Embodied Reasoning Through Planning with Language and Vision Foundation Models

Georgia Tech CS 7643/4644: Deep Learning

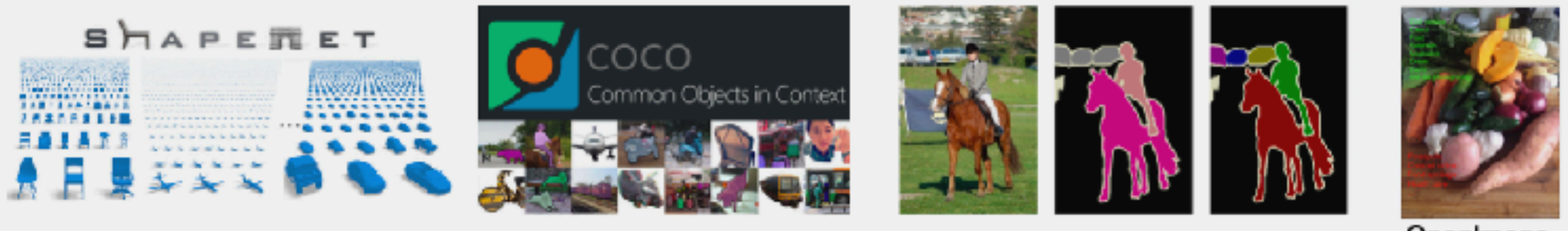
Fei Xia, Google DeepMind

11/7/2023

From “Internet AI” to “Embodied AI”

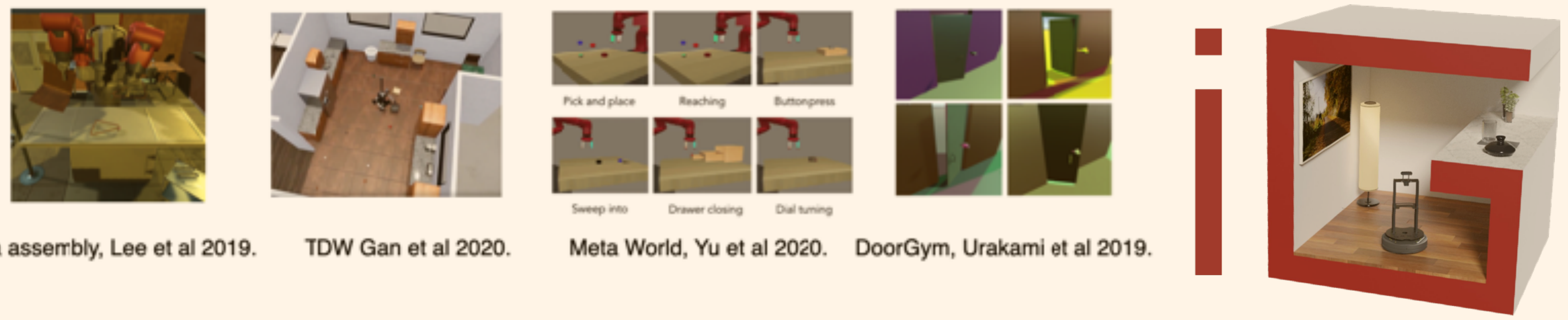
Datasets

ImageNet, Deng et al 2009. Visual Genome, Krishna et al 2017.



ShapeNet, Chang et al 2015. MS COCO, Lin et al 2014. Pascal VOC, Everingham et al 2012. OpenImage, Krasin et al 2016.

RLBench, James et al 2020. AI2Thor, Kolve et al 2017. SAPIEN, Xiang et al 2020.



Ikea assembly, Lee et al 2019. TDW Gan et al 2020. Meta World, Yu et al 2020. DoorGym, Urakami et al 2019.

Tasks

Classification

Segmentation

Detection

Generation

Captioning

...

Internet AI

Visual Navigation

Manipulation

Rearragement

Embodied-QA

Mobile Manipulation

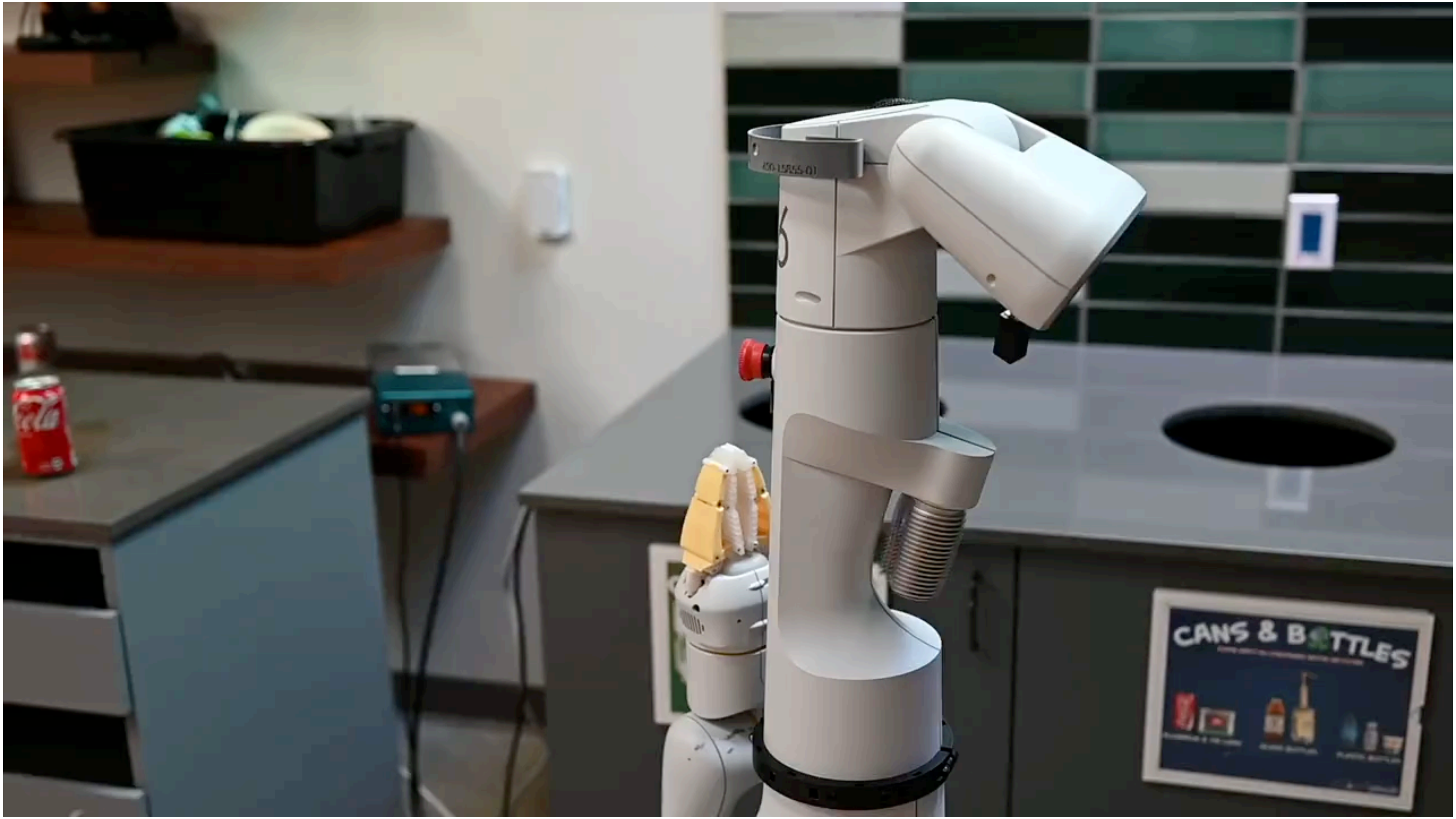
Instruction Following

...

Embodied AI

Do as I Can, Not as I Say (SayCan): Grounding Language In Robotic Affordances

Say-Can.github.io


CANS & BOTTLES

How do we make robot learning more useful?

I spilled my drink, can you help?

I just worked out, can you bring me a
snack and a drink to recover?

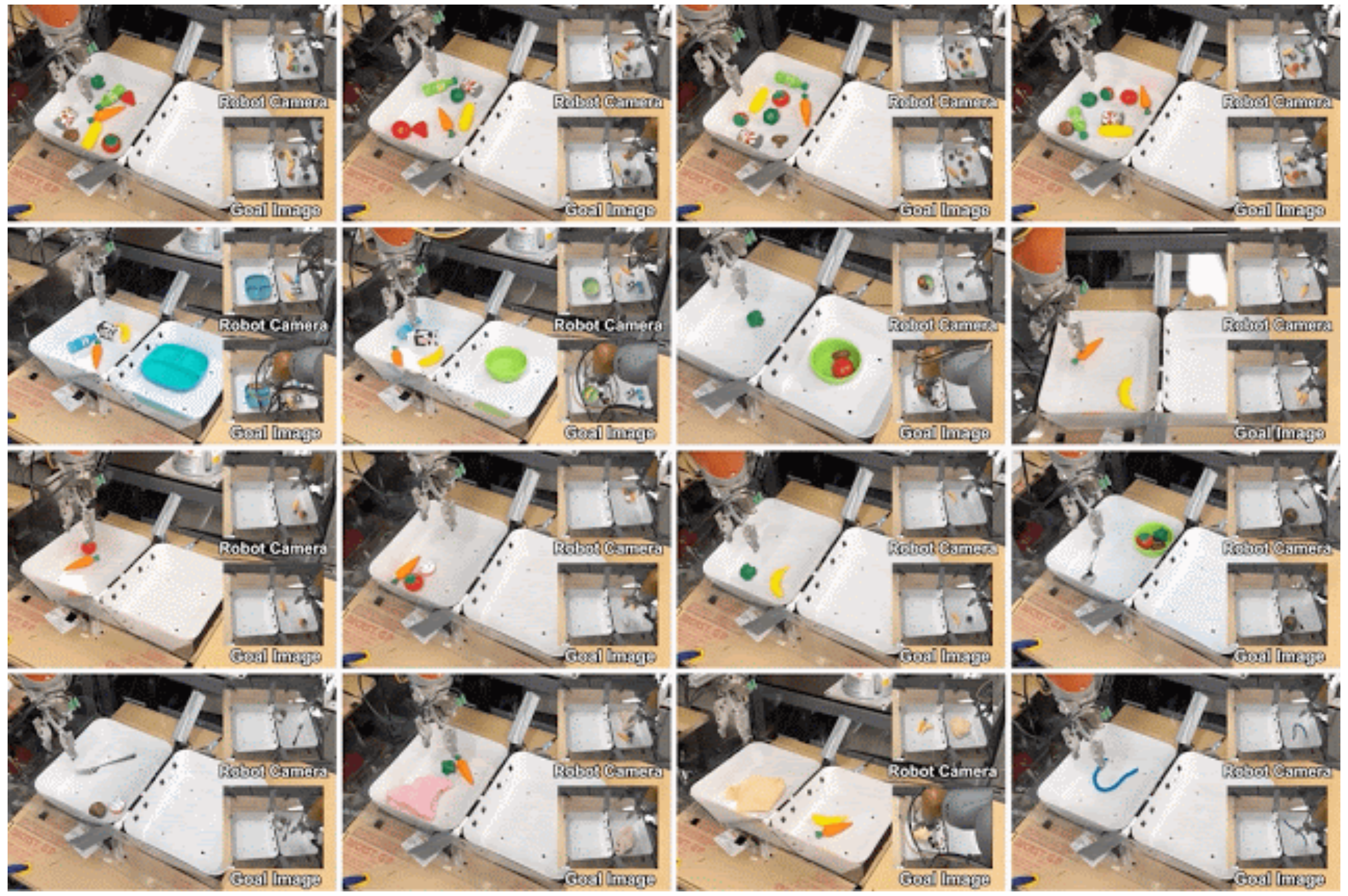
I’m feeling tired, can you make me a latte?

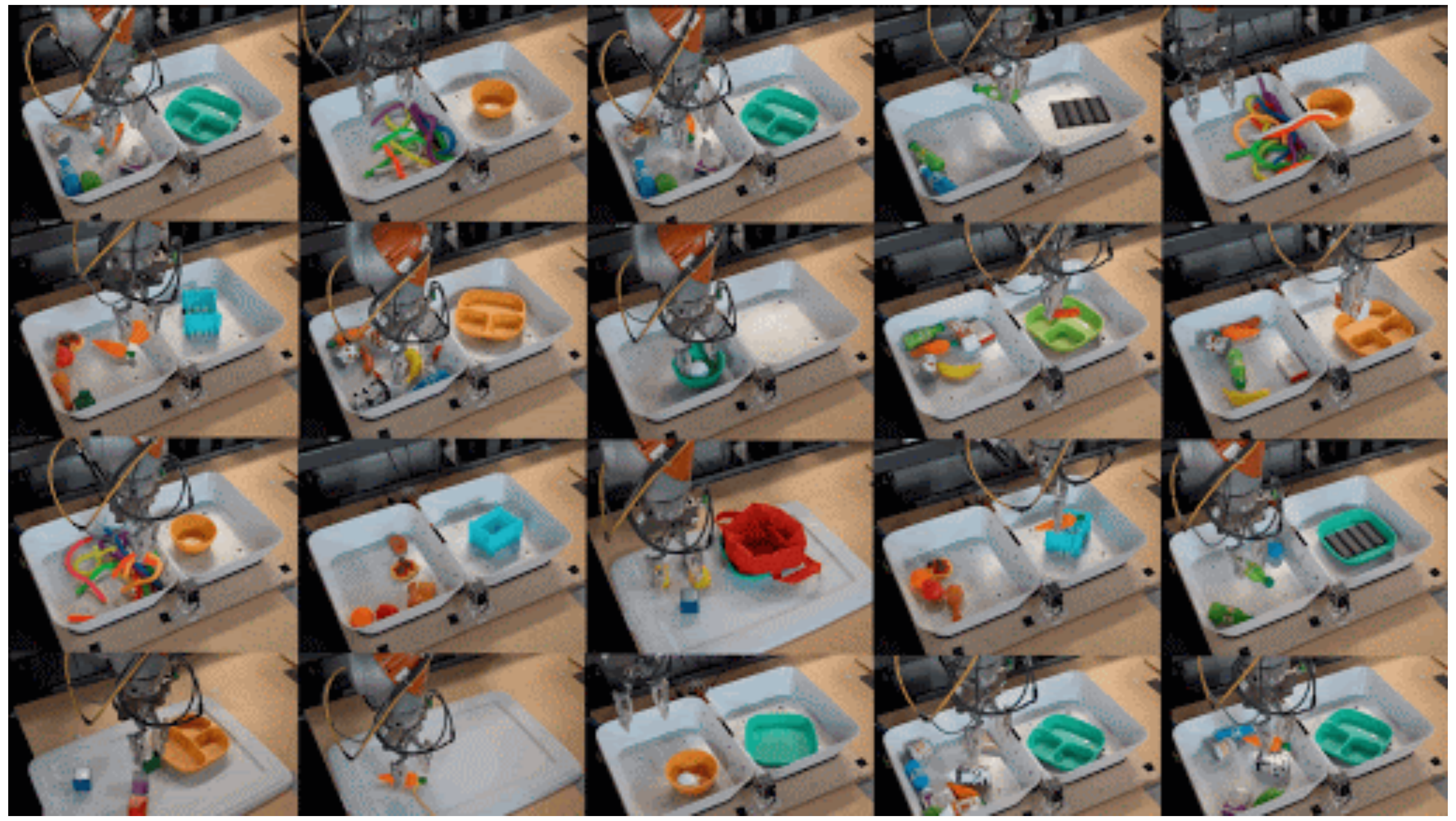
How do we make robot learning more useful?

I spilled my drink, can you help?

I just worked out, can you bring me a snack and a drink to recover?

I’m feeling tired, can you make me a latte?

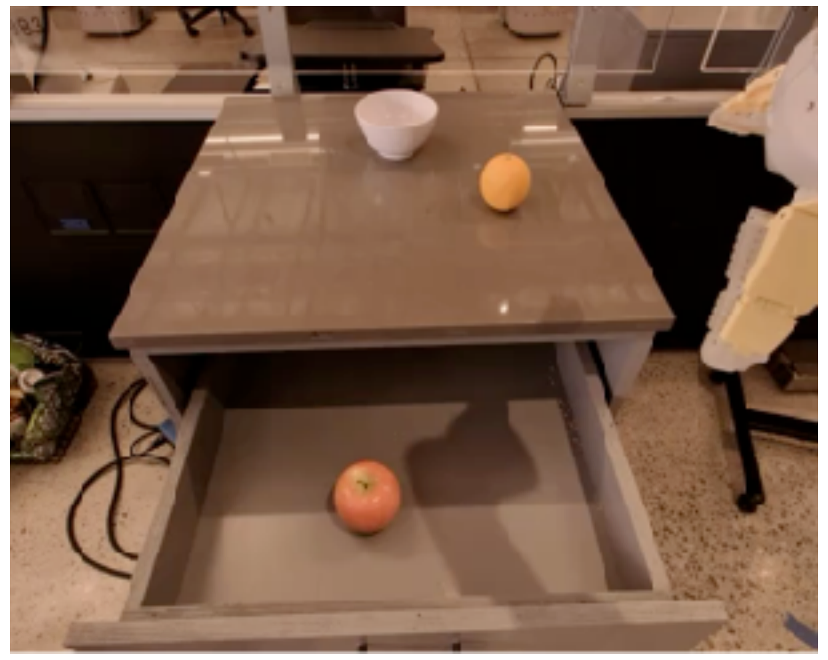




Language Conditioned Robot Behavior

Naive language conditioned imitation learning works on short horizon tasks but struggles with long-horizon tasks and complex instructions.

“I spilled my drink, can you help with that?”



“I just worked out, can you bring me a snack and a drink to recover?”

