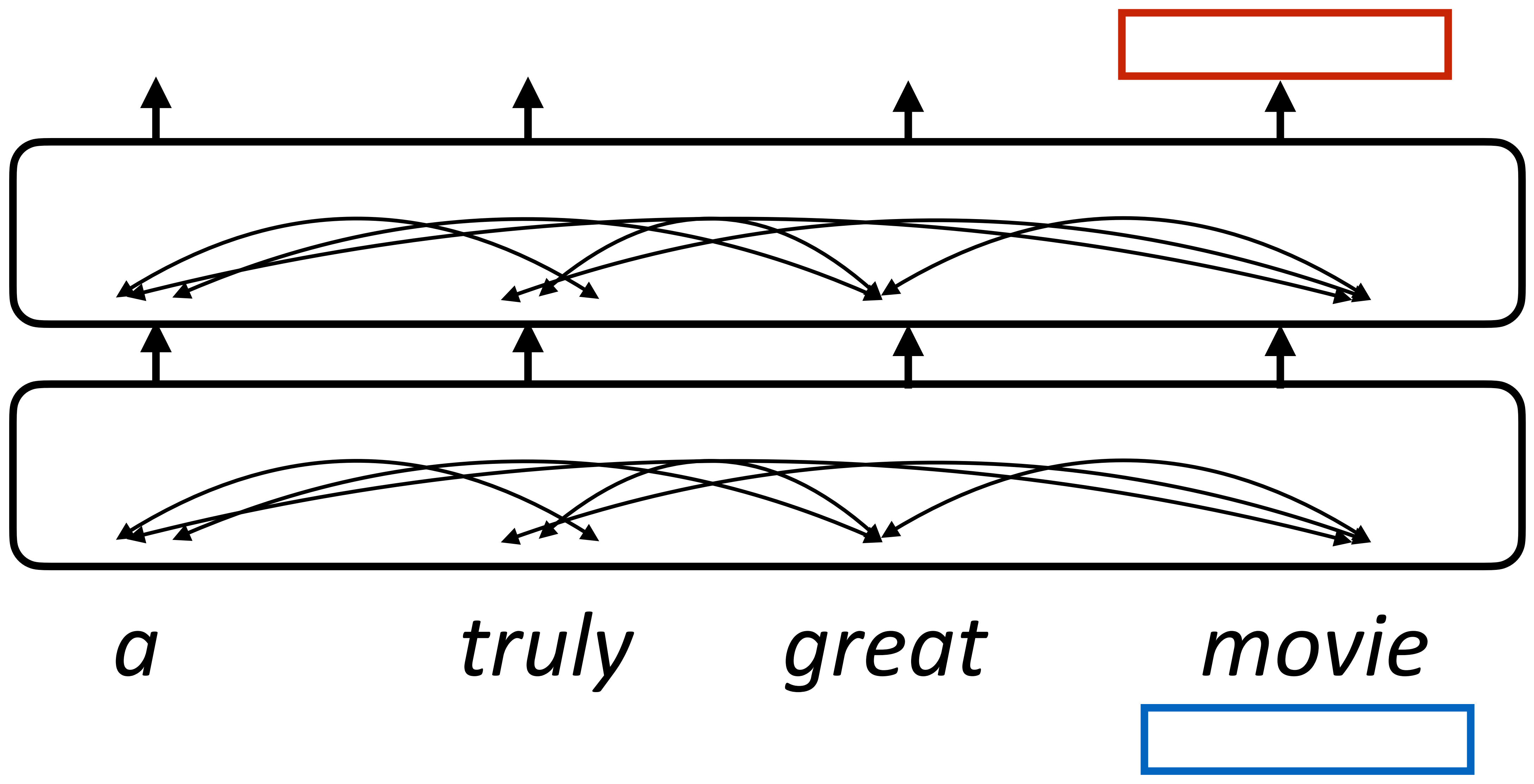
Classic Grounding

Language Grounding

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



non-contextualized



contextualized

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?

Harnad (1990) *The Symbol Grounding Problem*

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.

Searle (1980)

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.

Bender and Koller (2020) *Climbing towards NLU*

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:

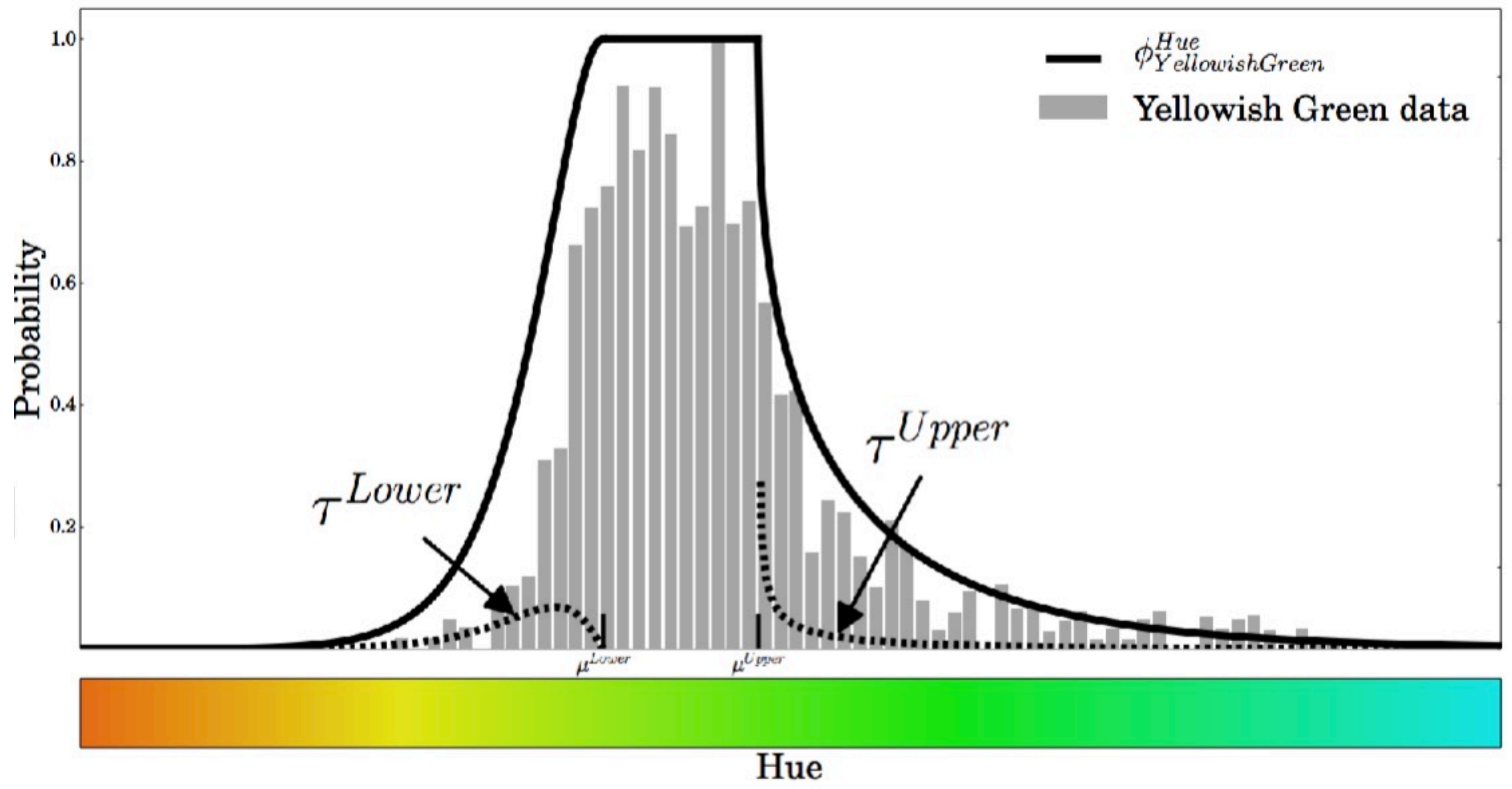
pure LM < LM fine-tuned on supervised data < vision+language LM < vision+language+manipulation LM < …