[BC-Z, E. Jang et al, 2021]

Mixing language and robotics

Large Language Models (LLMs)

Lots of recent breakthroughs, contain a wealth of knowledge and can handle sequences, memory, and more

User	What can you tell me about cell biology?
Gopher	Sure, I can tell you some basic facts. The nucleus is the central part c DNA. Each chromosome contains many genes. Cells divide when the duplicated before this happens. A group of similar cells is called a tis
User	What is the technical name for single-cell organisms?
Gopher	They're known as prokaryotes.
User	Are humans prokaryotes?
Gopher	No, they're eukaryotes.
User	What's an example of a prokaryote?
Gopher	Well, bacteria and archaea are prokaryotes.
User	What's a common species of bacteria that lives in the gut?
Gopher	Well, I believe it's Escherichia coli.
User	Is E. Coli always dangerous?
Gopher	Not always - but in some cases, it can be pathogenic. That means it
User	Which strains are dangerous?
Gopher	The most dangerous ones are serotype O157:H7 and serotype O111:H.
User	Can you give me a link where I can learn more?
Gopher	Yes, here you go: https://en.wikipedia.org/wiki/Escherichia_coli_O157

Explaining a joke

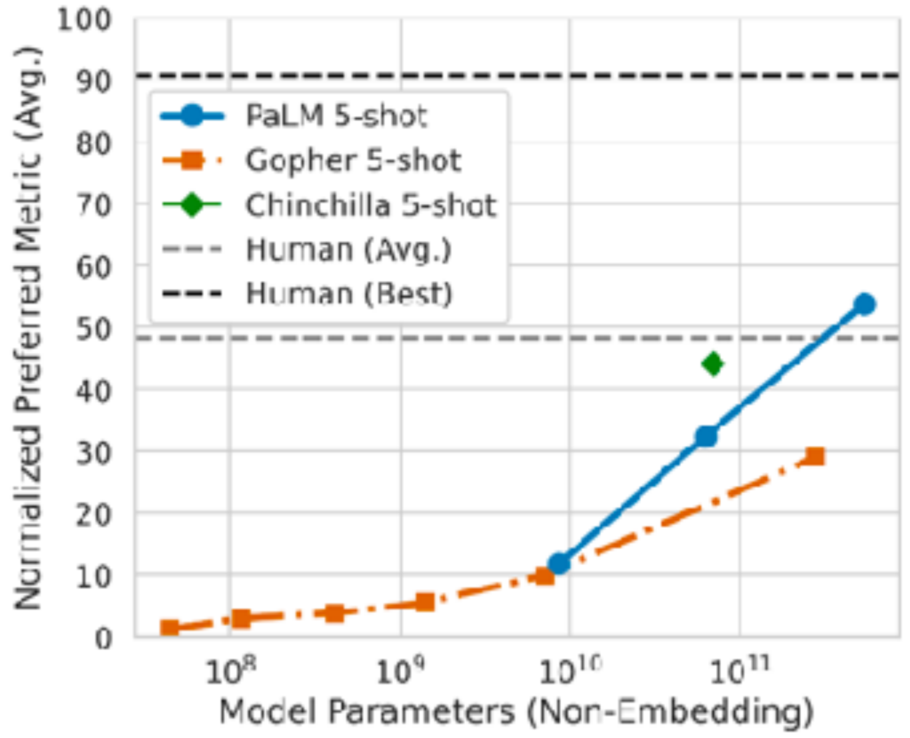
Prompt

```
Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for
their TPU team? It showed them how to communicate between two
different pods!
```

Model Response

```
Prediction: TPUs are a type of computer chip that Google uses
for deep learning. A "pod" is a group of TPUs. A "pod" is also a
group of whales. The joke is that the whale is able to
communicate between two groups of whales, but the speaker is
pretending that the whale is able to communicate between two
groups of TPUs.
```



PaLM,
Chowdhery et al, 2022

LLMs for robotics

Challenges:

1. Robot Language: Our robots can only do a fixed number of commands and need the problem broken down in actionable steps. This is not what LLMs have seen.

2. Grounding: LLMs have not directly “experienced” the physical world.

3. Safety, alignment, interpretability…

I spilled my drink, can you help?

I just worked out, can you bring me a snack and a drink to recover?

I’m feeling tired, can you make me a latte?

LLMs for robotics

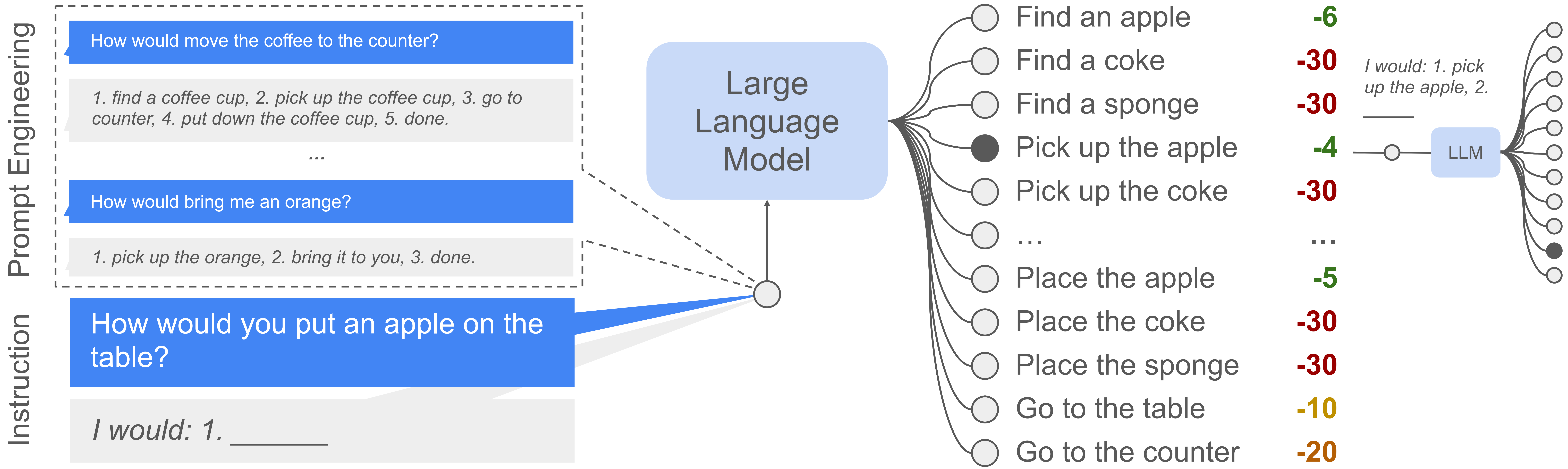
I spilled my drink, can you help?

GPT3	You could try using a vacuum cleaner.
LaMDA	Do you want me to find a cleaner?
FLAN	I'm sorry, I didn't mean to spill it.

Problem: Our robots can only do a fixed number of commands and need the problem broken down in actionable steps. This is not what LLMs have seen.

We need to get LLMs to speak “robot language”!

LLMs for robotics



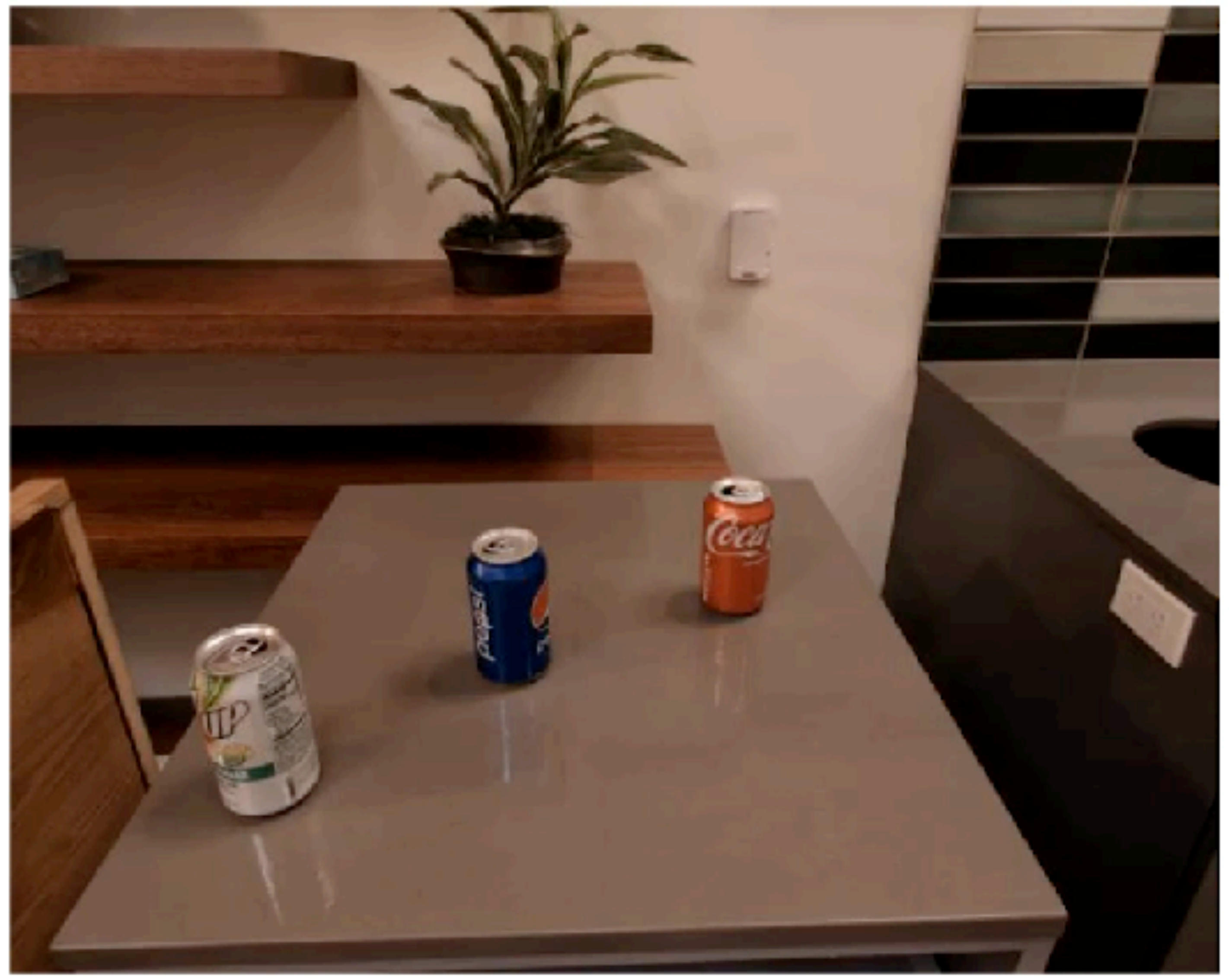
Problem: LLMs aren't grounded in the real-world. They don't know what's actually possible from a state with a given embodiment.

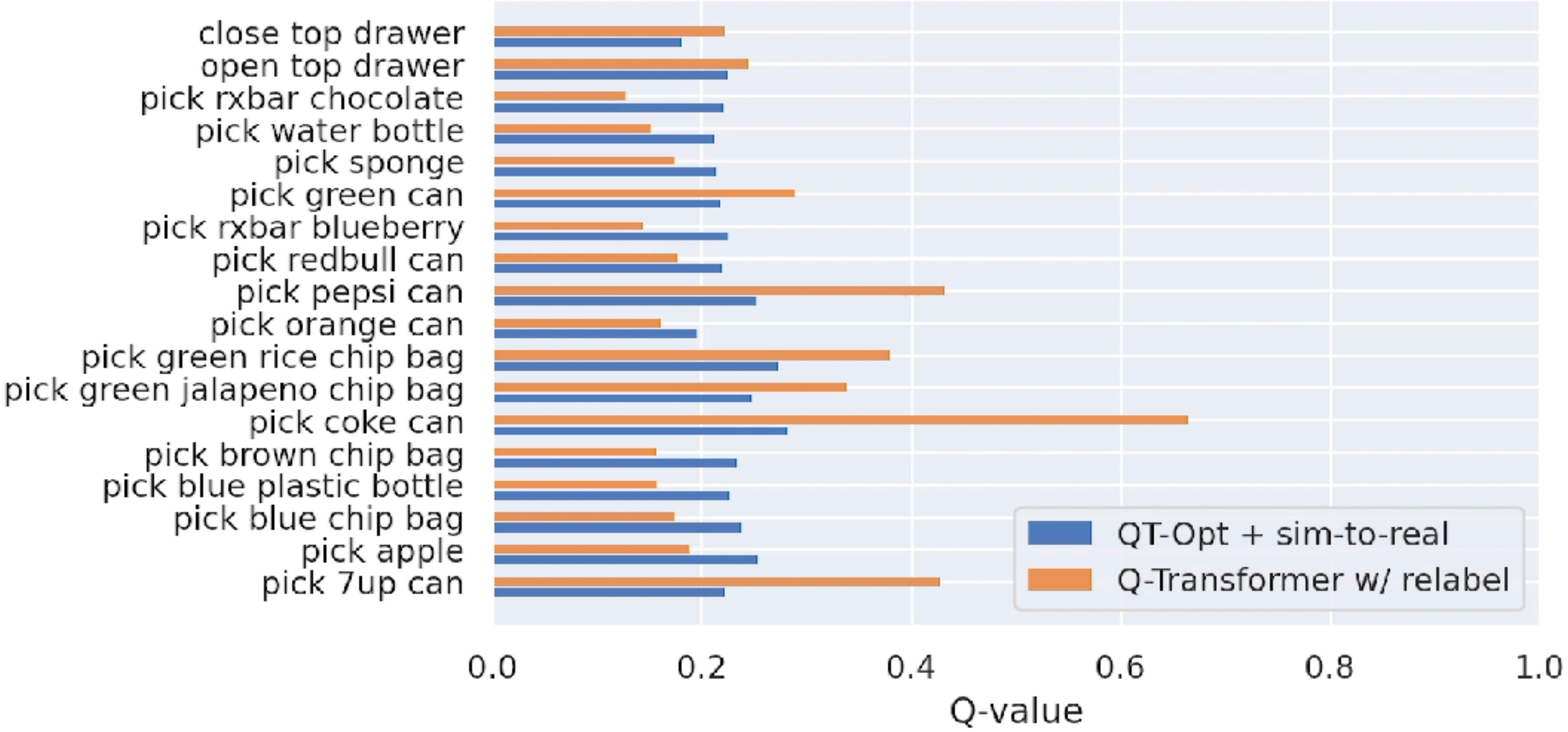
We need to ground LLMs in robotic affordances!

Robotic affordances

Reinforcement learning already provides task-based affordances.

They are encoded in the value function!

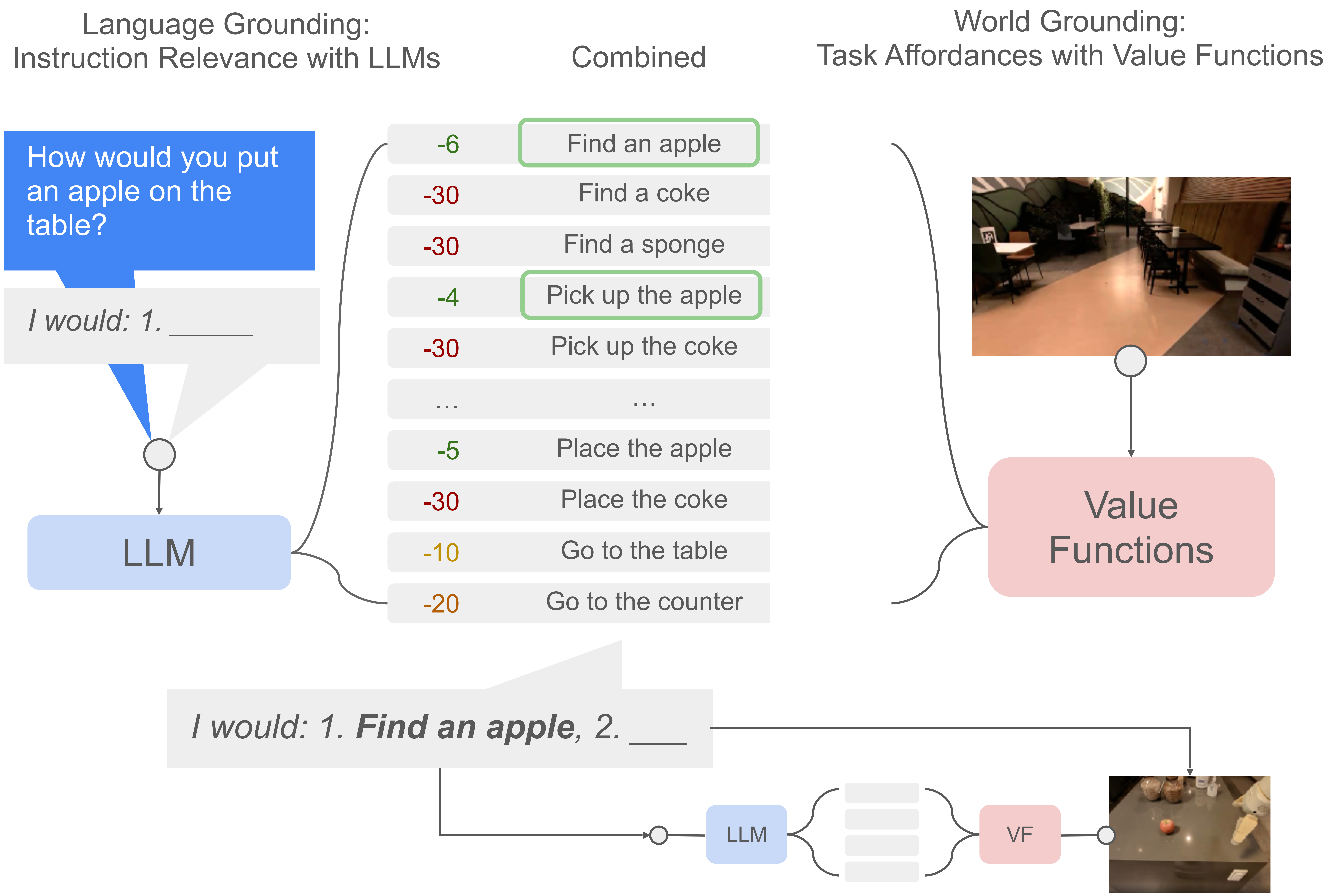




[Value Function Spaces, Shah, Xu, Lu, Xiao, Toshev, Levine, Ichter, ICLR 2022]

Q-Transformer, 2023.

LLMs for robotics and robotics for LLMs



Experiment Overview

70% planning rate

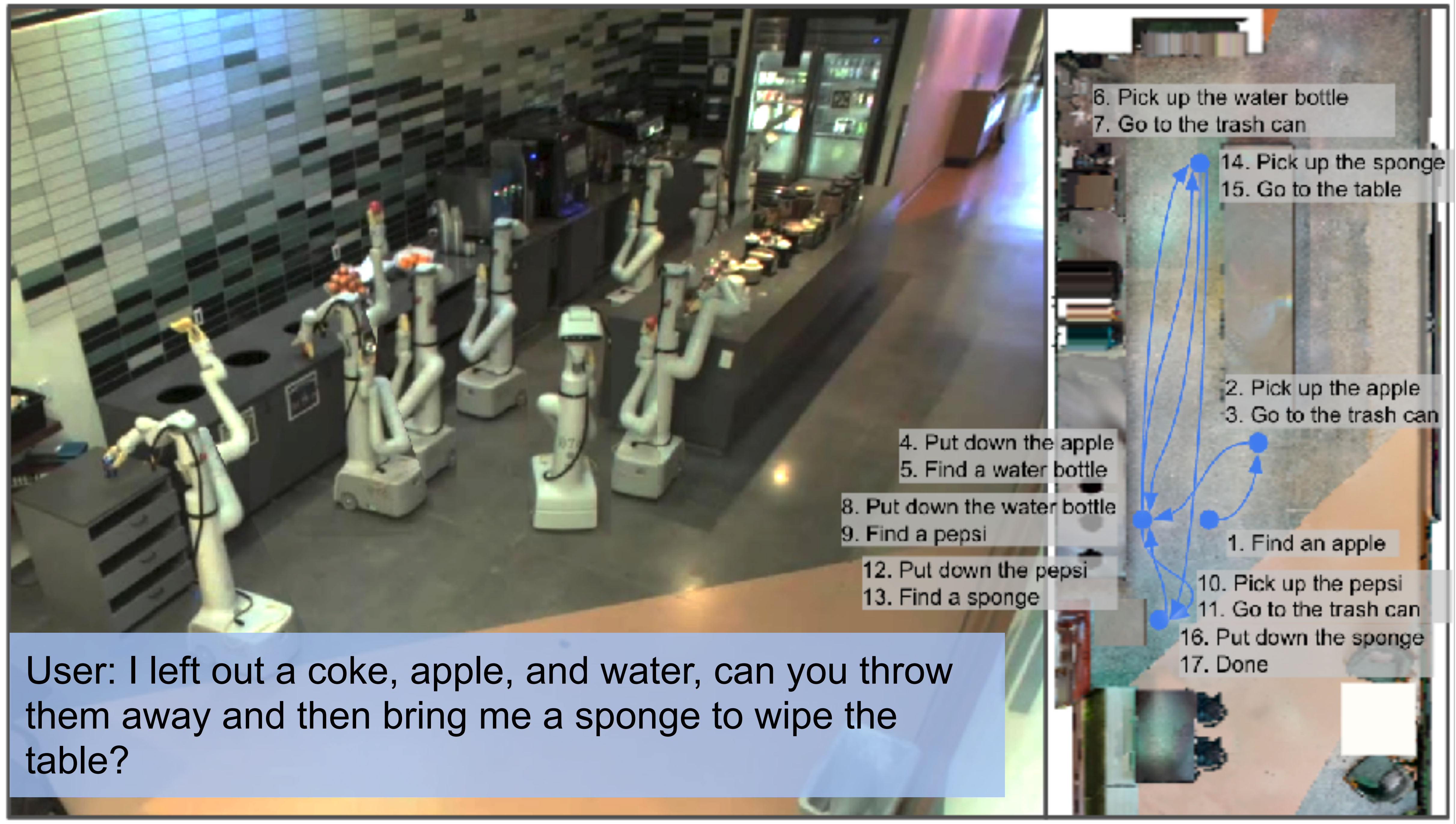
61% execution rate

101 long-horizon instructions

10+ navigation and manipulation skills in a row

Without grounding nearly halves performance

Instruction Family	Num	Plan	Execution
Natural Language Single Primitive	15	67%	67%
Natural Language Nouns	15	60%	53%
Natural Language Verbs	15	80%	67%
Structured Language	15	100%	87%
Embodiment	11	64%	55%
Crowd Sourced	15	73%	67%
Long-Horizon	15	47%	33%
Total	**101**	**70%**	**61%**

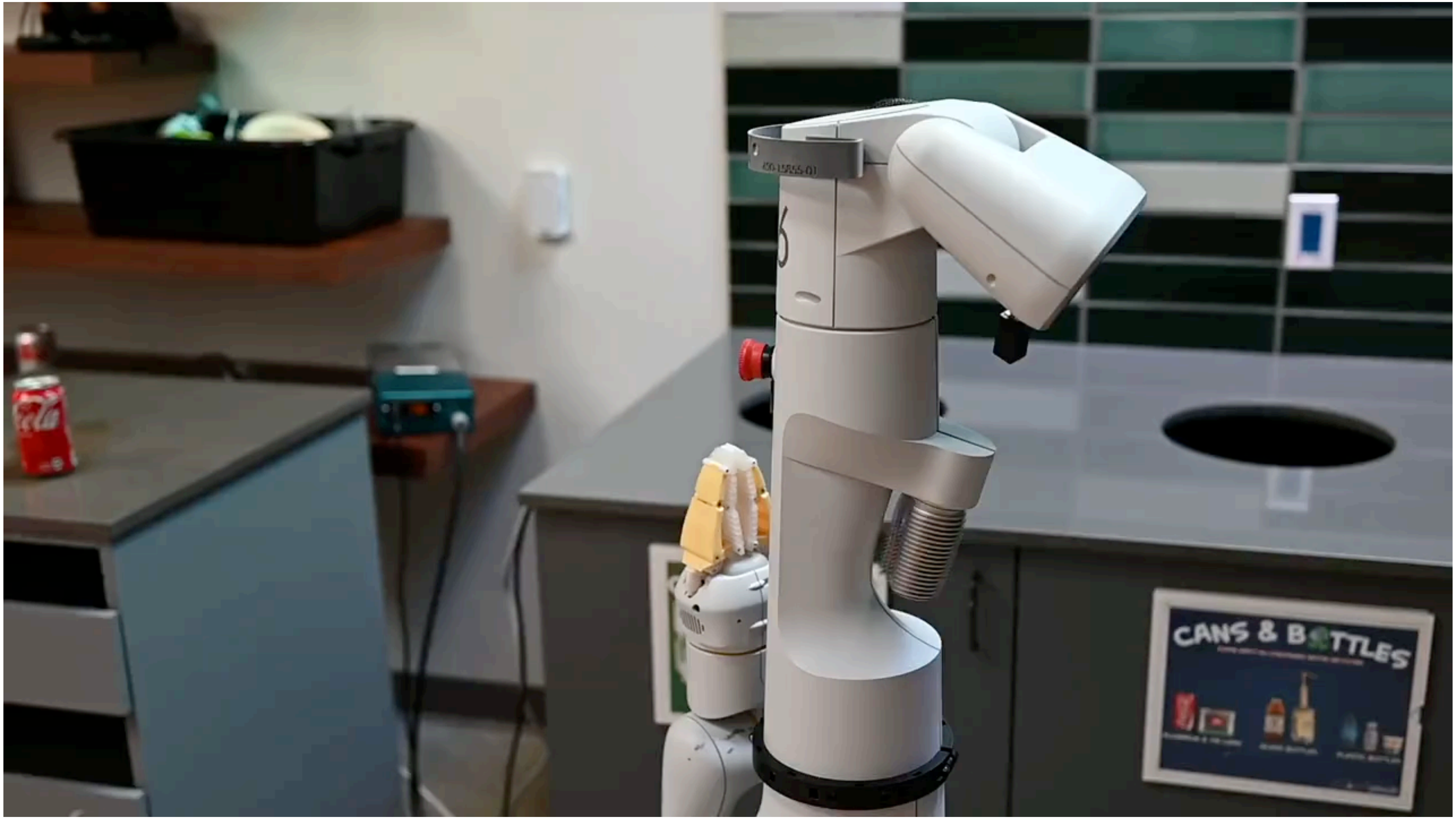


PaLM-SayCan vs FLAN-SayCan

		plan	execute	plan	execute
PALM		**PaLM**		**FLAN**	
Family	Num	SayCan	SayCan	SayCan	SayCan
NL Single	15	100%	100%	67%	67%
NL Nouns	15	67%	47%	60%	53%
NL Verbs	15	100%	93%	80%	67%
Structured	15	93%	87%	100%	87%
Embodiment	11	64%	55%	64%	55%
Crowd Sourced	15	87%	87%	73%	67%
Long-Horizon	15	73%	47%	47%	33%
Total	101	**84%**	**74%**	**70%**	**61%**

+14% Planning success rate overall
+26% Planning success rate on long-horizon tasks

SayCan: Grounding Language in Robotic Affordances



SayCan: Grounding Language in Robotic Affordances

SayCan: Takeaways

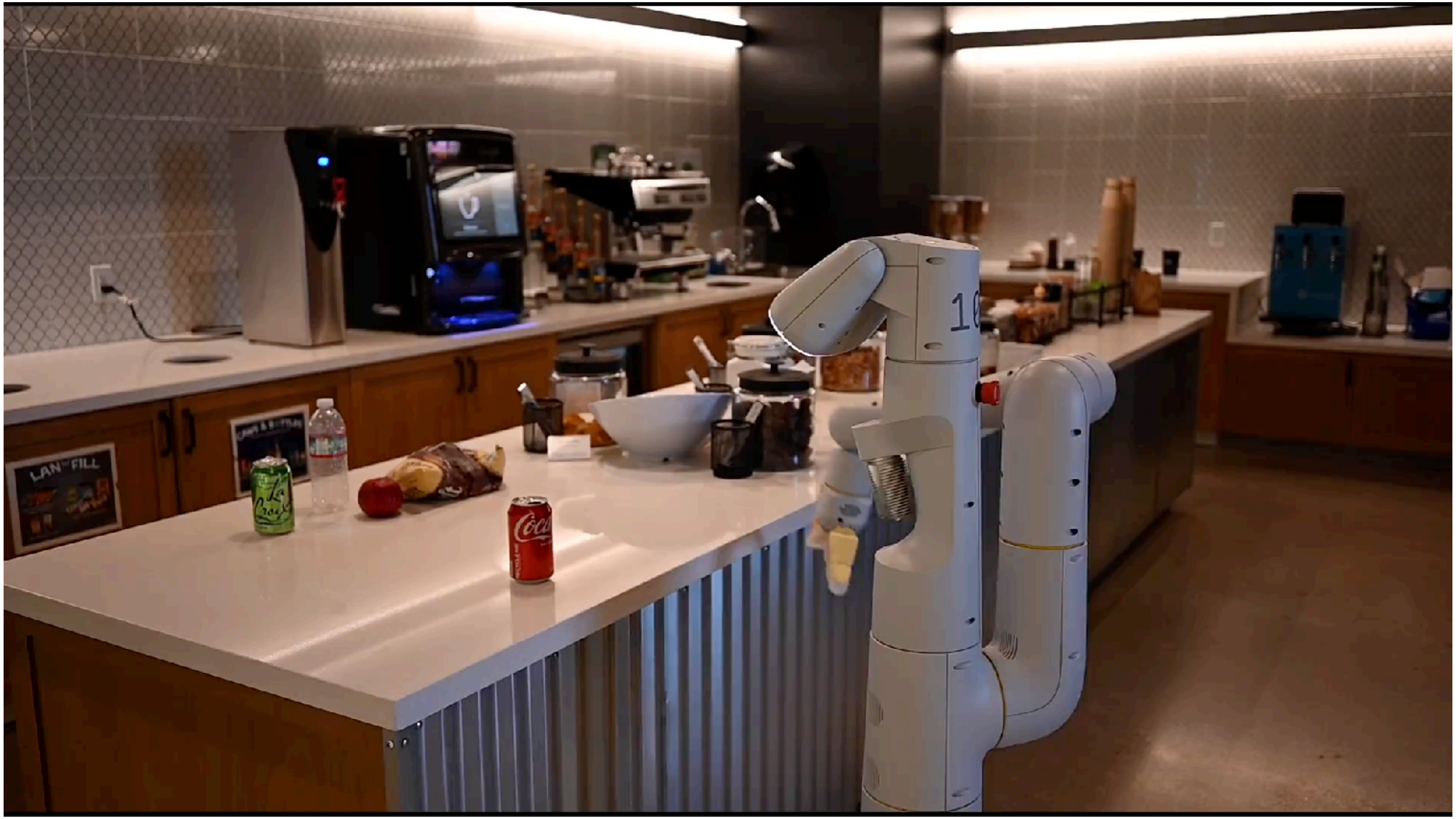
- LLMs can provide task grounding
- (Robotic) value functions provide real-world grounding
- This is compatible with any policy as long as there is an affordance

Challenge:

- One bottleneck is still on the skills
- Language-conditioned affordance model

RT-1: Robotics Transformer for Real-World Control at Scale

ROSIE: Scaling Robot Learning with Semantically Imagined Experience



Discussions

PaLM-E: An Embodied Multimodal Language Model

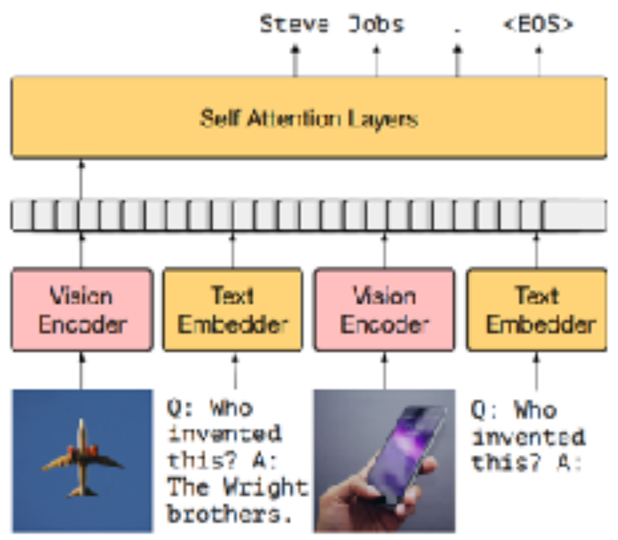
Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, Pete Florence

Google Research

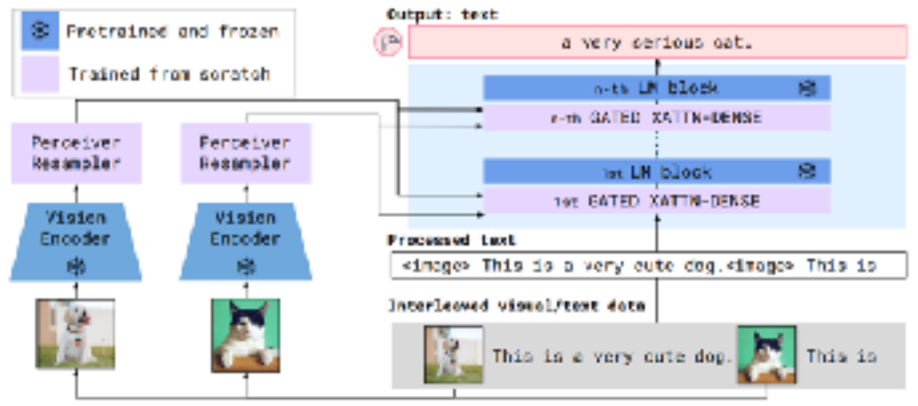
“One model”
- Embodied robotics tasks
- Vision-language
- Language
- *… across multiple robot embodiments*
- *… across multiple modalities (vision, states, neural scenes)*
Positive transfer
PaLM-E: An Embodied Multimodal Language Model
Given <emb> … <img> Q: How to grasp blue block? A: First, grasp yellow block
?
ViT
Large Language Model (PaLM)
Control
A: First, grasp yellow block and …
Vision-language generalist
Given <img>. Q: What's in the image? Answer in emojis.
A: 🍏🍌🍇🍐🍑🍈🍒.
Emergent visual-language capabilities
Zero-shot multimodal CoT, multi-image reasoning
Given <img>. Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step. 1. do not enter. 2. except bicycles. 3. do not entry except bicycles. 4. yes.
DO NOT ENTER
EXCEPT
Closed-loop end-to-end planning
(“Given <img>... Bring me the rice chips from the drawer ”)
Long-horizon tasks
(“Given <img>.... Sort the blocks by colors into corners”)
Zero-shot generalization
(unseen object pairings, or objects)

Multimodal Language Models

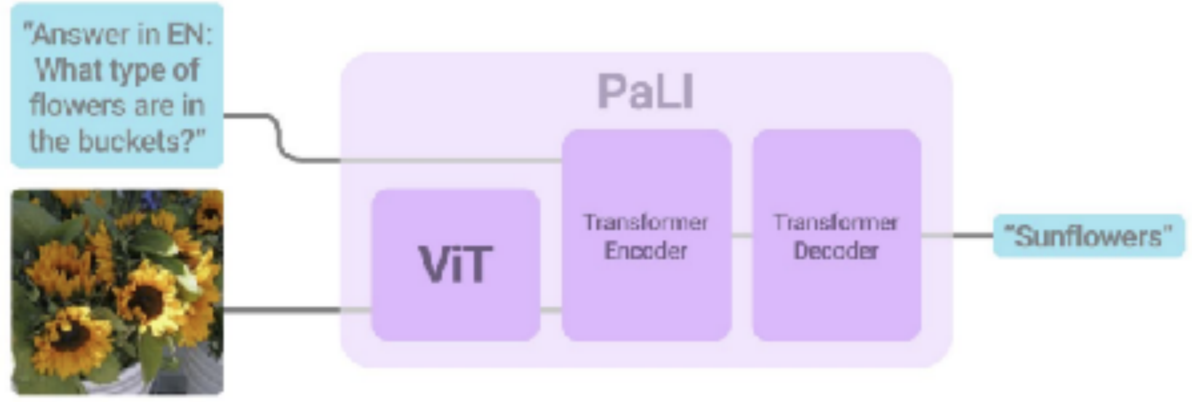
- “Frozen”, Tsimpoukelli et al.



- Flamingo, Alayrac et al.



- PaLI, Chen et al.