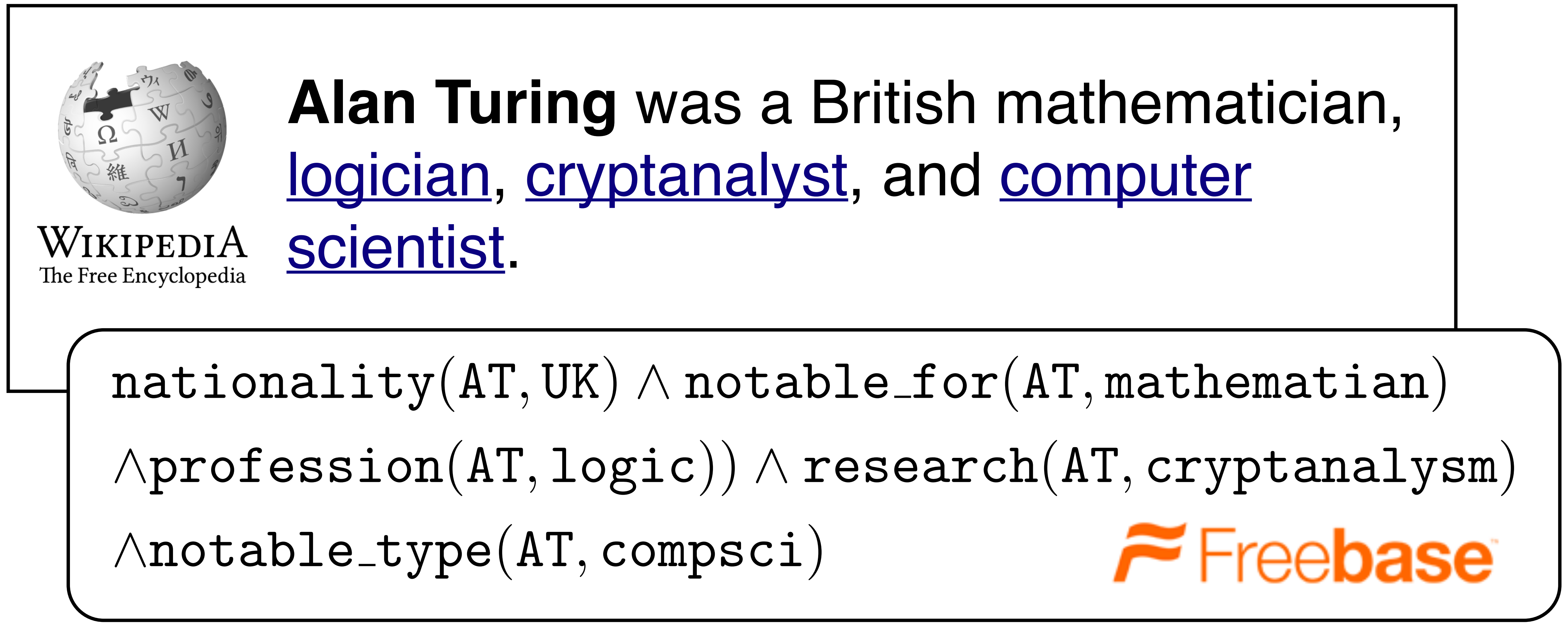
GPT-4 is here

PaLM-E (later)

Language Grounding

- There are many things that we can ground language in! Focus on vision today.
- How to associate words with sensory-motor experiences
- How to associate words with meaning representation



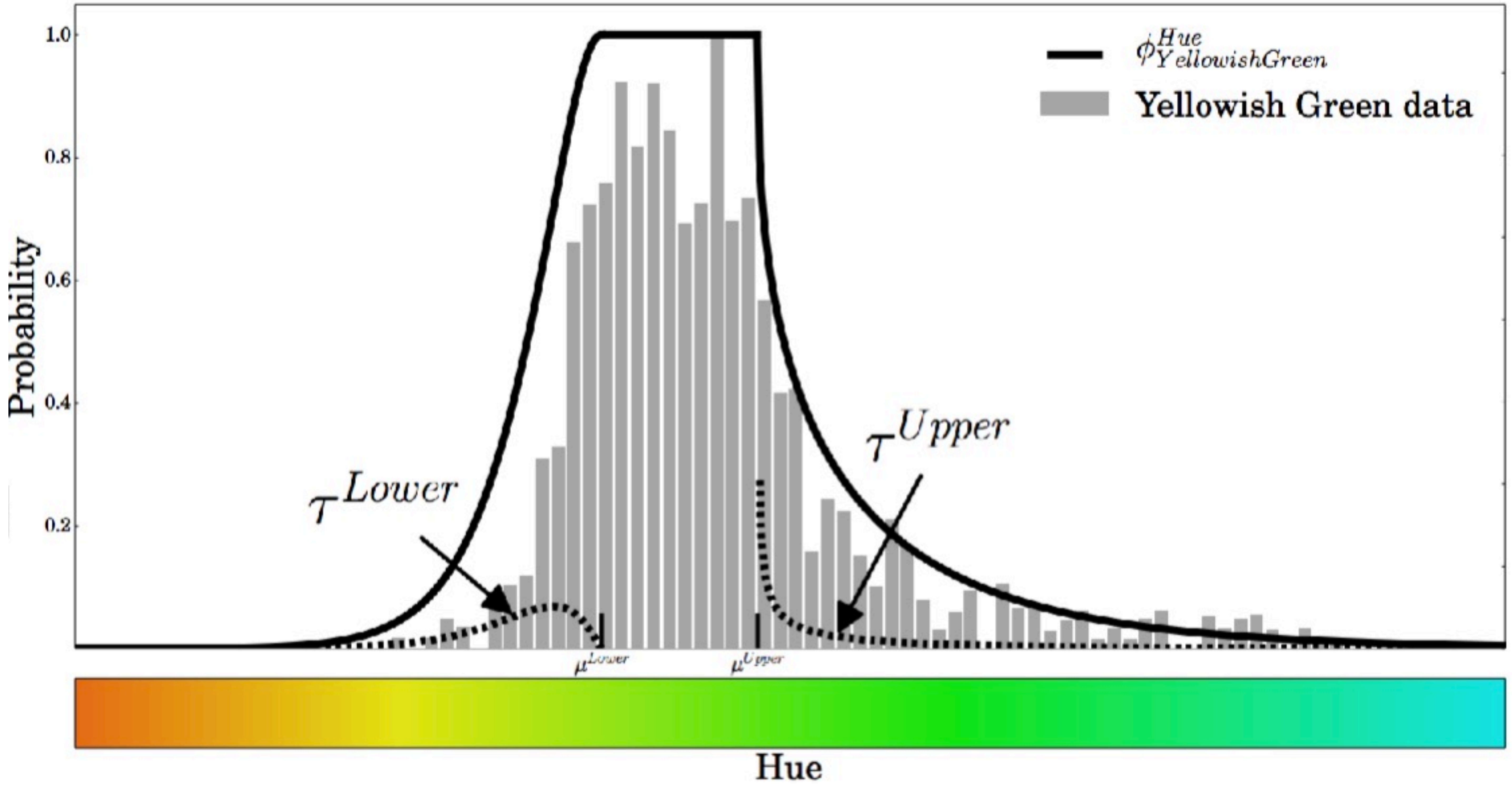


Multimodality, Language Grounding

some slides from Eunsol Choi

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)

Perception

- Visual: *green* = [0,1,0] in RGB
- Auditory: *loud* = >120 dB
- Taste: sweet = some threshold level of sensation on taste buds
- High-level concepts:

cat

dog

running

eating

Learning from Interaction

1. Use feedback from control application to understand language

Walk across the bridge

Reward +1

Alleviate dependence on large scale annotation

2. Use language to improve performance in control applications

Score: 7

+

1. **Ghosts** chase and try to kill you
2. Collect all the **pellets**
3. …

Score: 107

Other Grounding

- **Temporal concepts**
  - *late evening* = after 6pm. Ground in a time interval
  - *fast, slow* = describing rates of change
- **Functional:**
  - *Jacket:* keeps people warm
  - *Mug*: holds water
- **Spatial Relations**
  - *left, on top of, in front of*: how should we ground these?
- **Size:**
  - Whales are *larger* than lions
- **Focus today: grounding in images**

Language and Vision Models

Grounding in Images

- How would you describe this image?

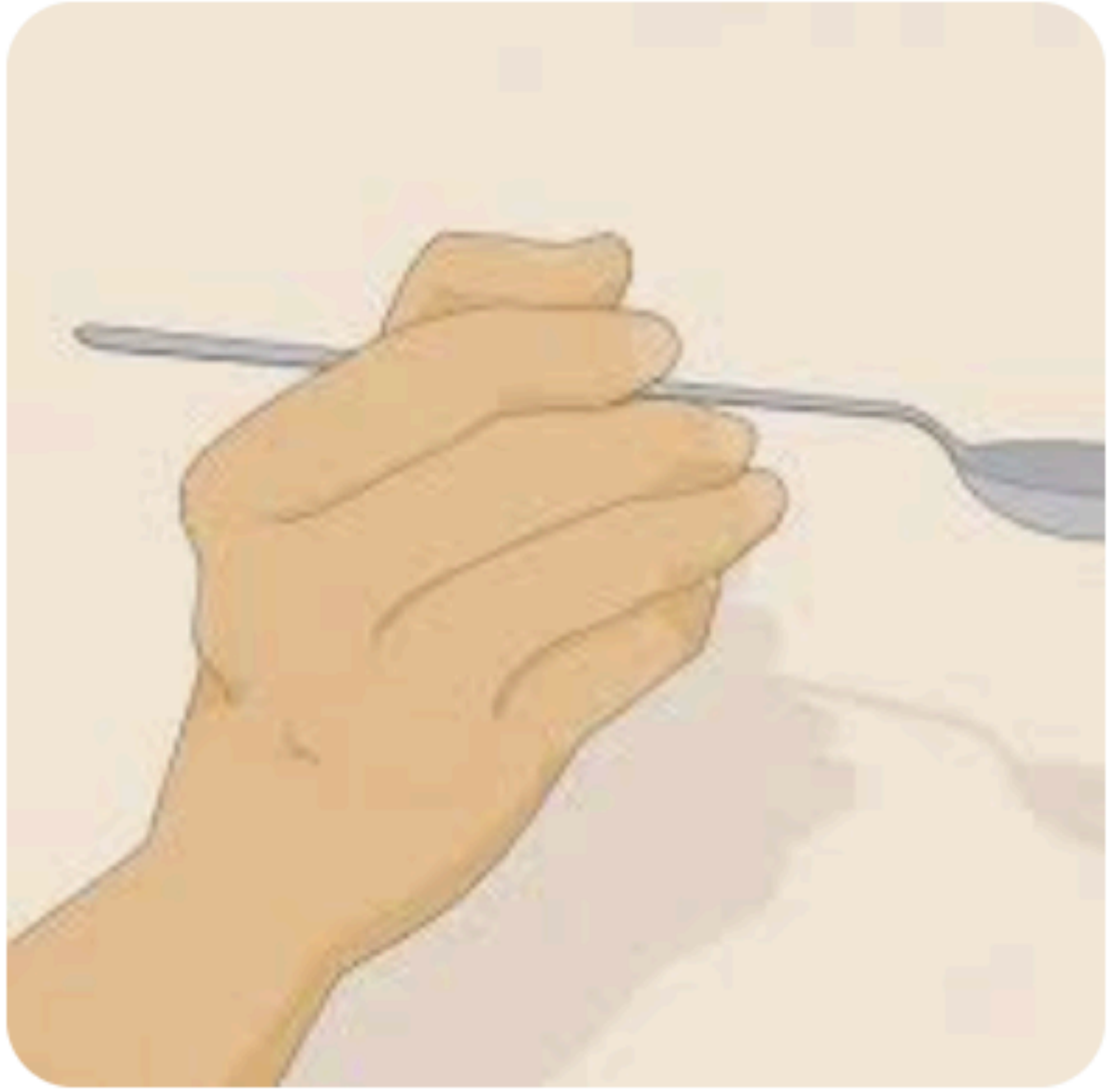
- What does the word "*spoon*" evoke?

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

the girl is licking the spoon of batter

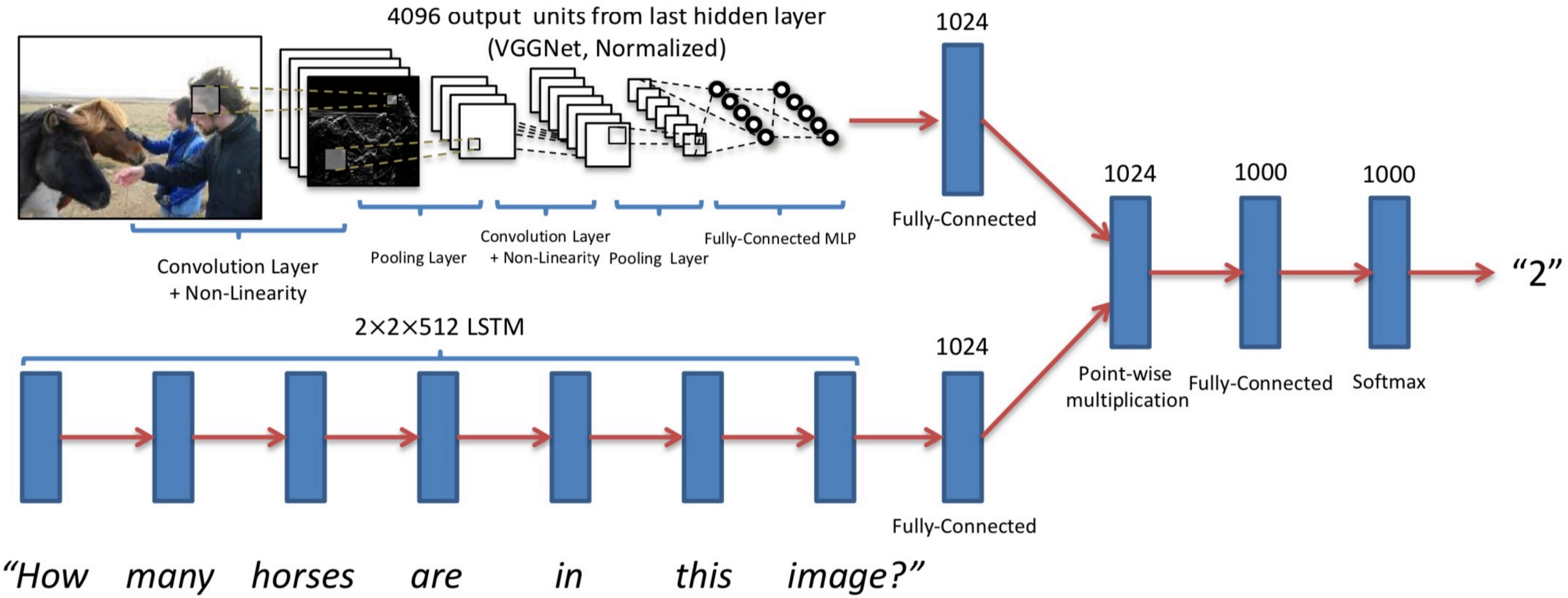
Image encoder
(CNN, Transformer)

Language encoder
(LSTM, Transformer)