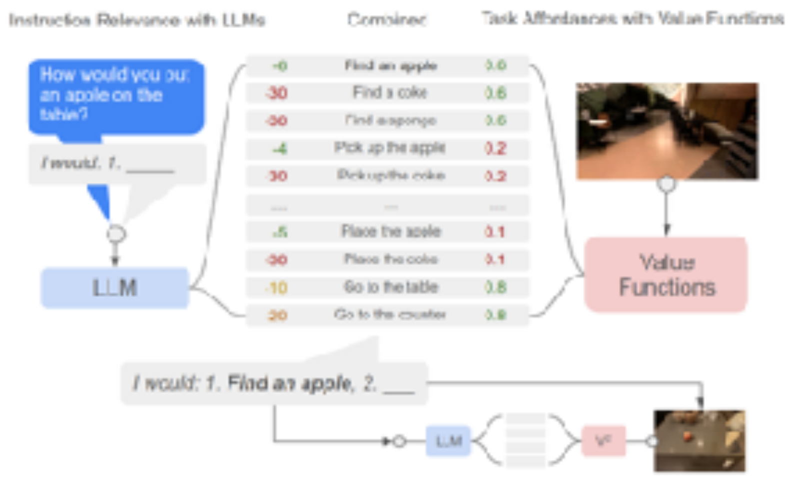


- BLIP-2, Kosmos-1, GPT-4, …

Language + Robotics

LLMs + robots for high-level planning

SayCan



Socratic Models

InnerMonologue

text only LLM

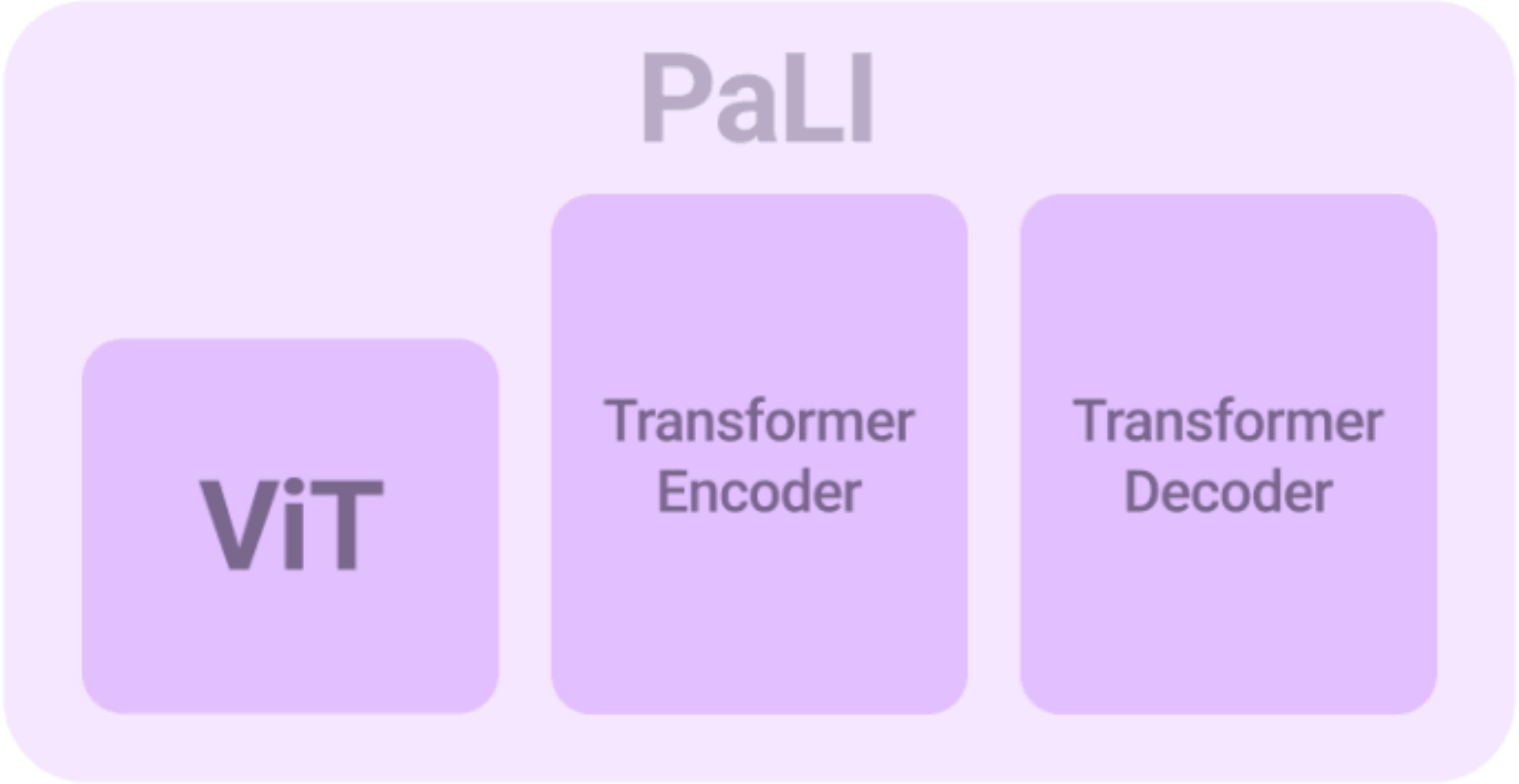
PaLM-E

multimodal LLM

Language conditioned policies

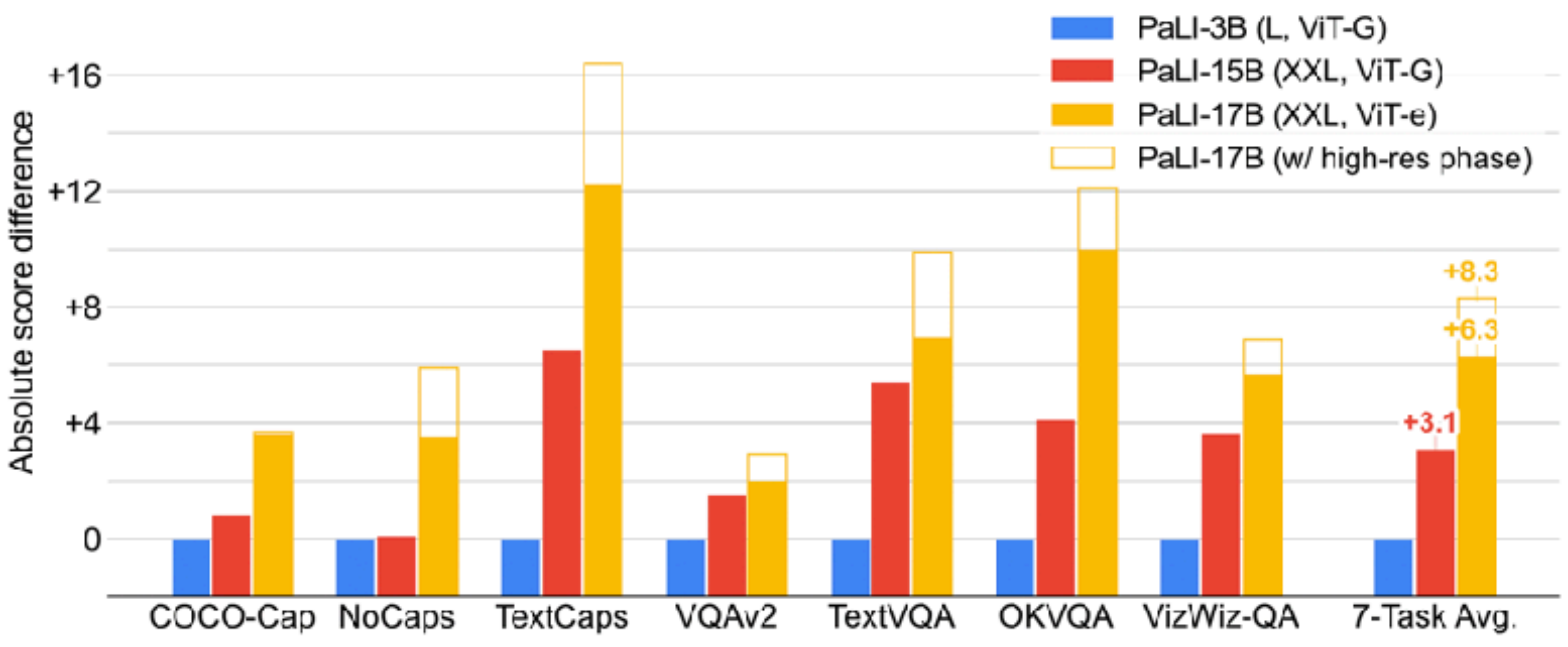
- Interactive Language, Lynch et al.
- RT-1, Brohan et al.

PaLI (Google 2022)



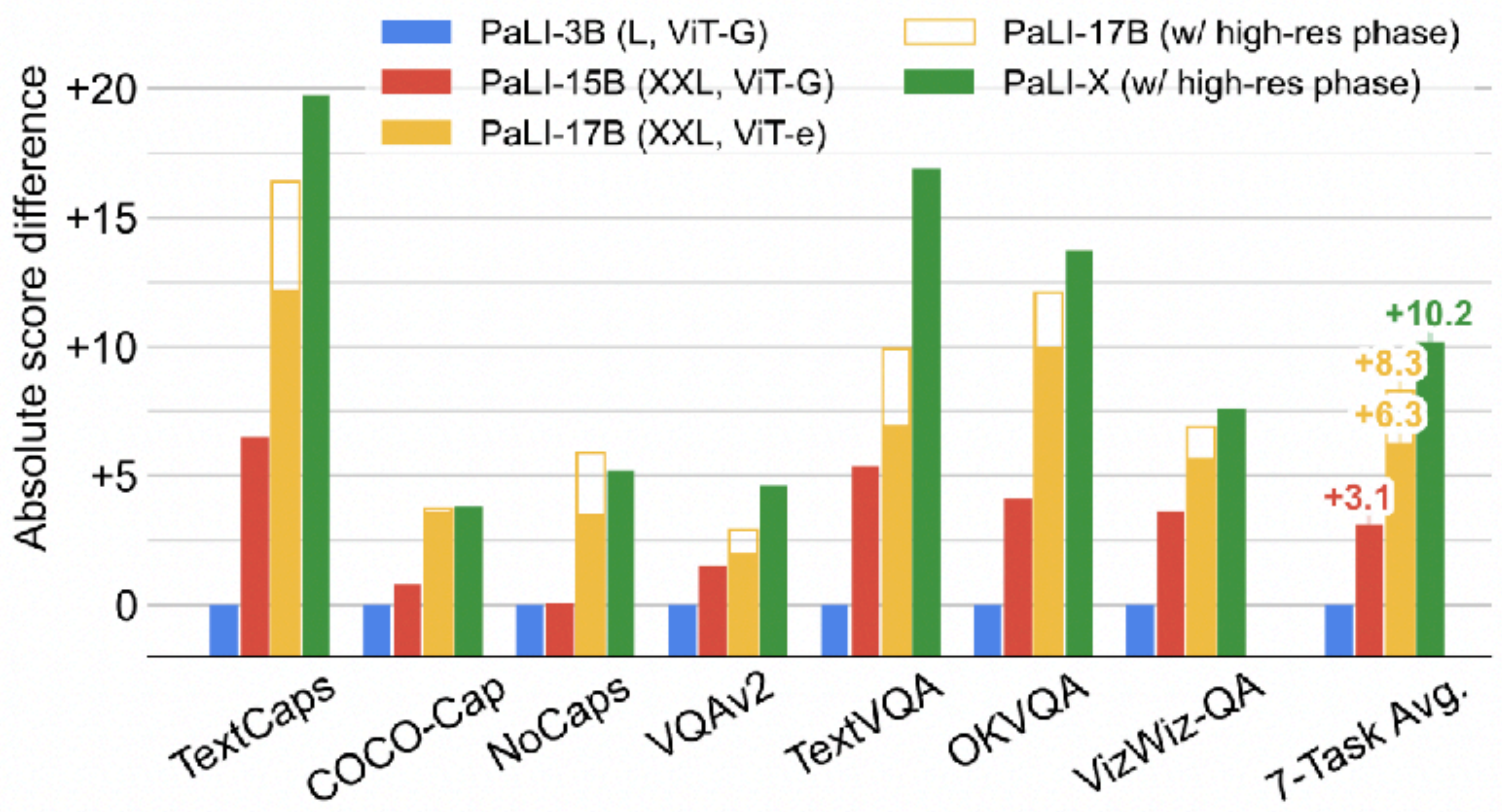
PaLI (Google 2022)

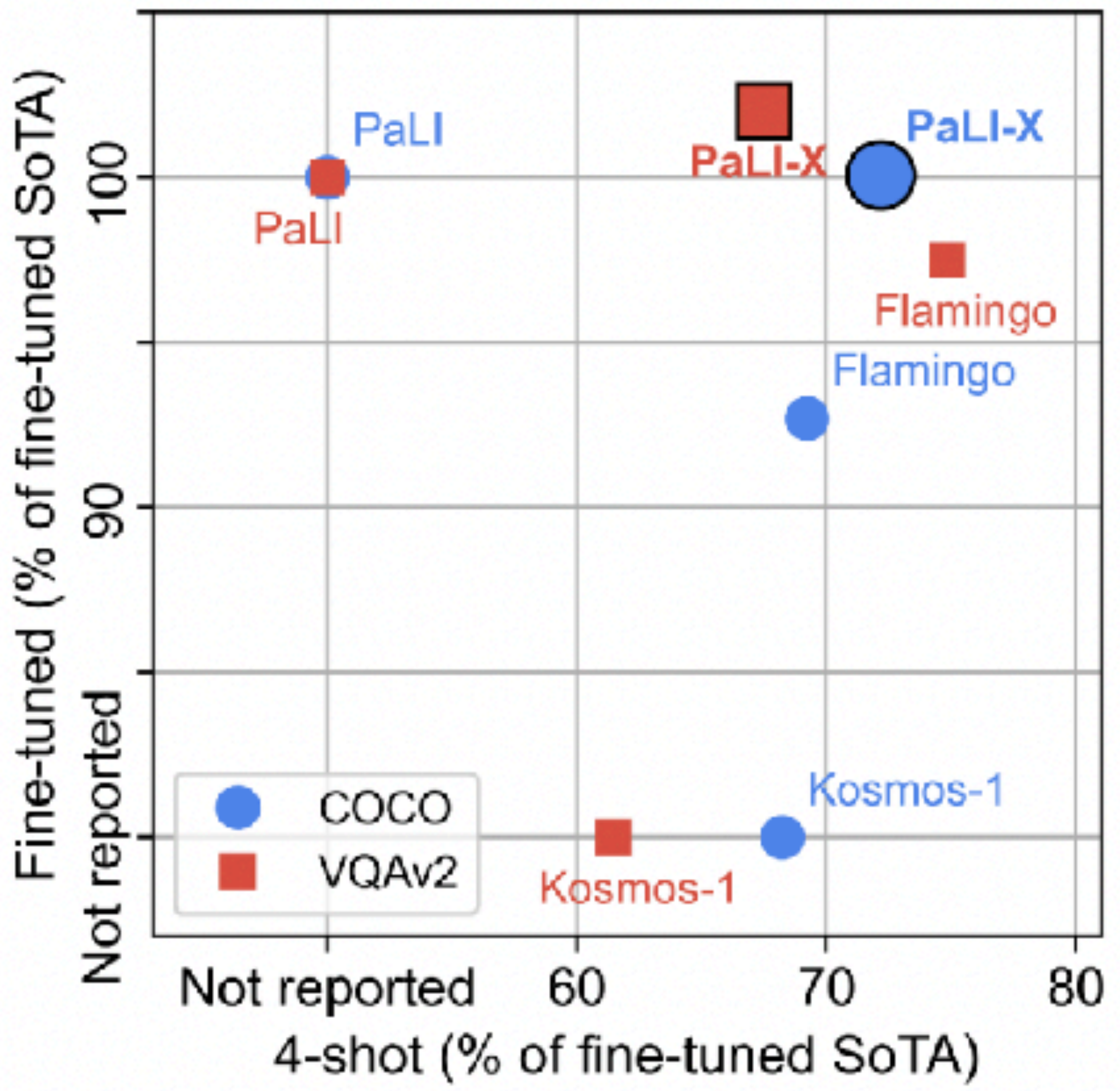
Method	VQAv2		OKVQA	TextVQA		VizWiz-QA		ST-VQA	
	test-dev	test-std	val	val	test	test-dev	test	val	test
SimVLM	80.03	80.34	-	-	-	-	-	-	-
CoCa (2.1B)	82.3	82.3	-	-	-	-	-	-	-
GIT (0.7B)	78.56	78.81	-	59.93	59.75	68.0	67.5	69.1	69.6
GIT2 (5.1B)	81.74	81.92	-	68.38	67.27	70.97	70.1	75.1	75.8
OFA (0.9B)	82.0	82.0	-	-	-	-	-	-	-
Flamingo (80B)	82.0	82.1	57.8*	57.1	54.1	65.7	65.4	-	-
BEiT-3 (1.9B)	84.2	84.0	-	-	-	-	-	-	-
KAT	-	-	54.4	-	-	-	-	-	-
Mia	-	-	-	-	73.67†	-	-	-	-
PaLI-3B	81.4	-	52.4	60.12	-	67.5	-	67.5	69.7
PaLI-15B	82.9	-	56.5	65.49	-	71.1	-	73.2	76.5
PaLI-17B	**84.3**	**84.3**	**64.5**	**71.81**	73.06	**74.4**	**73.3**	**77.1**	**79.9**



Model	COCO	NoCaps		TextCaps		VizWiz-Cap	
	Karpathy-test	val	test	val	test	test-dev	test-std
LEMON (0.7B)	139.1	117.3	114.3	-	-	-	-
SimVLM	143.3	112.2	110.3	-	-	-	-
CoCa (2.1B)	143.6	122.4	120.6	-	-	-	-
GIT (0.7B)	144.8	125.5	123.4	143.7	138.2	113.1	114.4
GIT2 (5.1B)	145.0	**126.9**	**124.8**	148.6	145.0	119.4	120.8
OFA (0.9B)	145.3	-	-	-	-	-	-
Flamingo (80B)	138.1	-	-	-	-	-	-
BEiT-3 (1.9B)	147.6	-	-	-	-	-	-
PaLI-3B	145.4	121.1	-	143.6	-	117.2	-
PaLI-15B	146.2	121.2	-	150.1	-	121.7	-
PaLI-17B	**149.1**	**127.0**	124.4	**160.0**	**160.4**	**123.0**	**124.7**

PaLI-X (Google 2023)





Credits: Watermelon/Cat; Sarah Pflug (burst), Bowls; ariesandrea (flickr), Wall; Matthew Henry (burst)

Figure 2: Examples demonstrating multilingual, OCR and other capabilities transferred to detection.