Cross-attention/joint layer

Prediction

Visual Question Answering



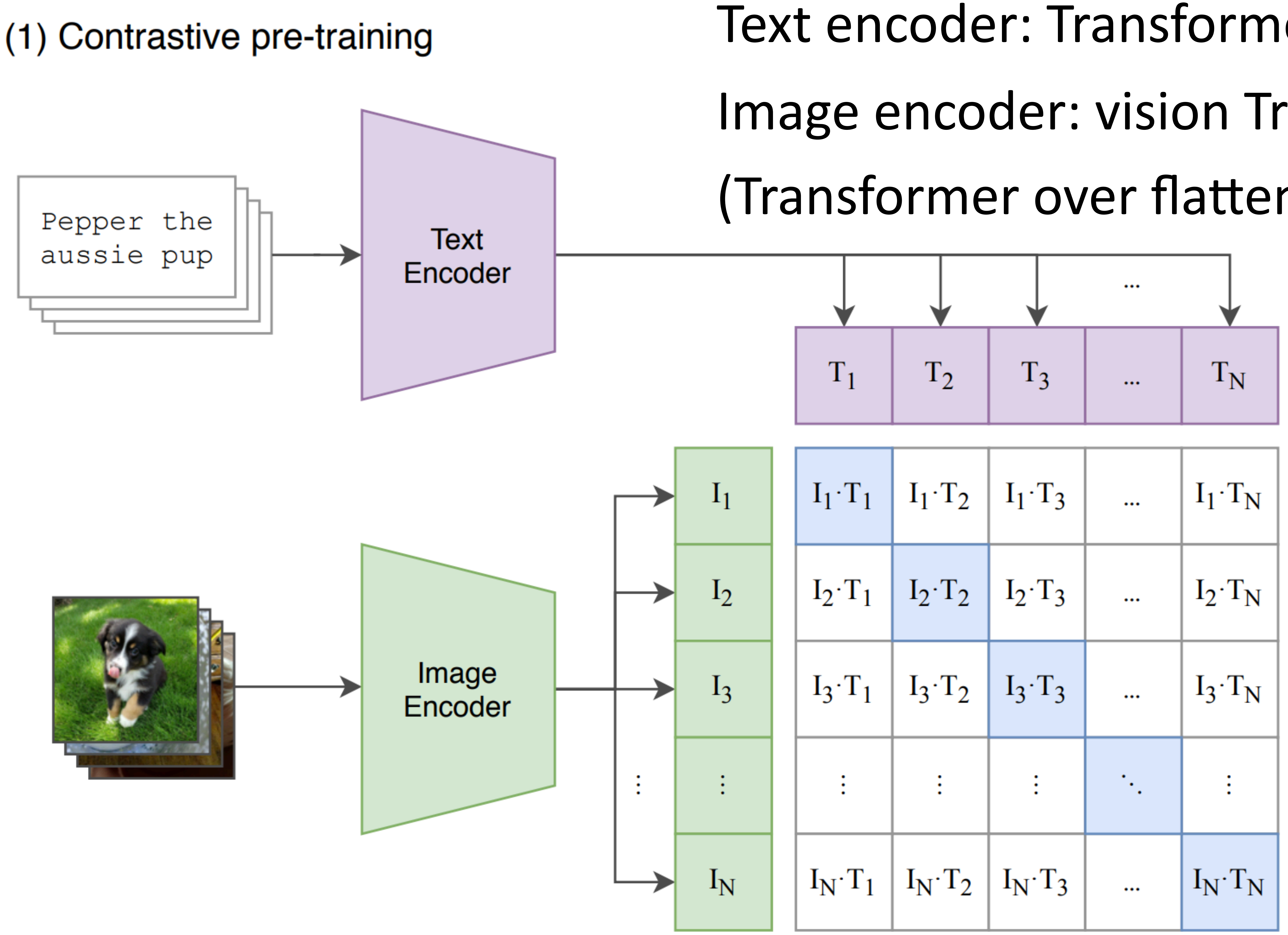
Agrawal et al., 2015

Language-vision Pre-training

Text encoder: Transformer

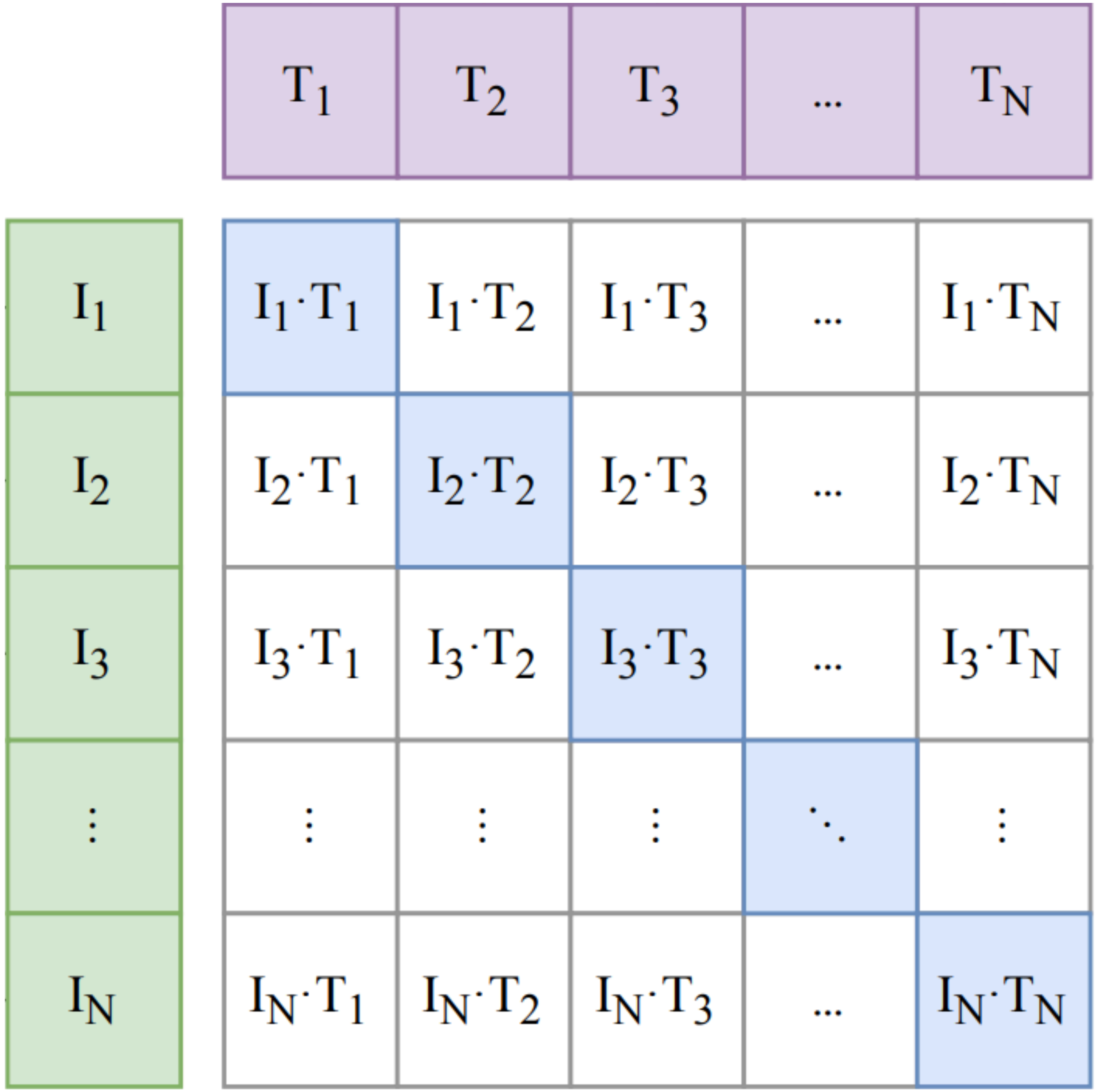
Image encoder: vision Transformer

(Transformer over flattened patches)