	LVIS AP	LVIS AP_{Rare}
ViLD [74] (tuned on non-rare LVIS)	29.3	26.3
Region-CLIP [75] (tuned on non-rare LVIS)	32.3	22.0
OwLViT-L/16 [28] (tuned on non-rare LVIS)	34.7	25.6
OwLViT-L/16 [28] (with Object365 and VG datasets)	34.6	31.2
PaLI-X (Zeroshot)	12.36	12.16
PaLI-X (Detection-tuned)	30.64	31.42

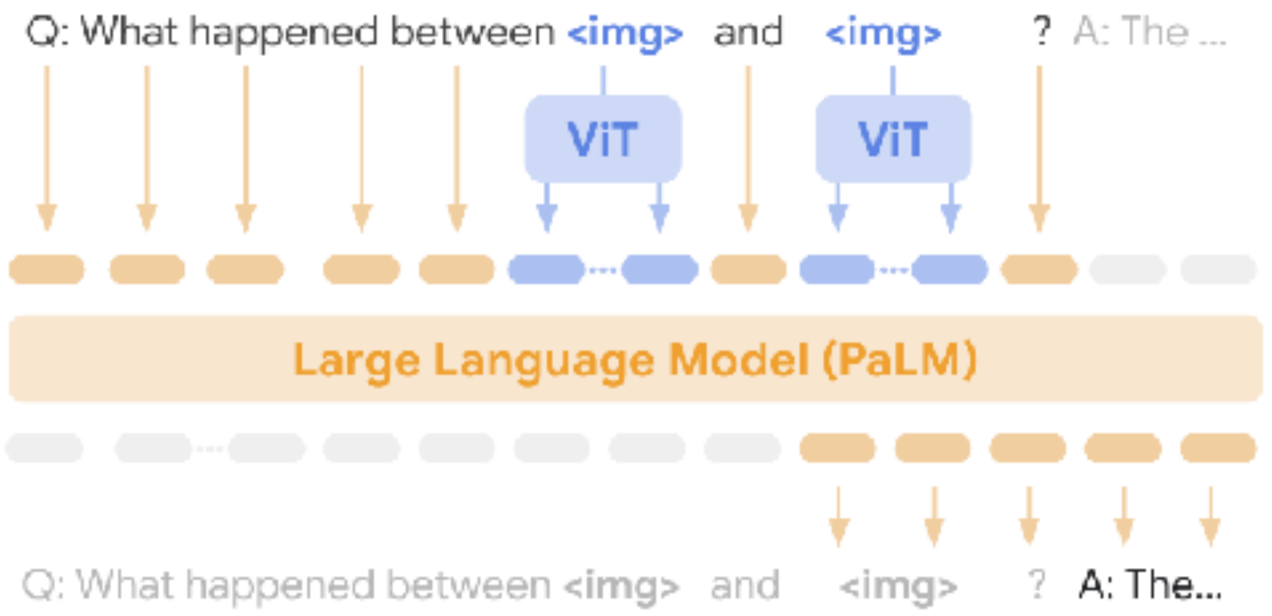
PaLI-3: Smaller, Faster, Stronger

Model	COCO	VQAv2		OKVQA	TallyQA	
	Karp.-test	test-dev	test-std	val	Simple	Complex
SimVLM	143.3	80.03	80.34	-	-	-
CoCa (2.1B)	143.6	82.3	82.3	-	-	-
GIT (0.7B)	144.8	78.56	78.81	-	-	-
GIT2 (5.1B)	145.0	81.74	81.92	-	-	-
OFA (0.9B)	145.3	82.0	82.0	-	-	-
Flamingo (80B)	138.1	82.0	82.1	57.8*	-	-
BEiT-3 (1.9B)	147.6	84.2	84.0	-	-	-
PaLM-E (562B)	138.7	80.0	-	**66.1** -	-	
MoVie	-	69.26	-	-	74.9	56.8
PaLI-17B	149.1	84.3	84.3	64.5	81.7	70.9
PaLI-X (55B)	**149.2**	**86.0**	**86.1**	**66.1**	**86.0**	**75.6**
PaLI-3 (5B)	145.9	<u>85.0</u>	<u>85.2</u>	60.1	<u>83.3</u>	70.5

Contrastive or classification pretraining for ViT?

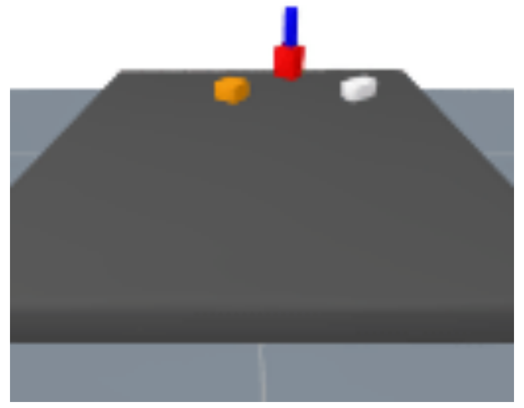
		Probe	**Captioning**			**VQA**			**RefCOCO**		
		8 tasks	COCO	XM3600		v2	OK	Text	val	+	g
G/14	Classif	88.1	139.9	94.5	44.7	76.7	57.2	31.9	51.6	43.5	43.4
	SigLIP	**-2.5**	**+0.4**	**+1.6**	**+0.7**	**+0.8**	**+1.4**	**+18.7**	**+15.1**	**+19.1**	**+17.7**
L/16	Classif	86.2	132.6	93.0	42.3	73.7	55.6	24.9	46.9	38.8	38.8
	SigLIP	**-2.8**	**+3.2**	**+1.4**	**+1.4**	**+1.9**	**+1.9**	**+16.2**	**+17.4**	**+20.9**	**+20.1**
B/16	Classif	83.7	127.7	91.7	40.7	72.3	54.7	22.5	46.3	38.1	38.4
	SigLIP	**-2.6**	**+3.6**	**-2.0**	**-0.2**	**+1.4**	**+0.9**	**+13.3**	**+16.8**	**+19.6**	**+19.3**

Method & detailed experiments



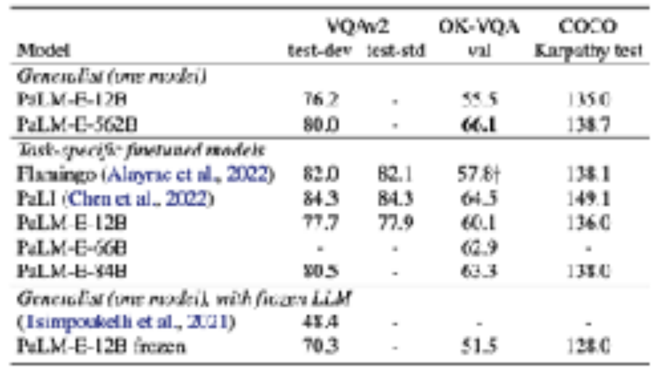
"Main" model: **PaLM-E-562B**

- Generalist visual-language model
- PaLM-540B and ViT-22B !
- Trained on: robot data, Internet-scale VQA, captioning

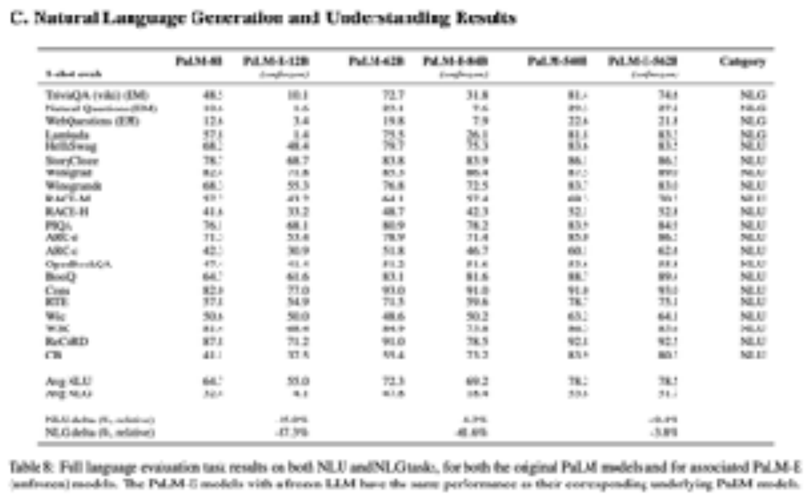


Also explored with PaLM-E:

- Neural 3D scene, and robot state encoders into the LLM
- Object-centric reasoning
- Arbitrary interleaving of text + multimodal modalities



Model	VQAv2 test-dev	VQAv2 test-std	OK-VQA val	COCO Karpathy test
Generalist (one model)				
PaLM-E-12B	76.2	-	55.5	135.0
PaLM-E-562B	80.0	-	**66.1**	138.7
Task-specific finetuned models				
Flamingo (Alayrac et al., 2022)	82.0	82.1	57.8†	138.1
PaLI (Chen et al., 2022)	84.3	84.3	64.5	149.1
PaLM-E-12B	77.7	77.9	60.1	136.0
PaLM-E-66B	-	-	62.9	-
PaLM-E-84B	80.5	-	63.3	138.0
Generalist (one model), with frozen LLM				
(Tsimpoukelli et al., 2021)	48.4	-	-	-
PaLM-E-12B frozen	70.3	-	51.5	128.0

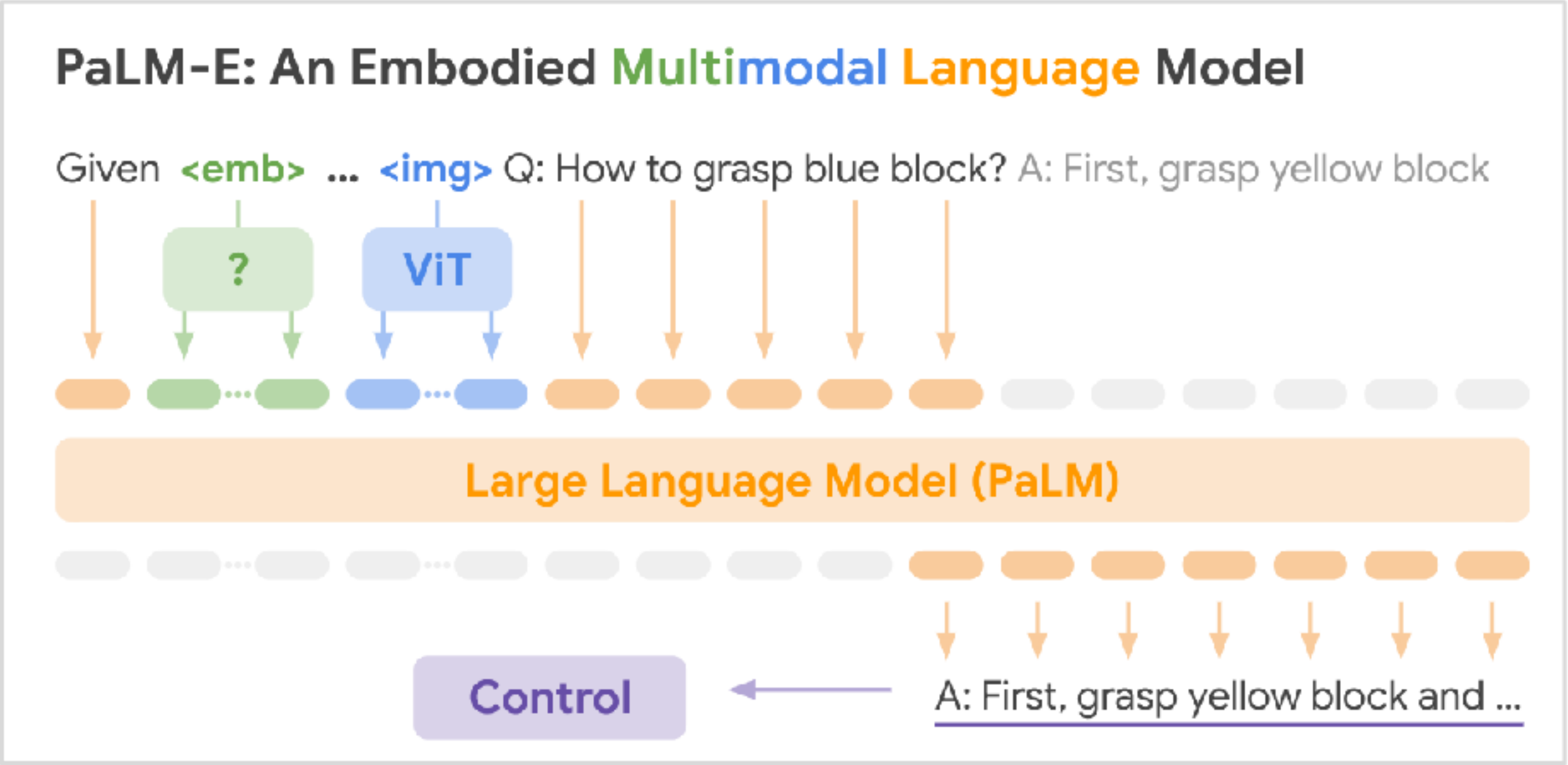


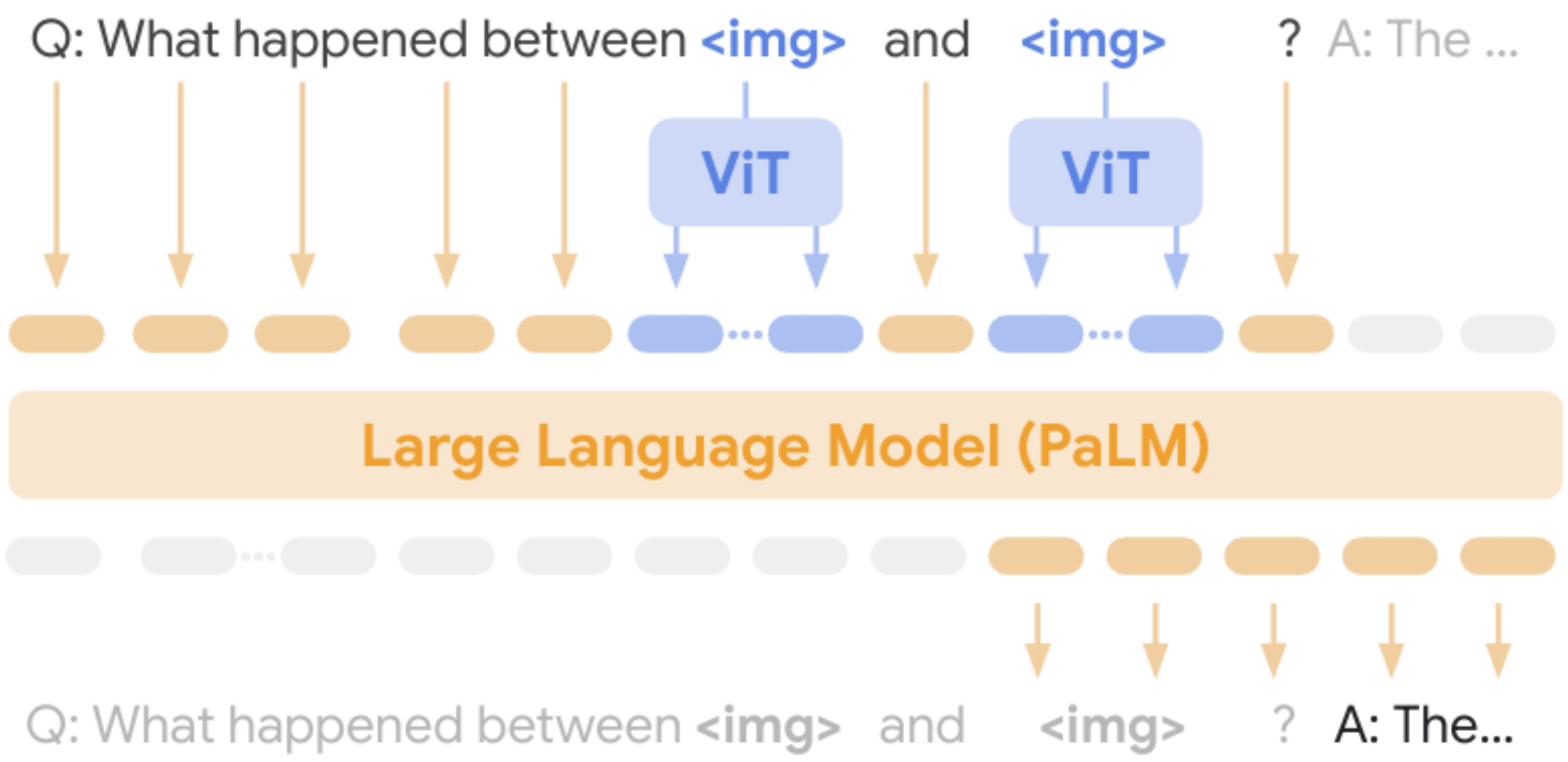
C. Natural Language Generation and Understanding Results

Experimentation

- Several different domains/categories of robot tasks
- Standard vision-language tasks
- Standard language-only tasks