Radford et al., 2021

Language-vision Pre-training

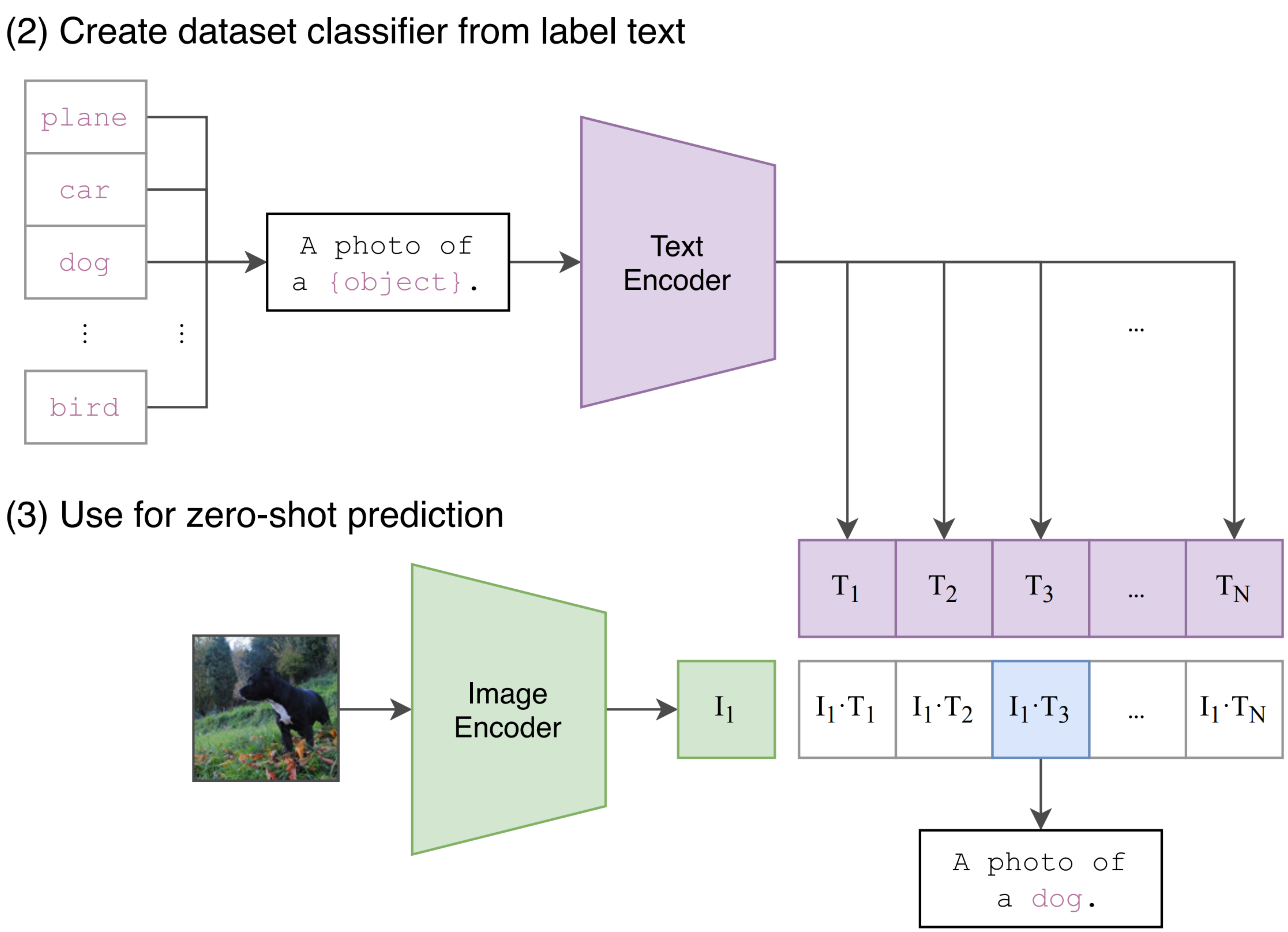


- Contrastive objective: each image should be more similar to its correspond caption than to other captions

$$\text{maximize softmax}(I_1^T T_i)[1] + \text{softmax}(I_2^T T_i)[2] + \ldots$$

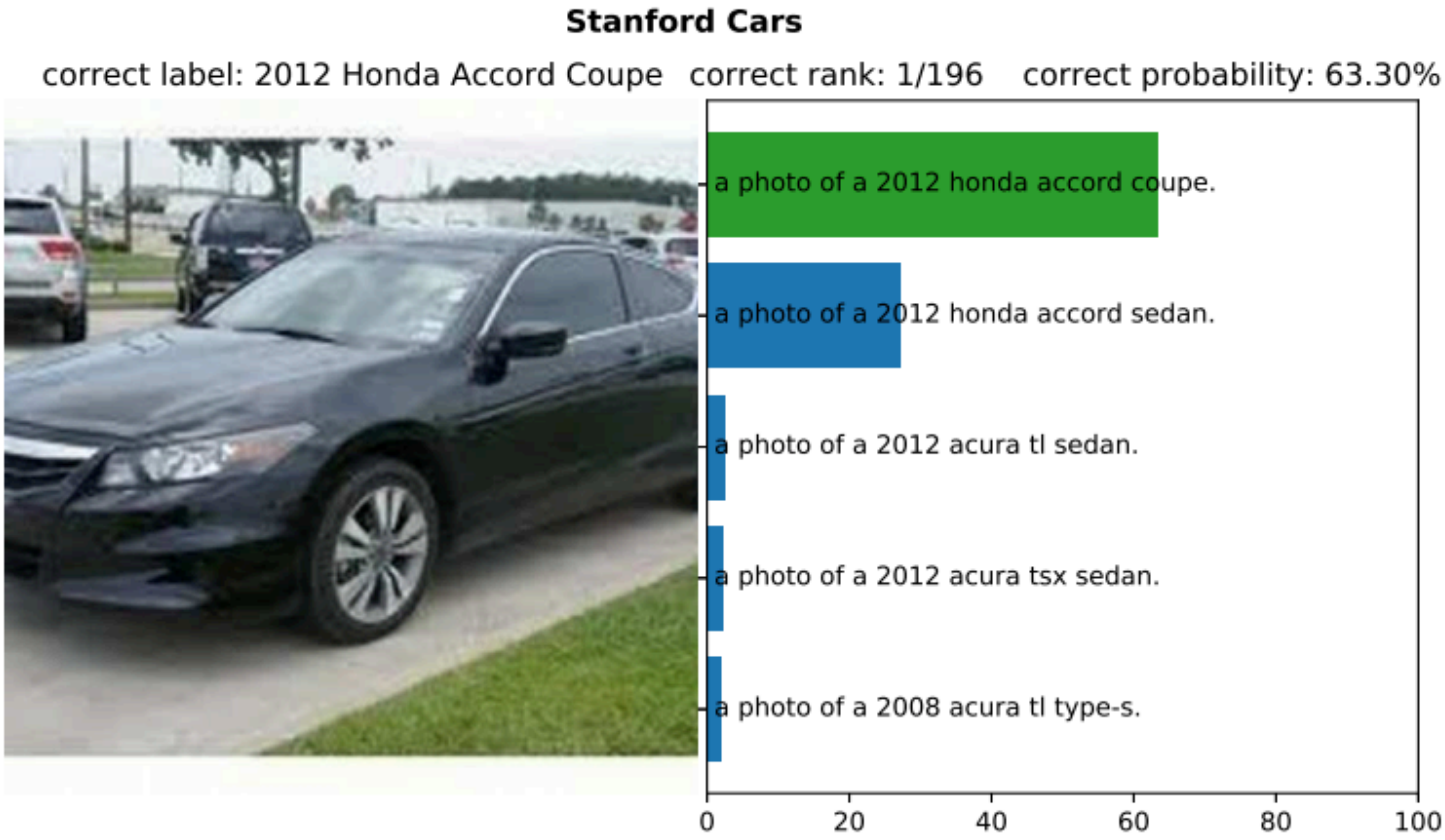
Radford et al., 2021

Language-vision Pre-training

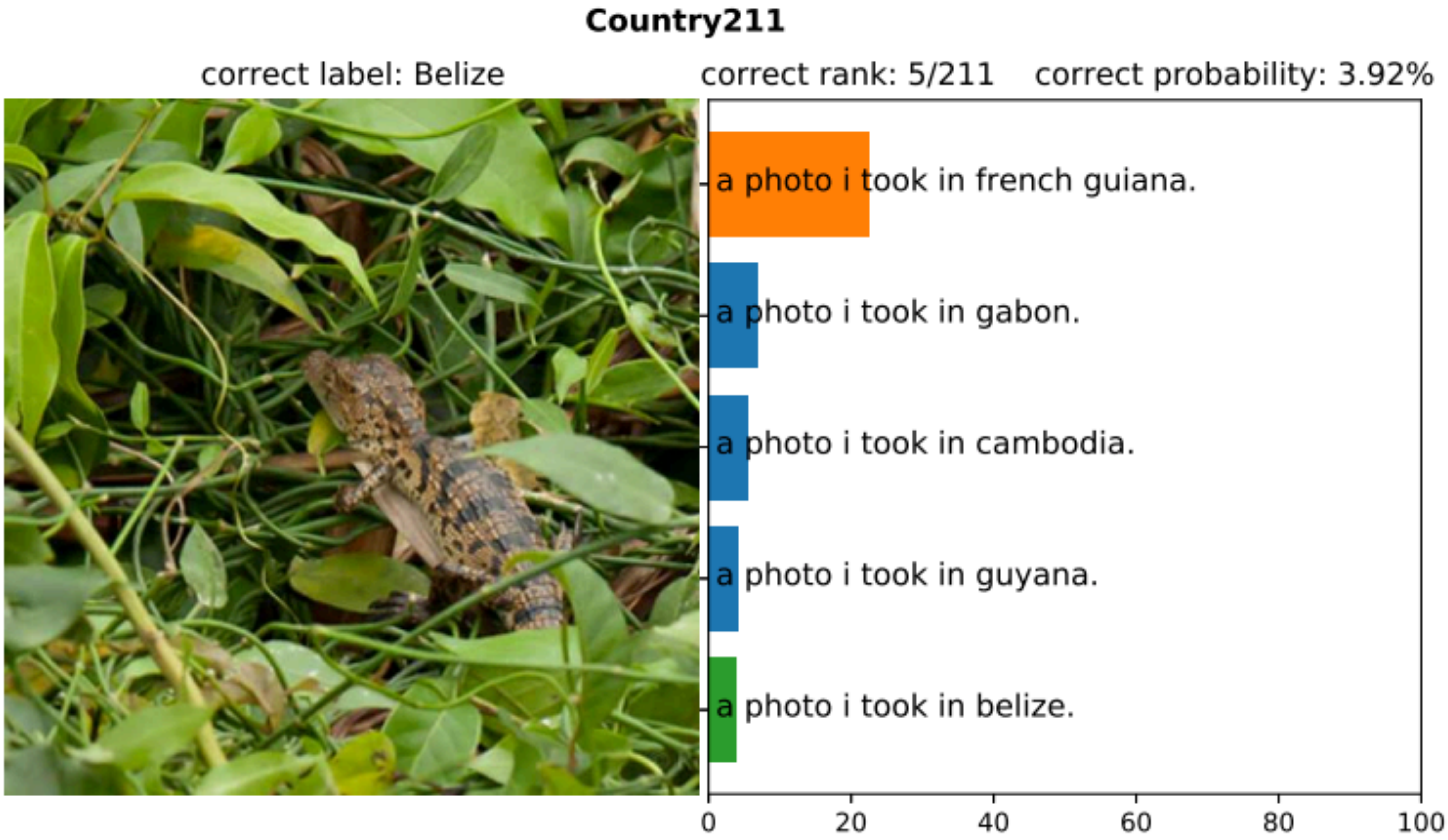


Radford et al., 2021

CLIP: Zero-shot Results



CLIP: Zero-shot Results



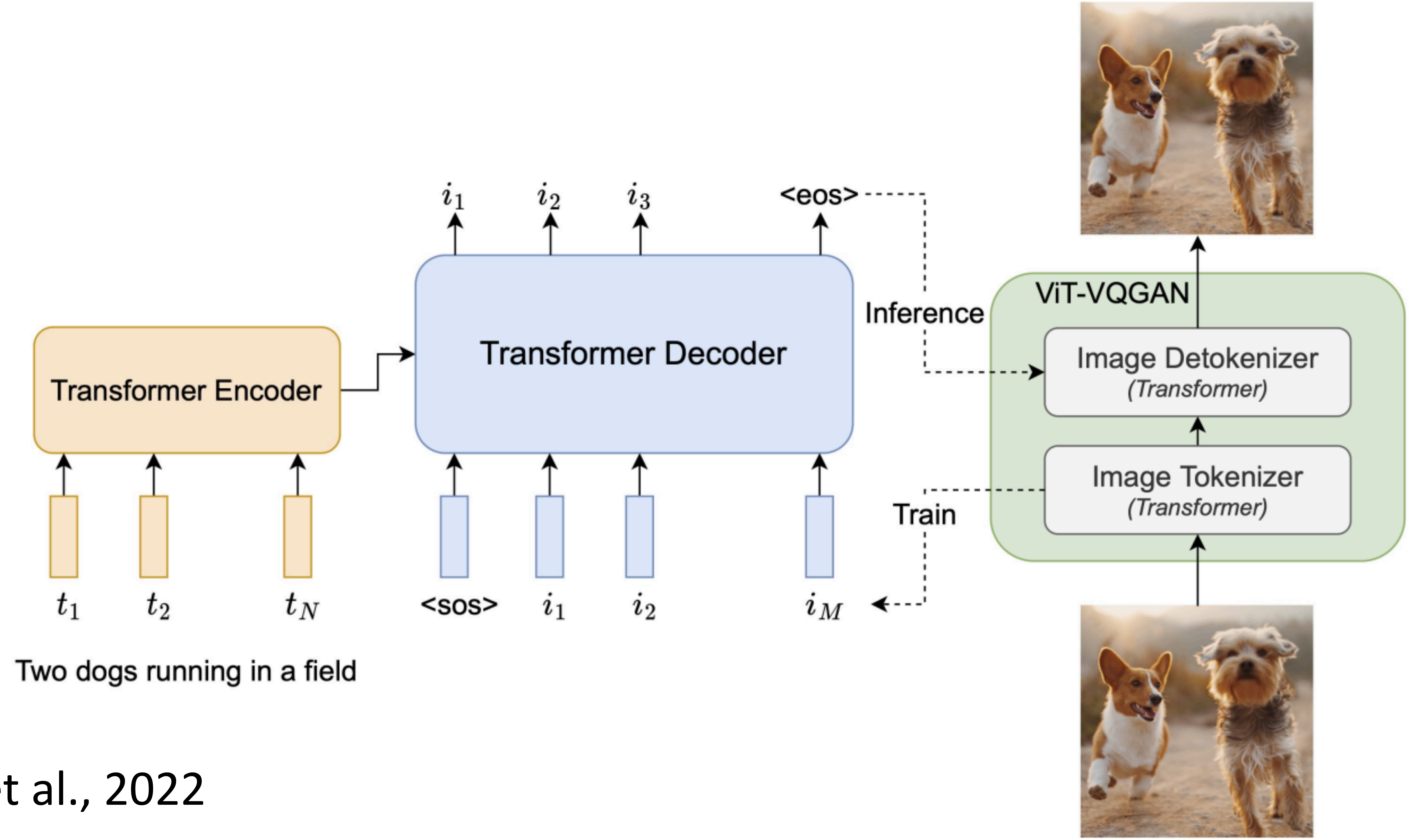
Parti

- Autoregressive text-to-image model (differs from the diffusion models you may have seen, like Stable Diffusion or DALL-E)

A. *A photo of a frog reading the newspaper named "Toaday" written on it. There is a frog printed on the newspaper too.*