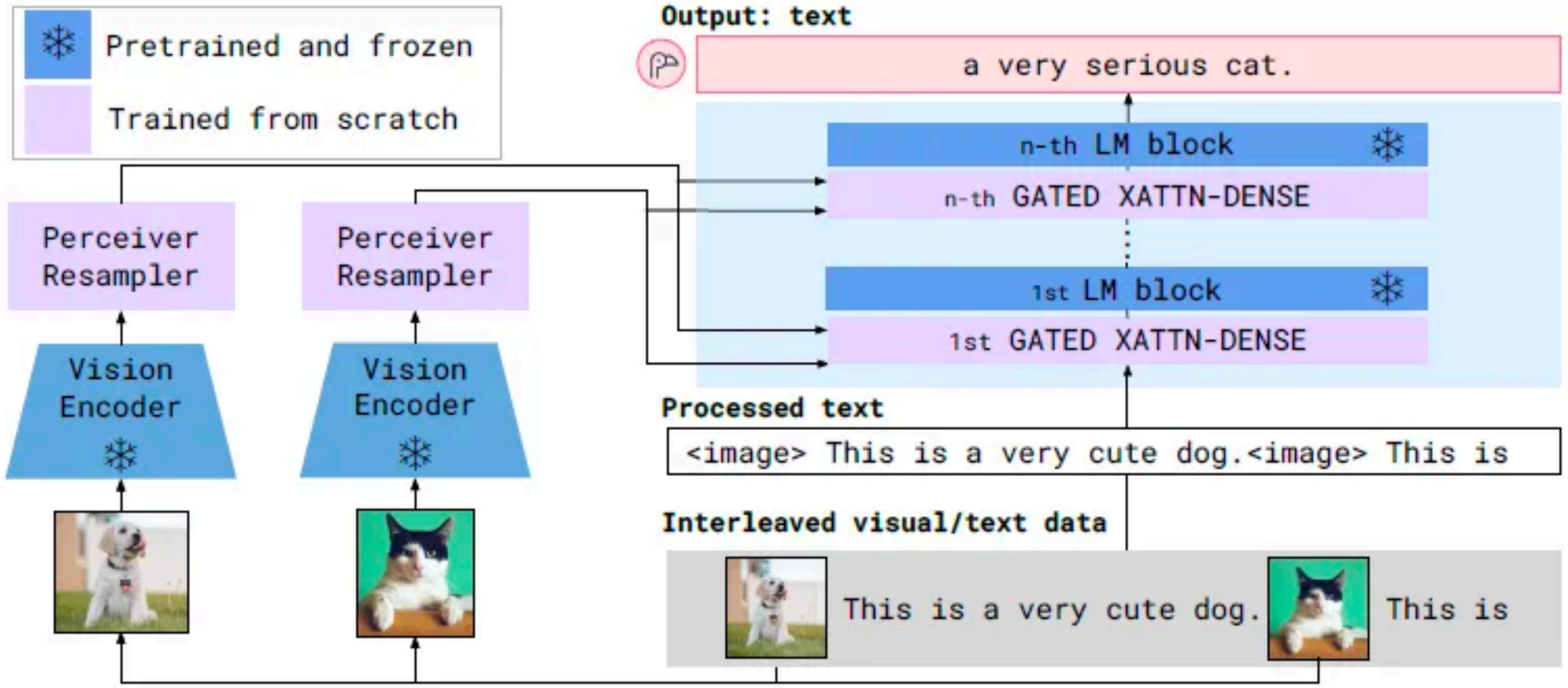
Simple Architecture of PaLM-E

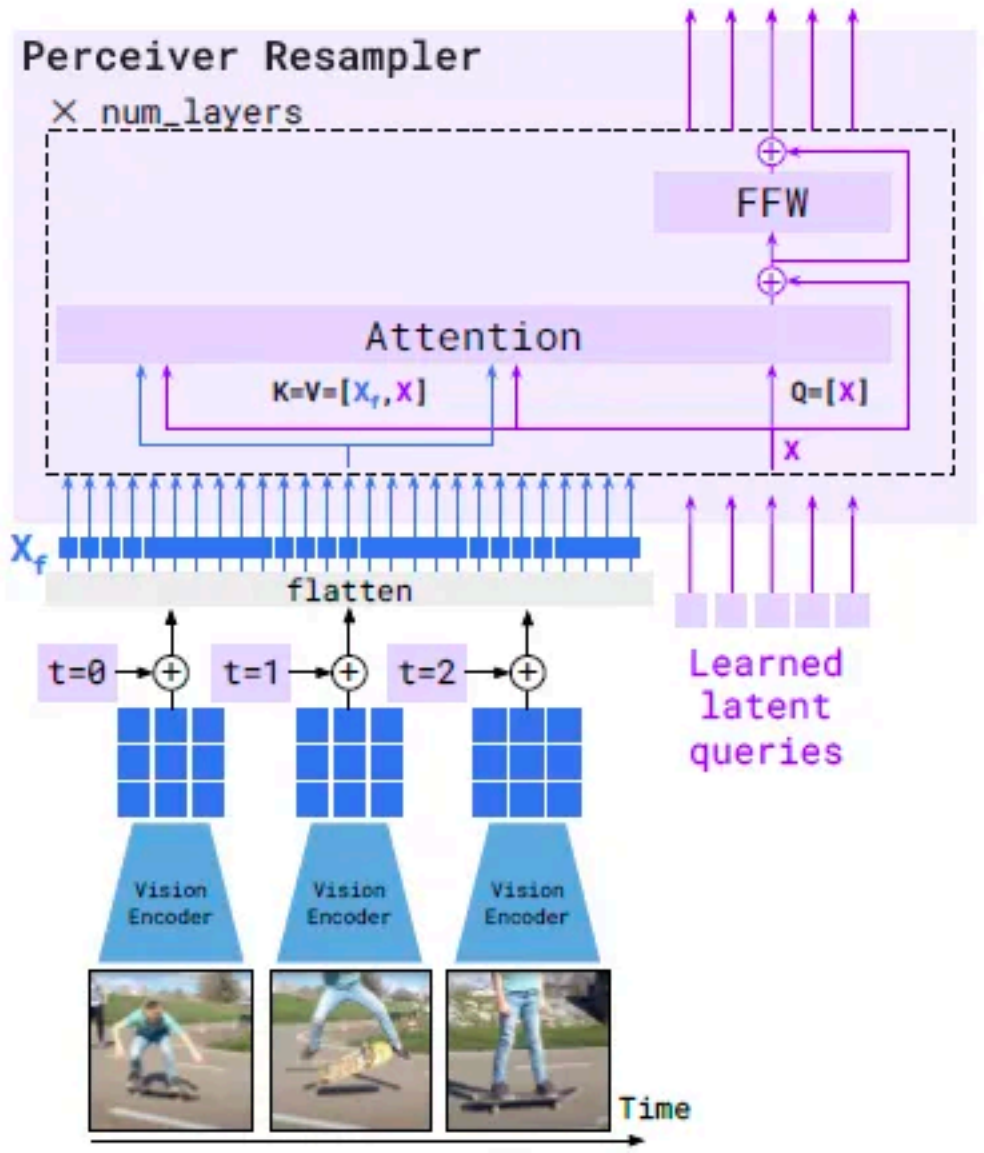




Arbitrary interleaving

Comparison to Flamingo





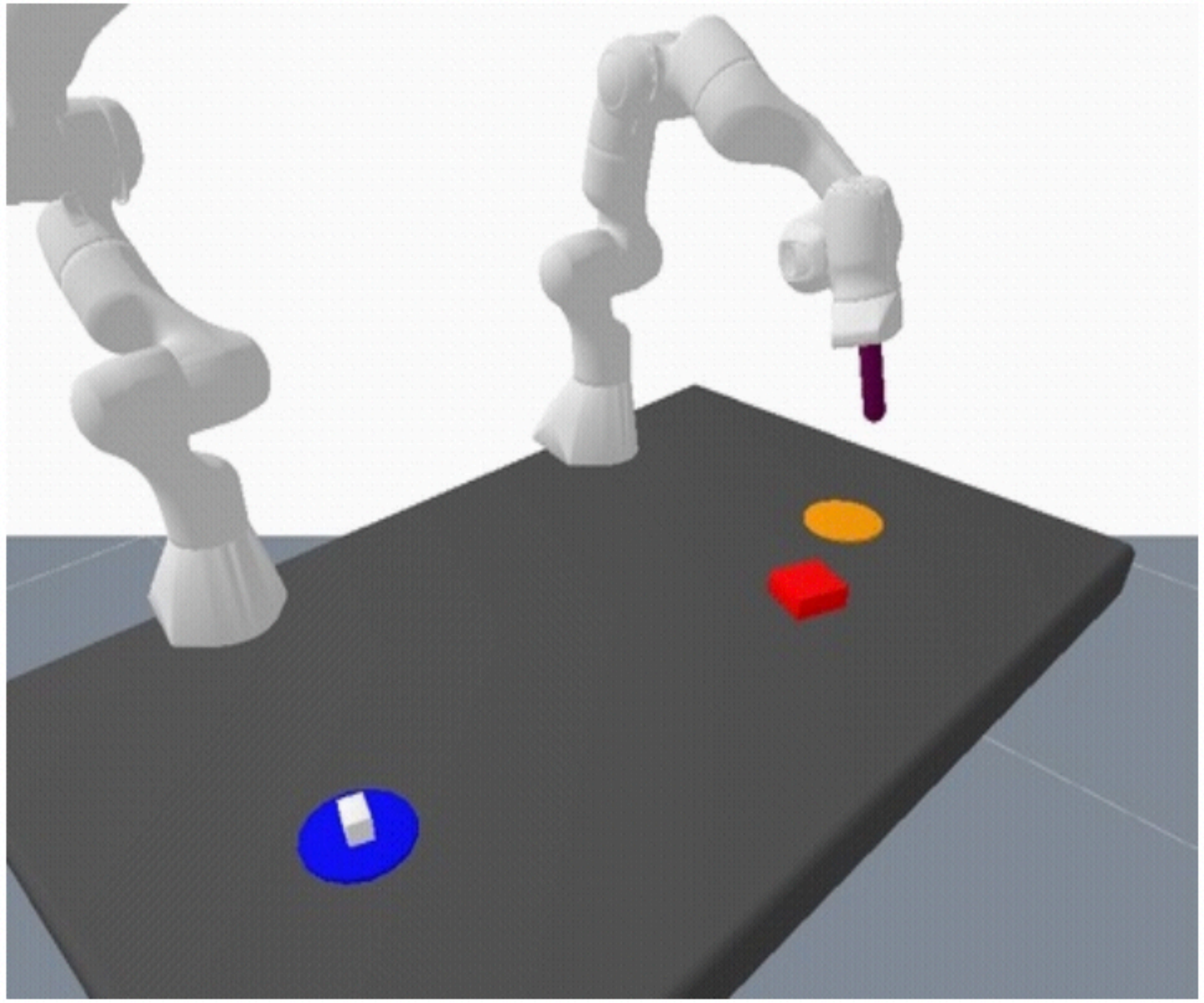

```python
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

Scene Representation: Object Scene Representation Transformer

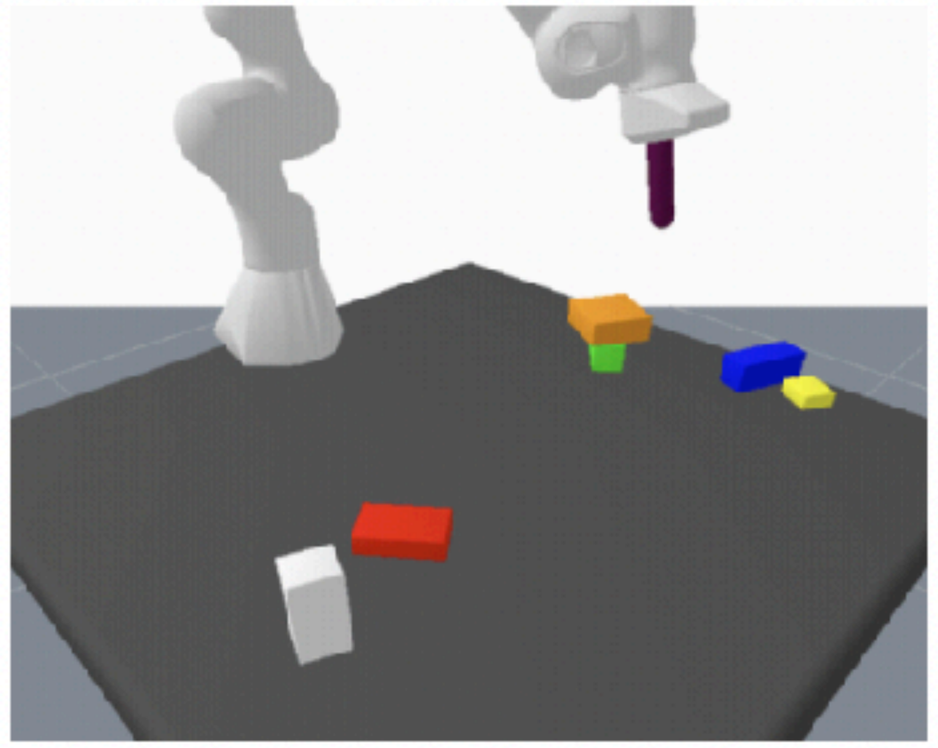
	Object-centric	LLM pre-train	Embodied VQA q_1	q_2	q_3	q_4	Planning p_1	p_2
SayCan (oracle afford.) (Ahn et al., 2022)		✓	-	-	-	-	38.7	33.3
PaLI (zero-shot) (Chen et al., 2022)		✓	-	0.0	0.0	-	-	-
PaLM-E (ours) w/ input enc:								
State	✓(GT)	✗	99.4	89.8	90.3	88.3	45.0	46.1
State	✓(GT)	✓	**100.0**	96.3	95.1	93.1	55.9	49.7
ViT + TL	✓(GT)	✓	34.7	54.6	74.6	91.6	24.0	14.7
ViT-4B single robot	✗	✓	-	45.9	78.4	92.2	30.6	32.9
ViT-4B full mixture	✗	✓	-	70.7	93.4	92.1	74.1	74.6
OSRT (no VQA)	✓	✓	-	-	-	-	71.9	75.1
OSRT	✓	✓	99.7	**98.2**	**100.0**	**93.7**	**82.5**	**76.2**

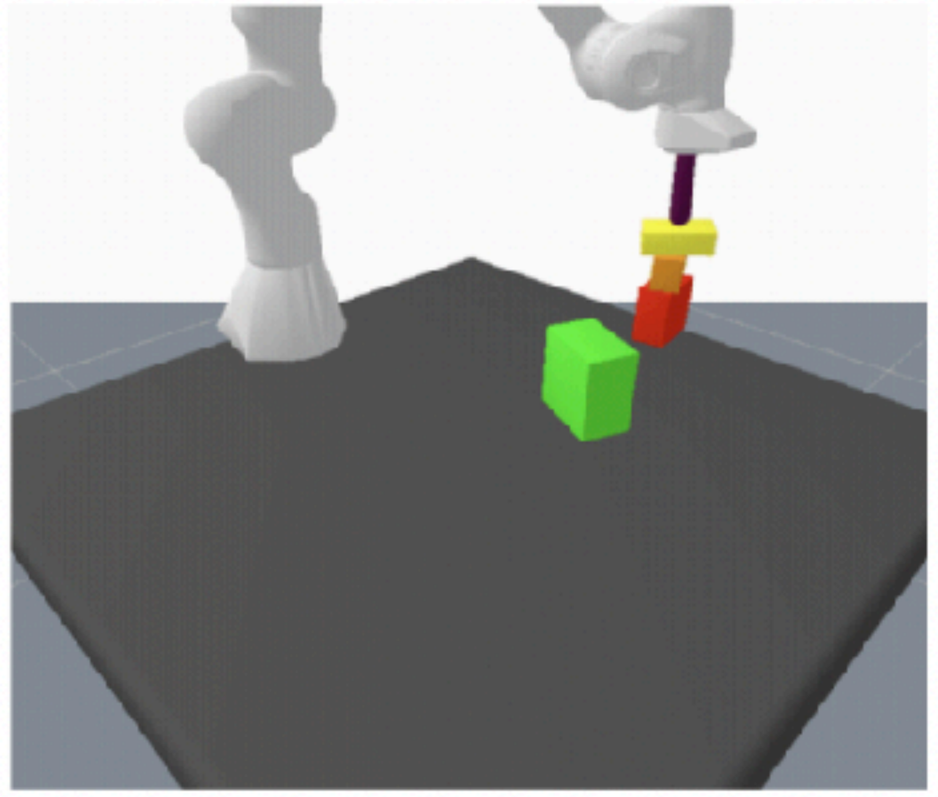
PaLM-E on TAMP-like tasks



Given **img**. Q: How to stack the white object on top of the red object?

A: First grasp the red object and place it on the table, then grasp the white object and place if on the red object.





Highlight: **Zero-shot Multimodal Chain-of-Thought**

Given <img>.
Q: In this photo, are there more cats, or more dogs?
A: Let's think step by step.
There are 2 dogs and 2 cats.
2 and 2 are equal.
Therefore there are equal number of cats and dogs.

Given <img>. Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step.
1. do not enter.
2. except bicycles.
3. do not entry except bicycles.
4. yes.

I see (image omitted). Description: A cow sitting on a rug. Joke: Why is the cow sitting on a rug? Because it wants to be a cow-ch! ... I see <img>. Description: a donkey is carrying a dog, cat, and rooster. Joke: what do you call a donkey with a rooster on his back. A rooster booster.

Given <img> .
Q: I am getting just two custom pizza for me and my friend. How much should I pay in total? A: Let's think step by step.