Yu et al., 2022

Parti



Yu et al., 2022

Manipulation: SayCan, PaLM-E

SayCan

- Most models like CLIP are just vision+language. What about interaction with the world?

SayCan

- Probability of taking an action decomposes as follows:

$$p(c_i|i, s, \ell_\pi) \propto p(c_\pi|s, \ell_\pi)p(\ell_\pi|i)$$

p(skill possible given world state)

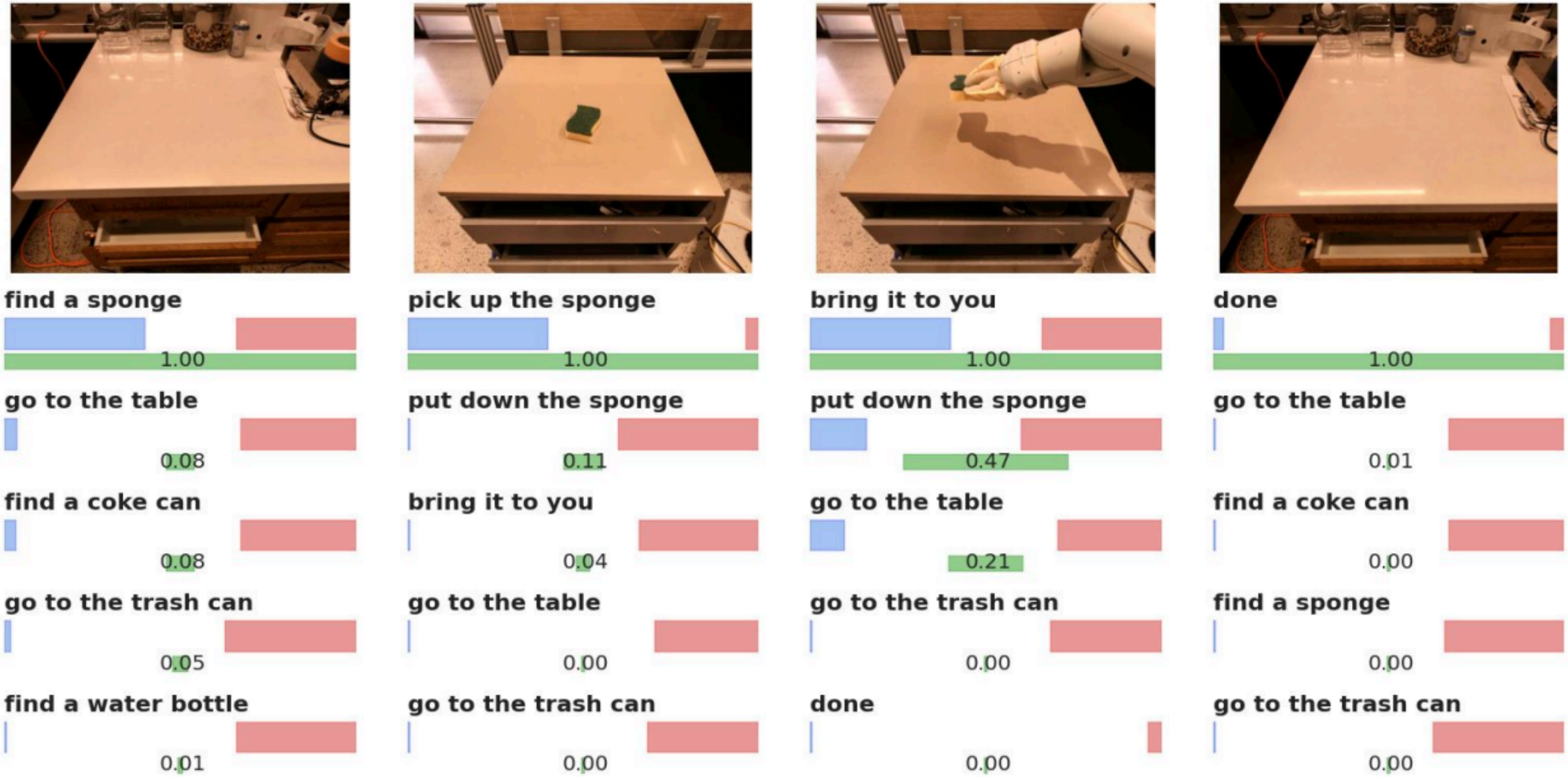
p(language description of skill | instruction)