1st step: 2 custom pizzas.
2nd step: 2 x $9.99.
3rd step: $19.98.

Highlight: **Complex Multi-image Reasoning**

Photo 1: **<img1>**. Photo 2: **<img2>**. Q: What is in Photo 1 but NOT in Photo 2? A: Let's think step by step and explain. Photo 1 has sunglasses on top of folded clothes. Photo 2 does not have sunglasses on top of folded clothes. Therefore, sunglasses are in Photo 1 but not in Photo 2.

Photo 1: **<img1>**. Photo 2: **<img2>**. Q: Which t-shirt from Photo 2 matches the sweatshirt from Photo 1? A: Let's think step by step and explain.

1. photo 1 has sweatshirt.
2. photo 2 has t shirts.
3. photo 1 has blue sweatshirt.
4. photo 2 has blue t shirt.
5. blue t shirt matches blue sweatshirt.

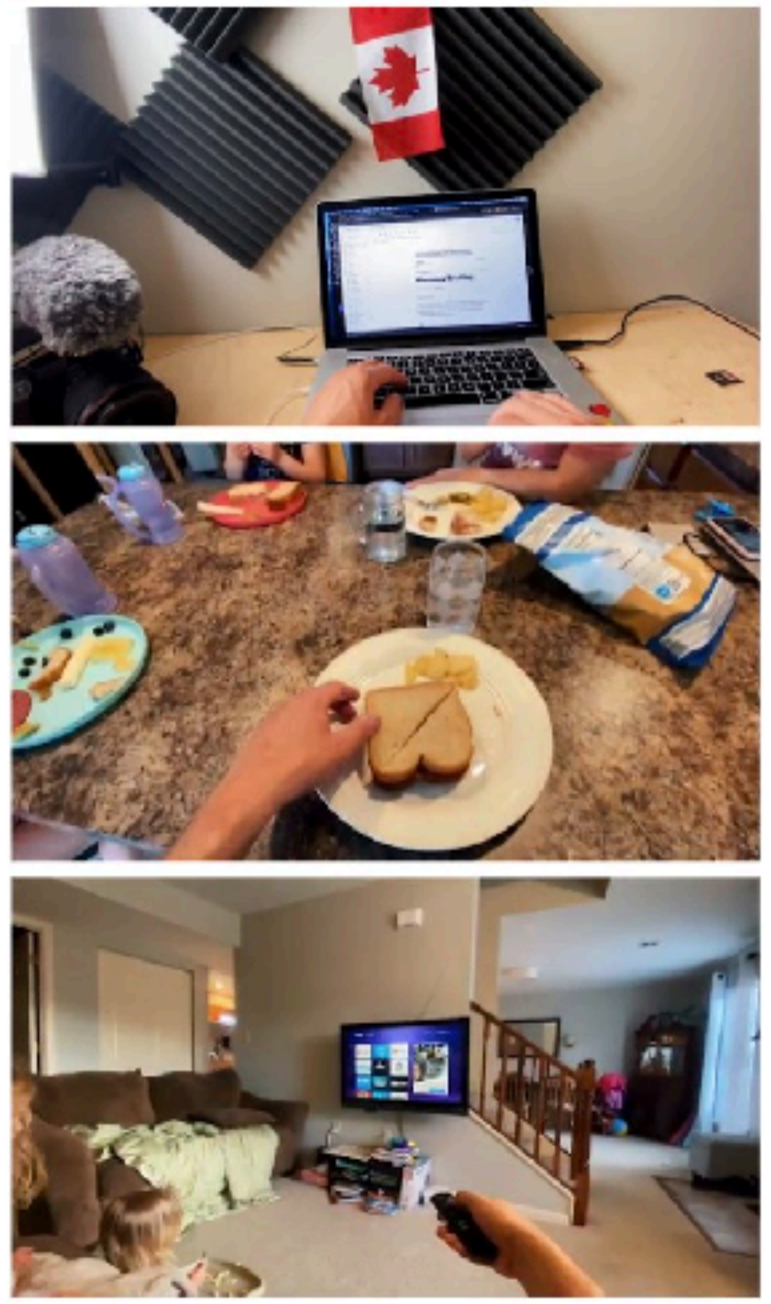
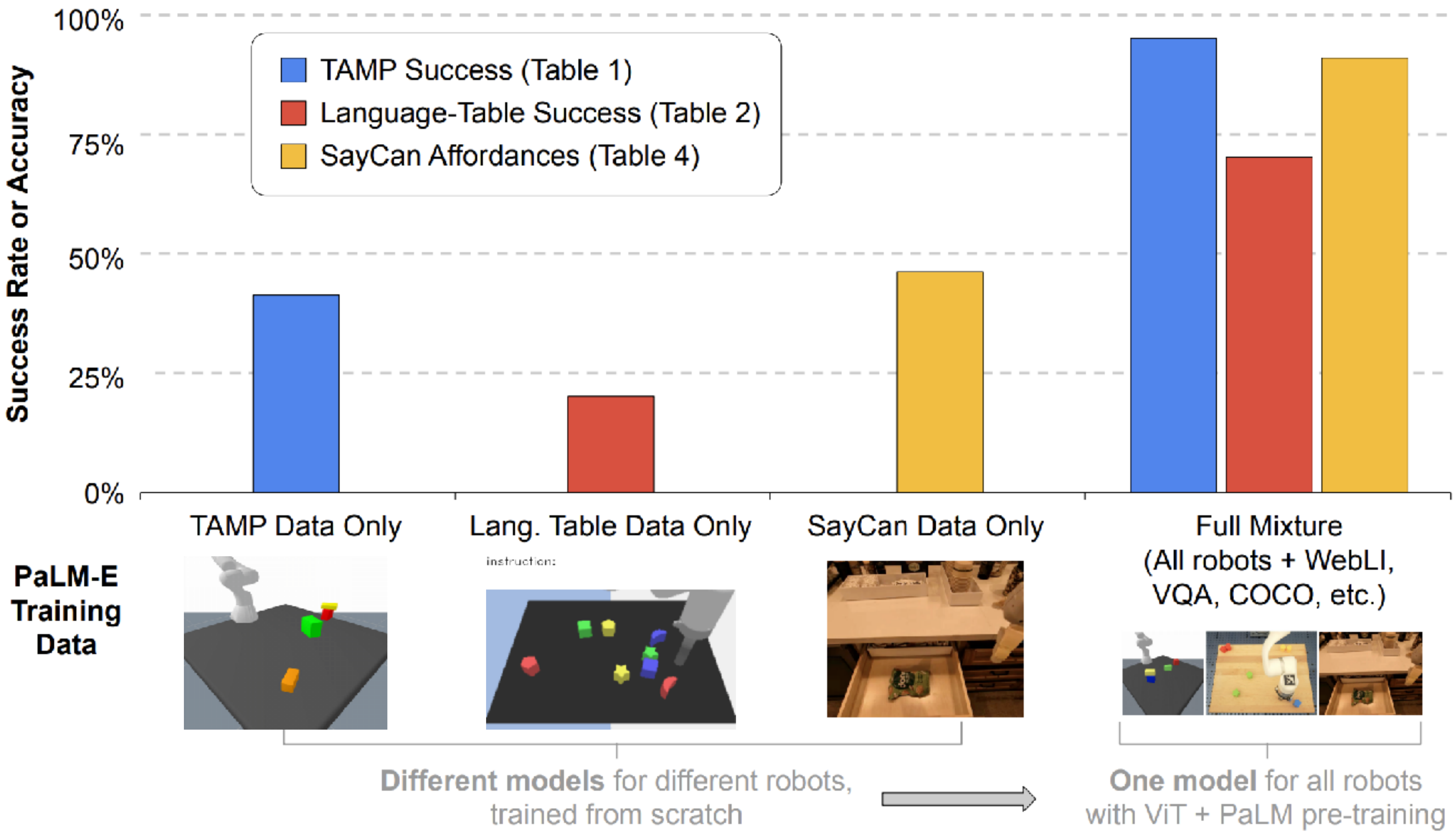


Photo 1, at 10:30 am: **<img1>**.
Photo 2, at 12:45 pm: **<img2>**.
Photo 3, at 3:45 pm: **<img3>**.
Q: I forget, what did I have for lunch, and what time was it? A: Let's think step by step.

1. you had a sandwich for lunch.
2. it was 12:45 pm.

PaLM-E: Positive Transfer

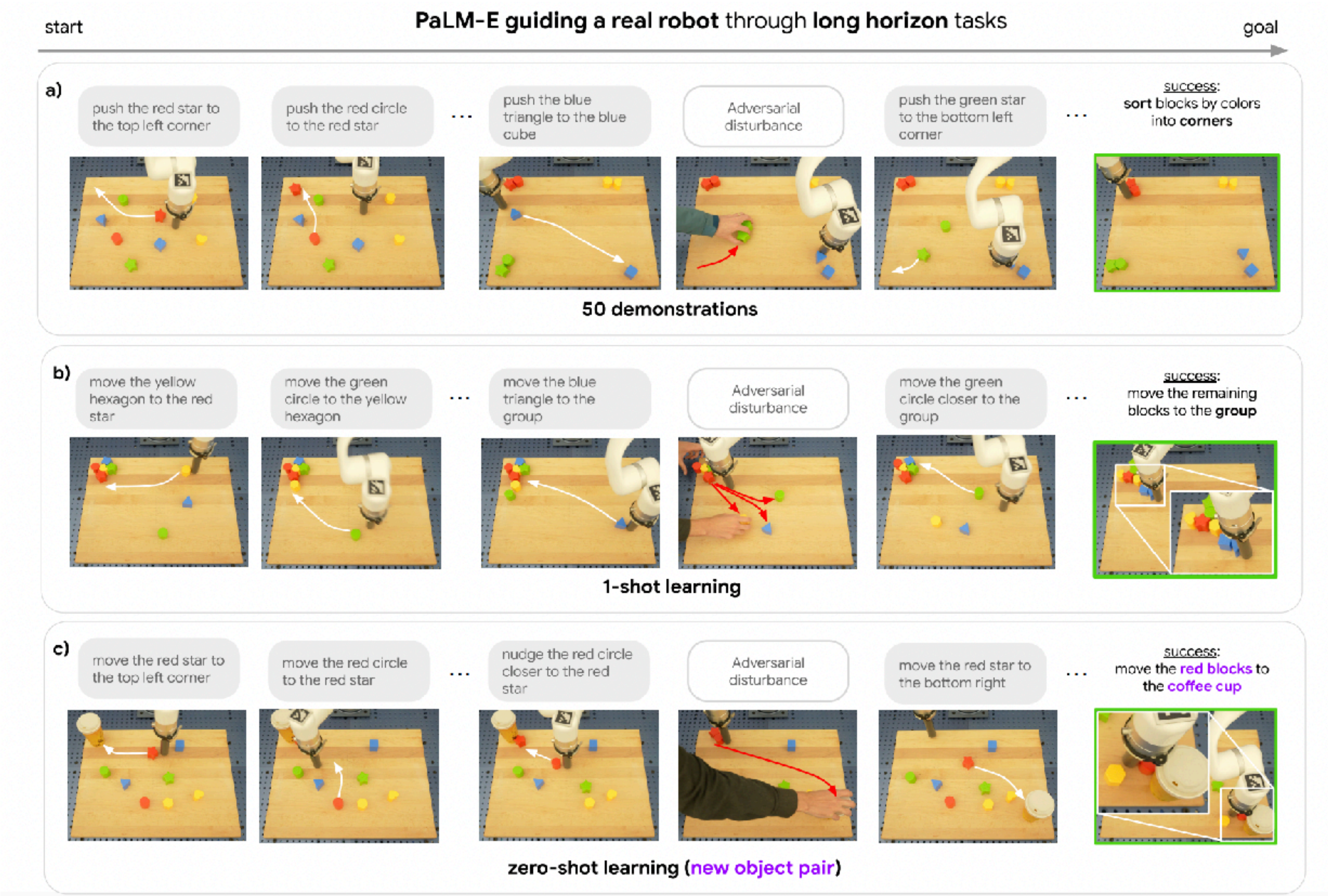


Real Robot Results

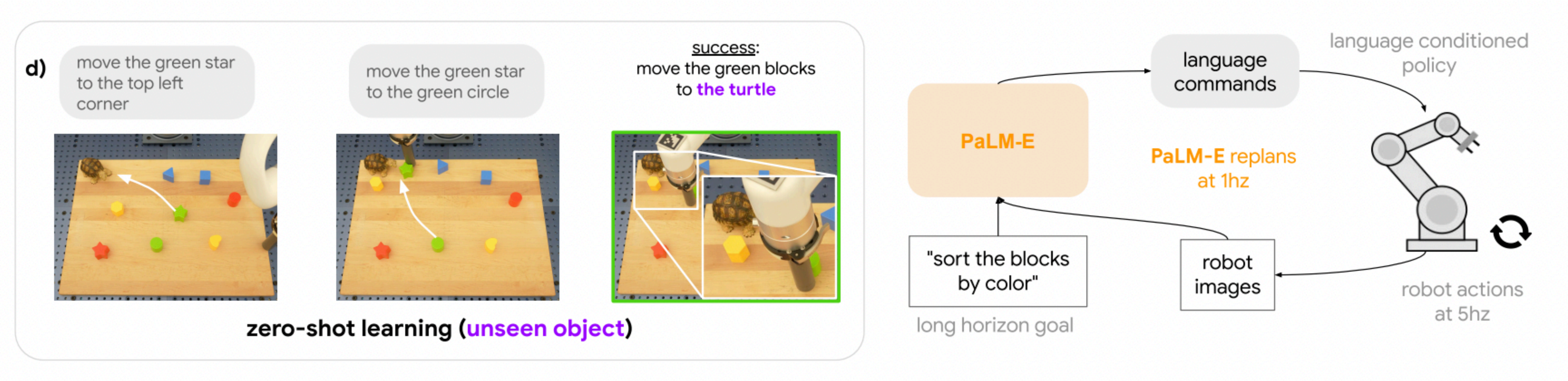
Given  . Q: How to sort the blocks by colors into the corners? A:

Same exact model checkpoint!
(PaLM-E can be a multi-embodiment robot brain)

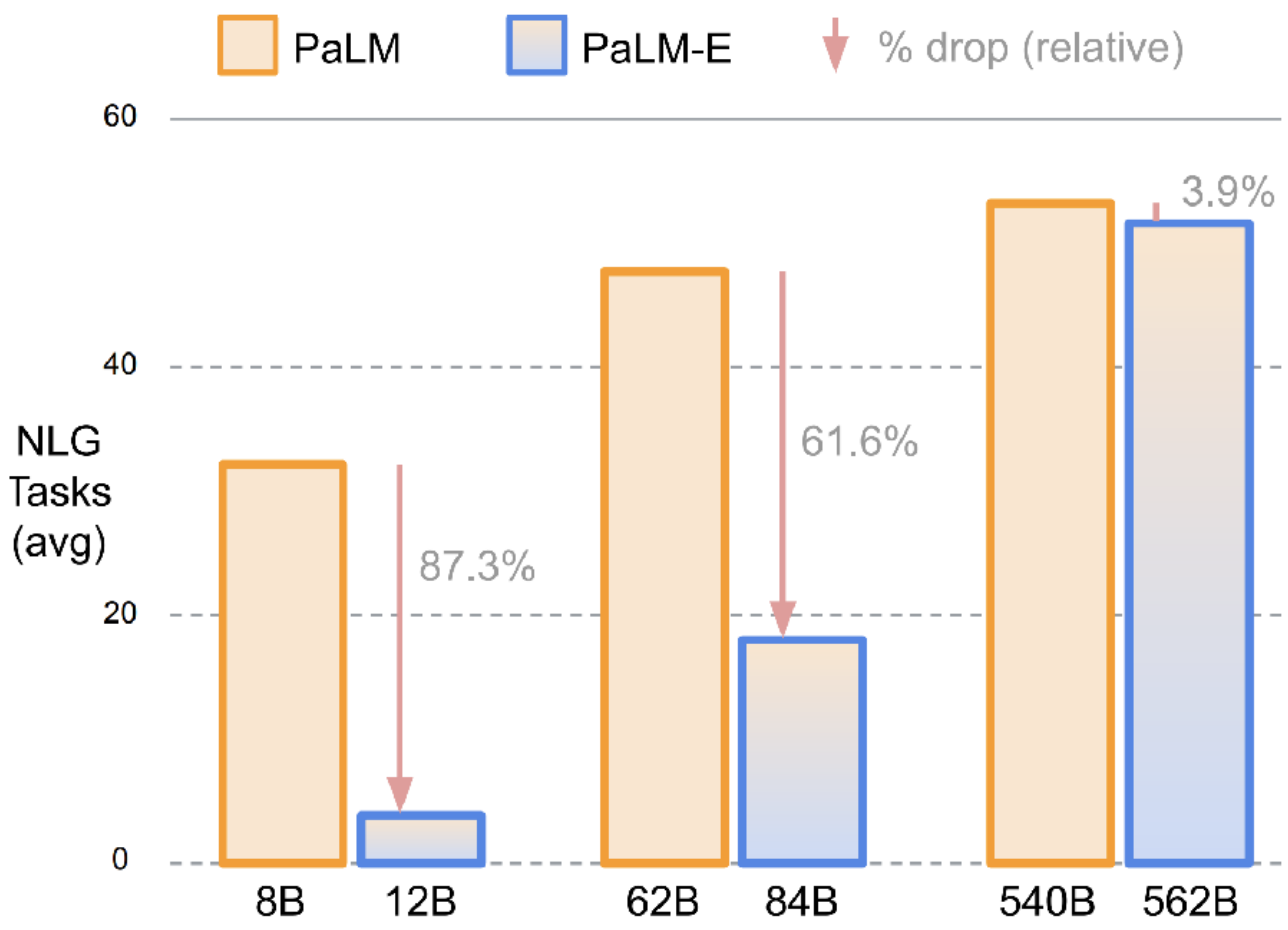
Sample-efficient learning



Sample-efficient learning



Language catastrophic forgetting reduced with scale

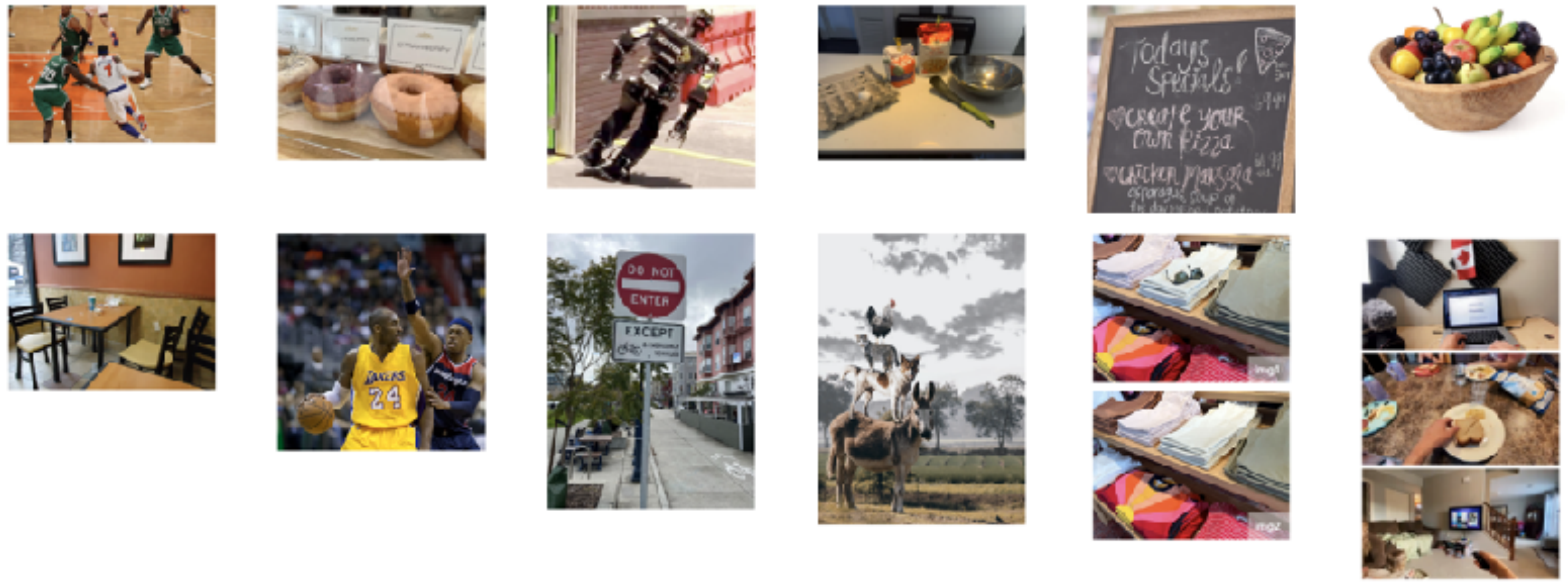


For more on PaLM-E

- See paper + videos + inference examples at palm-e.github.io

Demo

The examples below are all example completions (in orange) from PaLM-E. The prompt is the one or more images and the text in gray.



Given **img**. Q: I am getting just two custom pizza for me and my friend. How much should I pay in total? A: Let's think step by step.

1st step: 2 custom pizzas. 2nd step: 2 x $9.99. 3rd step: $19.98.

Hands-on VLMs