- Individual skills are learned in advance, form affordance models for that skill
- Train a single multi-task policy that conditions on the lang description
- Do you think this is a grounded language model?

SayCan

Human: I spilled my coke, can you bring me something to clean it up?

Robot: I would
1. Find a sponge
2. Pick up the sponge
3. Bring it to you
4. Done

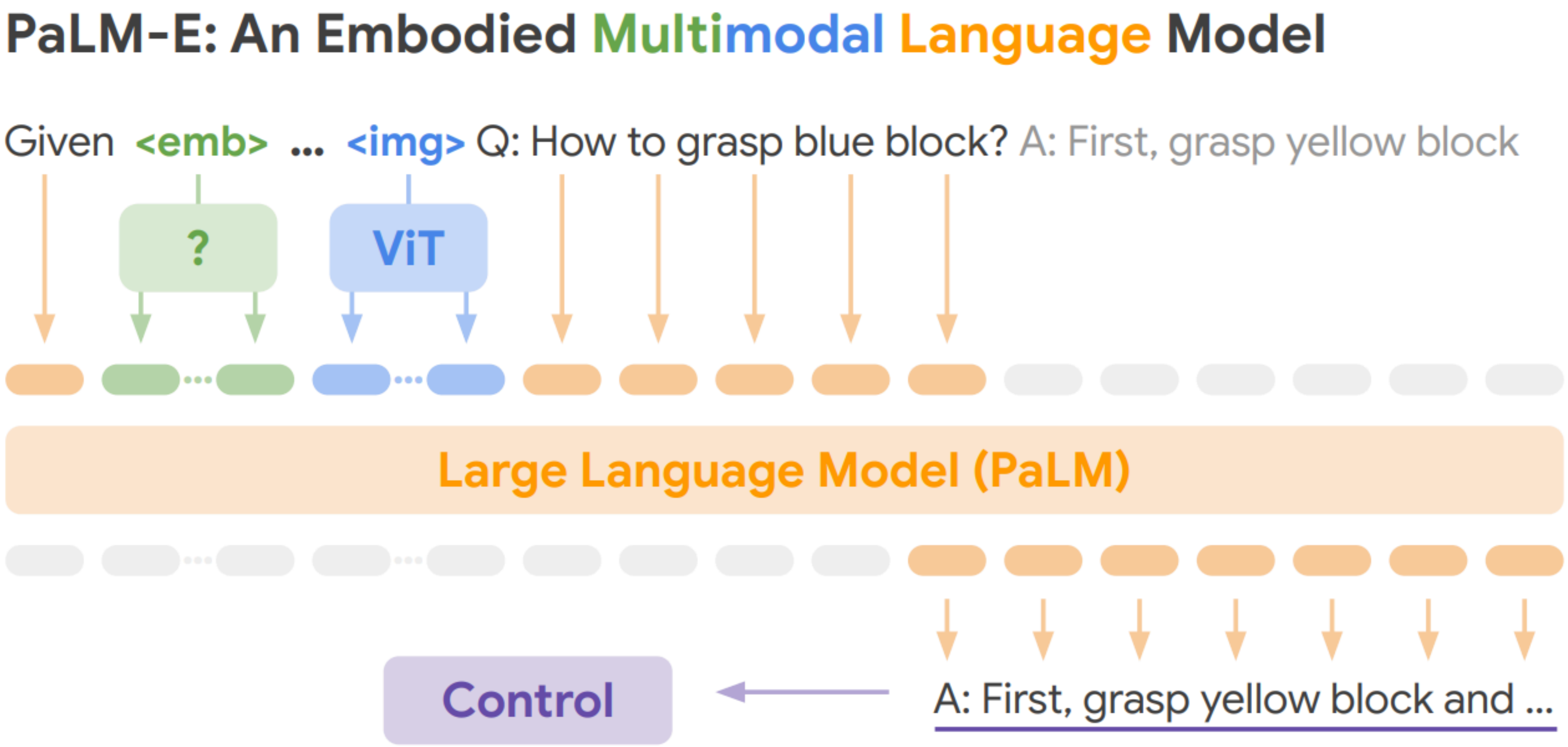


Language × Affordance

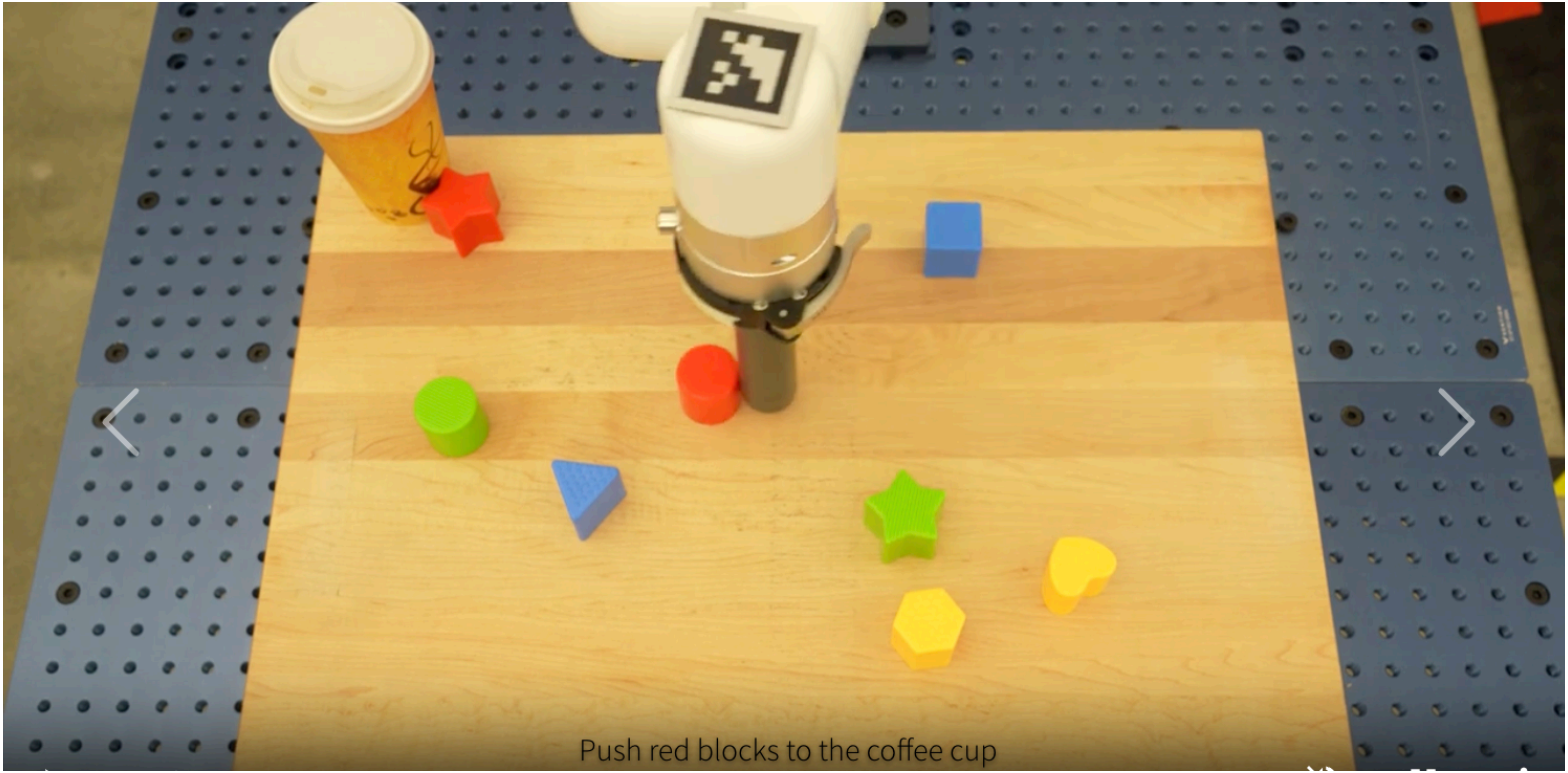
Combined Score

PaLM-E

- Most models like CLIP are just vision+language



PaLM-E



Where are we today

- Explosion of multimodal pre-training for {video, audio, images, interaction} x text

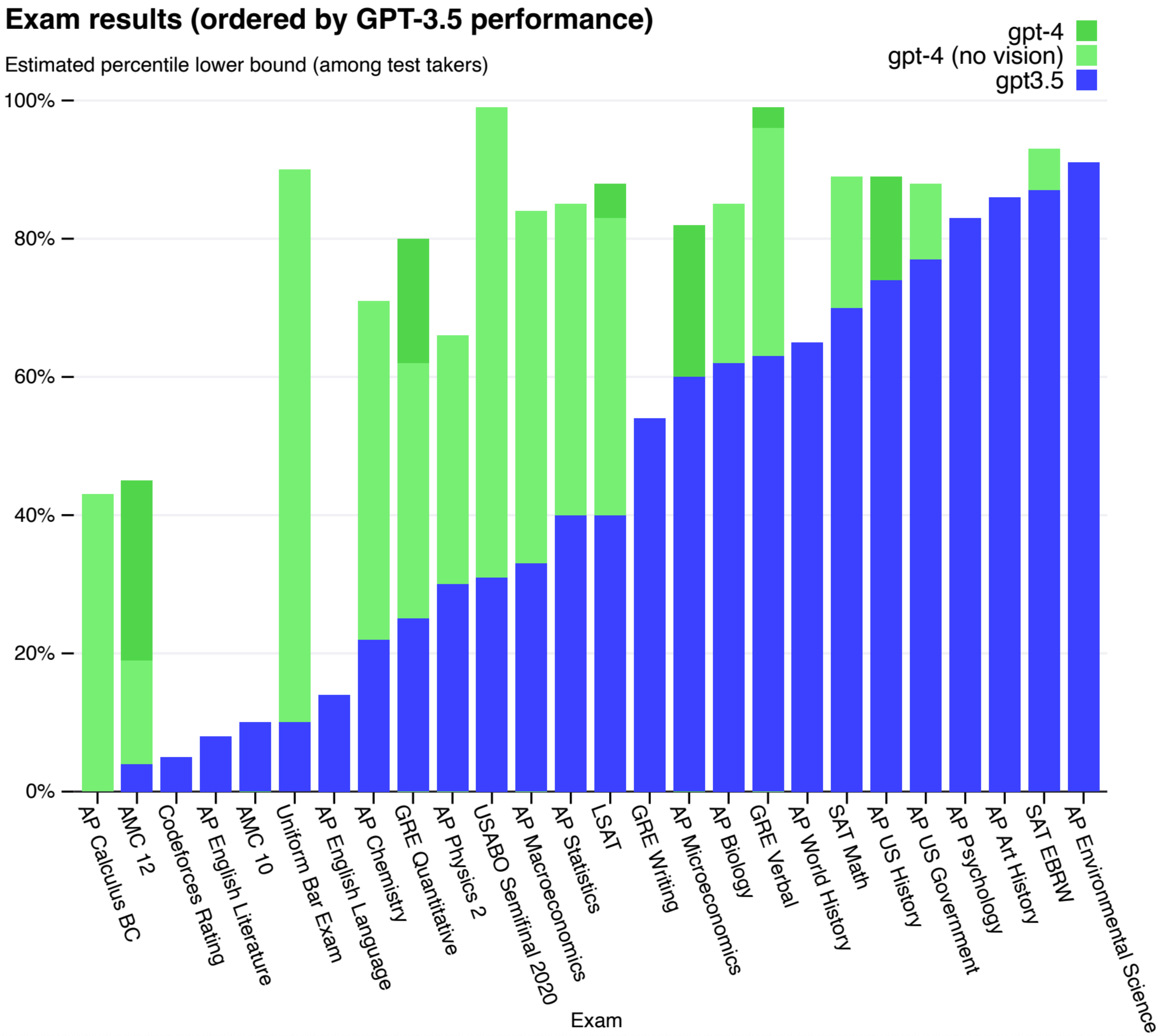
- Many of these methods are Transformer-based

- Still haven't seen large-scale multimodal pre-training of this form advance text-only tasks, but there's potential!

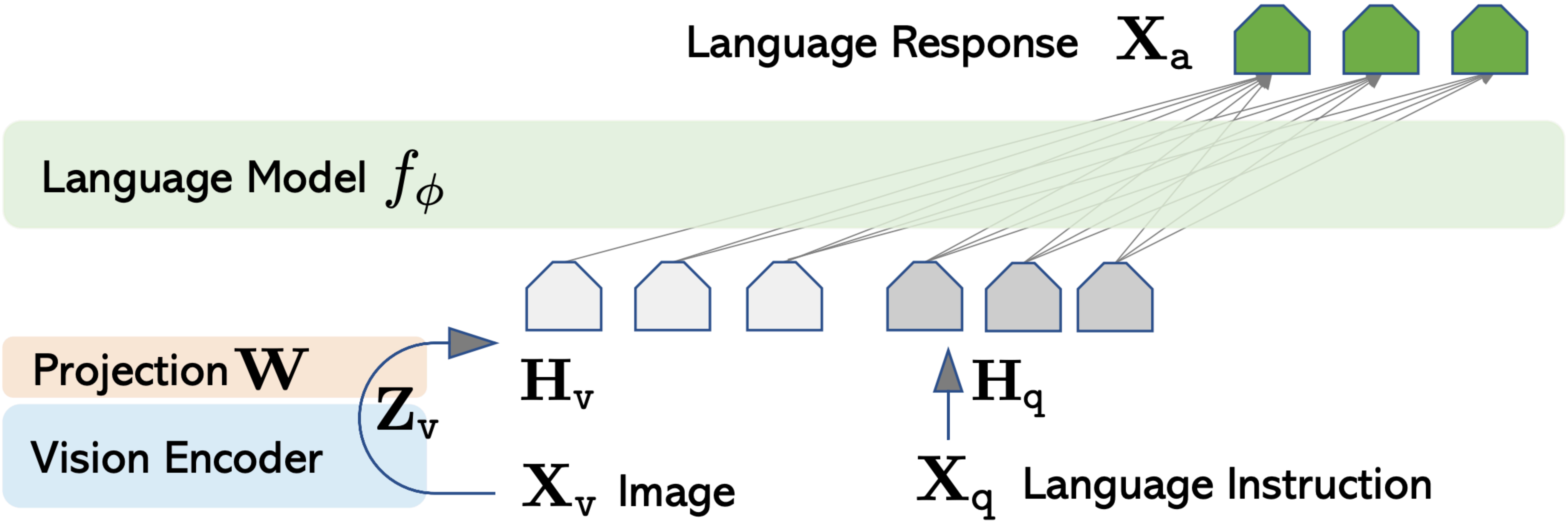
- Impact of images on GPT-4 is unclear

GPT-4

- Dark green: additional performance from vision pre-training
- This graph is hard to read and doesn't make sense…



LLaVA: Visual Instruction Tuning



Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

Haotian Liu et al., 2023

Takeaways

- Is the lack of grounding in text-only pre-trained models a problem?
- Multimodal methods can allow us to learn representations for images as well as text and provide a path towards language grounding
- Pre-training on text and other modalities is more and more common and unlocking new capabilities for models