LLaVA: Large Language and Vision Assistant

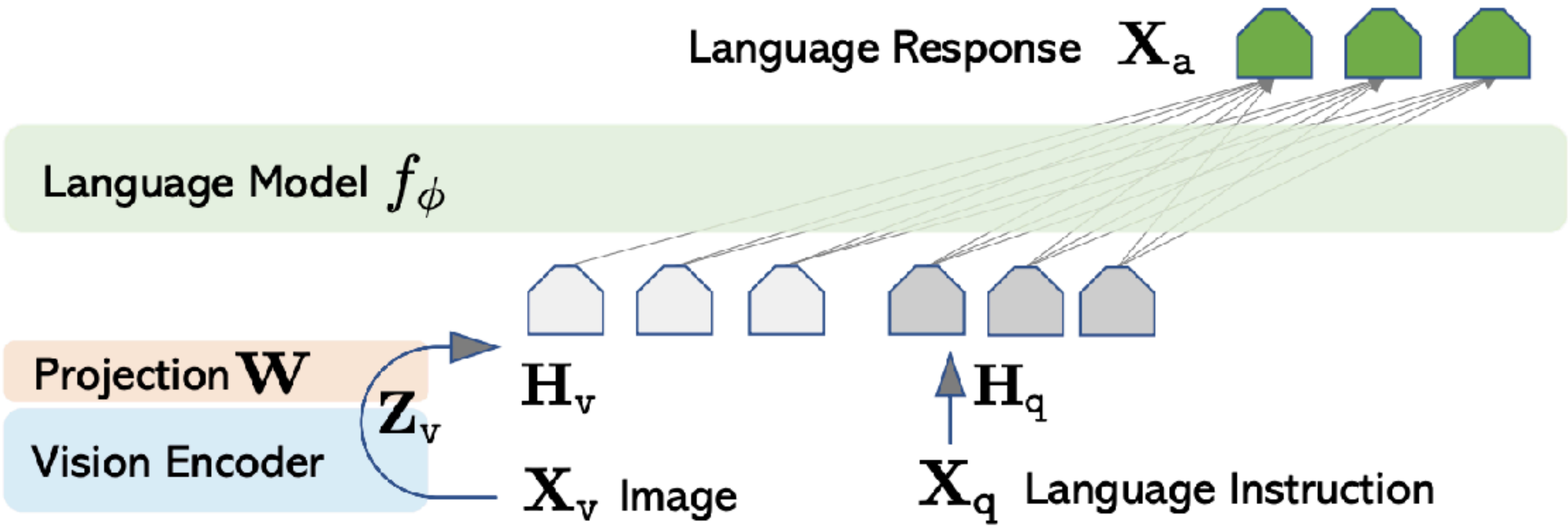
Visual Instruction Tuning

NeurIPS 2023 (Oral)

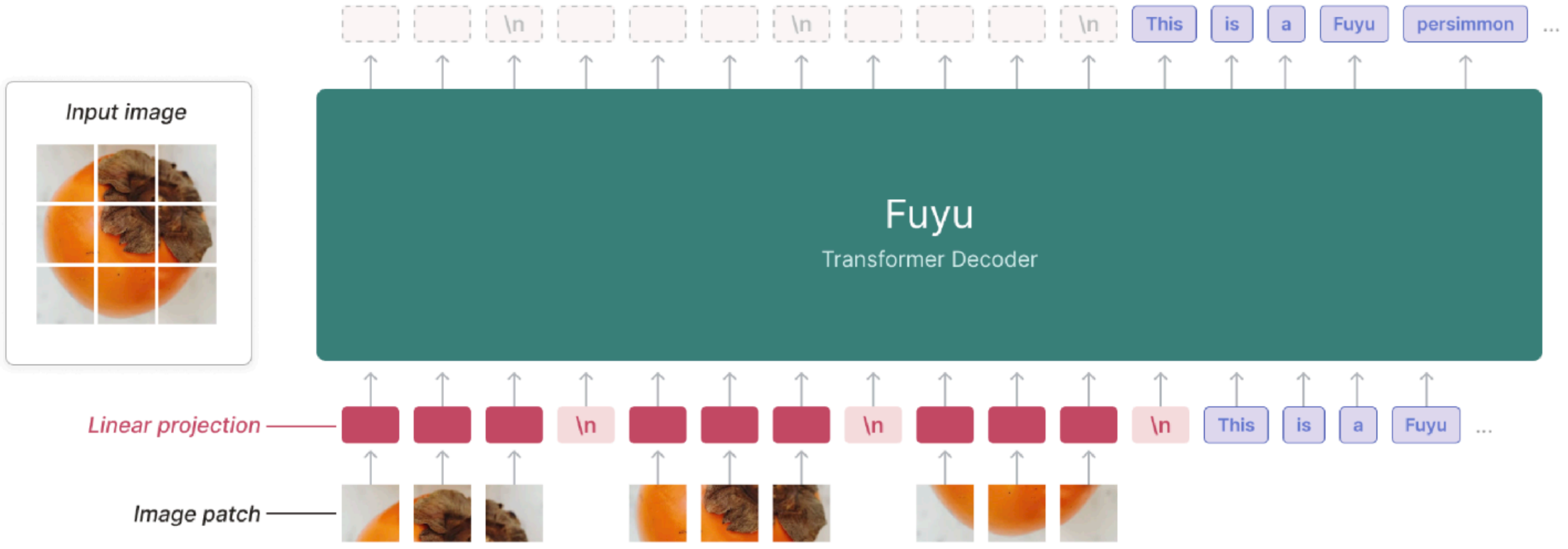
Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee

University of Wisconsin-Madison ▶ Microsoft Research ▶ Columbia University

*Equal Contribution



Hands-on VLMs, Fuyu-8b and open source PaLM-E



- A good programming exercise:
- Fix the bug in
- https://github.com/kyegomez/PALM-E/blob/main/palme/model.py

Discussions

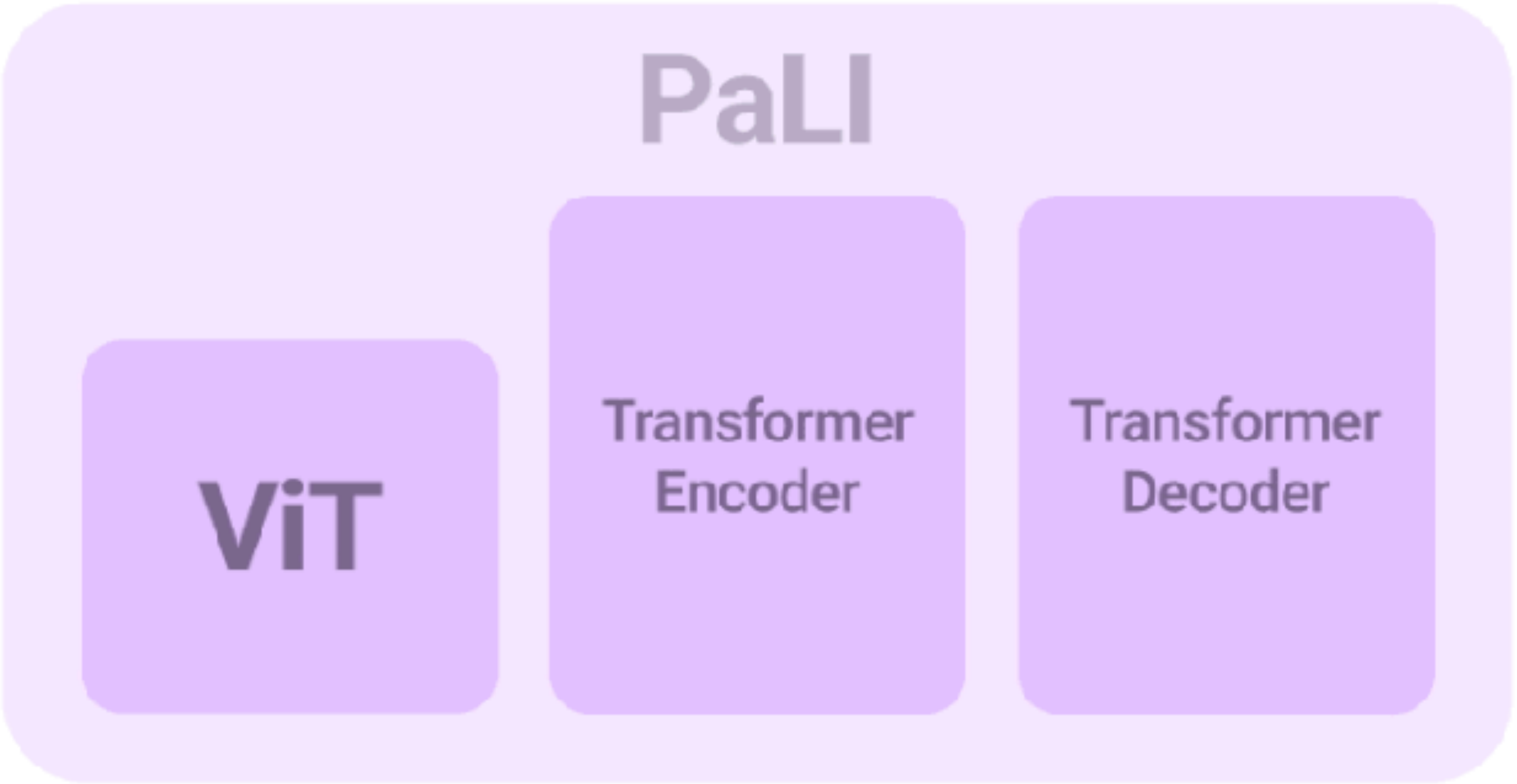
RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, Brianna Zitkovich


Robotics @ Google

Let’s dive into RT-2!

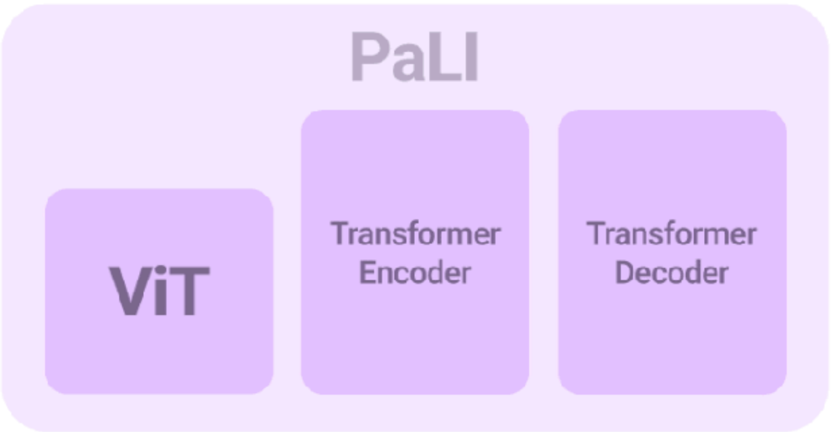
Vision-Language Models



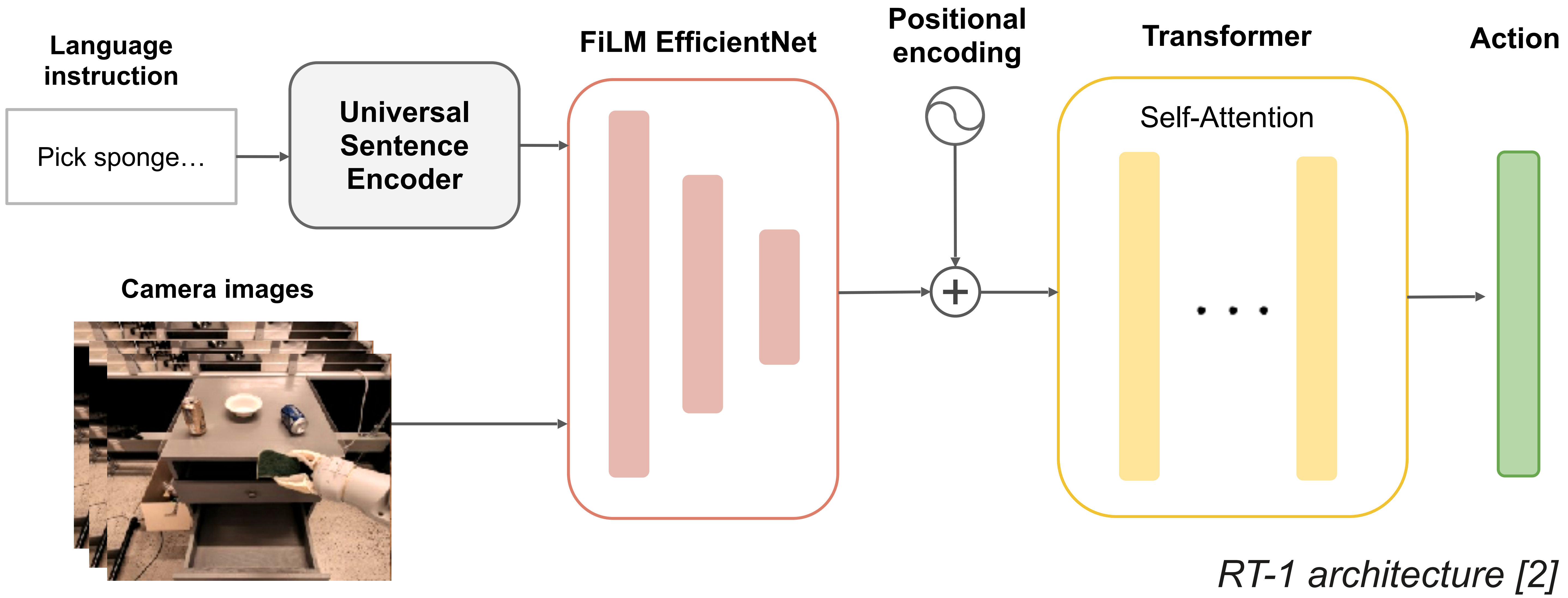
- VLMs encompass both **visual** and **semantic** understanding of the world
- In Robotics we have to deal a lot with **both** of these
- How do we leverage all of this knowledge?

[1] PaLI: A Jointly-Scaled Multilingual Language-Image Model. Chen et al. 2022.

VLMs as Robot Policies



PaLI architecture [1]



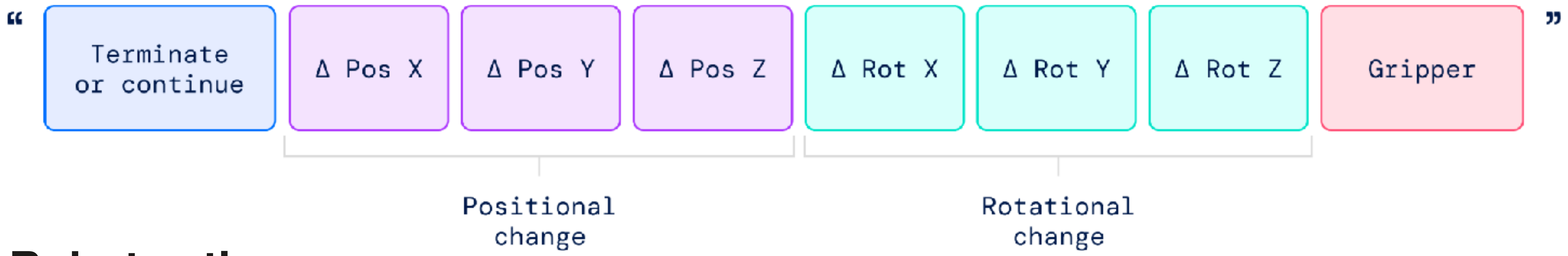
RT-1 architecture [2]

- **RT-1:** image + text → **discretized actions**
- Similar to a Visual-Language Model (VLM) with different **output tokens**
- Use large pre-trained VLMs directly as the **policy**!
- How do we **deal with actions** when using pre-trained VLMs?

[1] PaLI: A Jointly-Scaled Multilingual Language-Image Model. Chen et al. 2022.

[2] RT-1: Robotics Transformer for Real-World Control at Scale, Robotics at Google and Everyday Robots, 2022.

Representing Actions in VLMs



- **Robot actions:**
  - Moving the robot arm and gripper
  - Discretized into 256 bins
- **Actions in VLMs**
  - Convert to a string of numbers
  - Example: "1 127 115 218 101 56 90 255"
  - Alternatives:
    - *Float numbers* - more tokens needed
    - *Human language (left, right etc.)* - can't be directly executed on a robot

→ **Vision-Language-Action (VLA) model!**

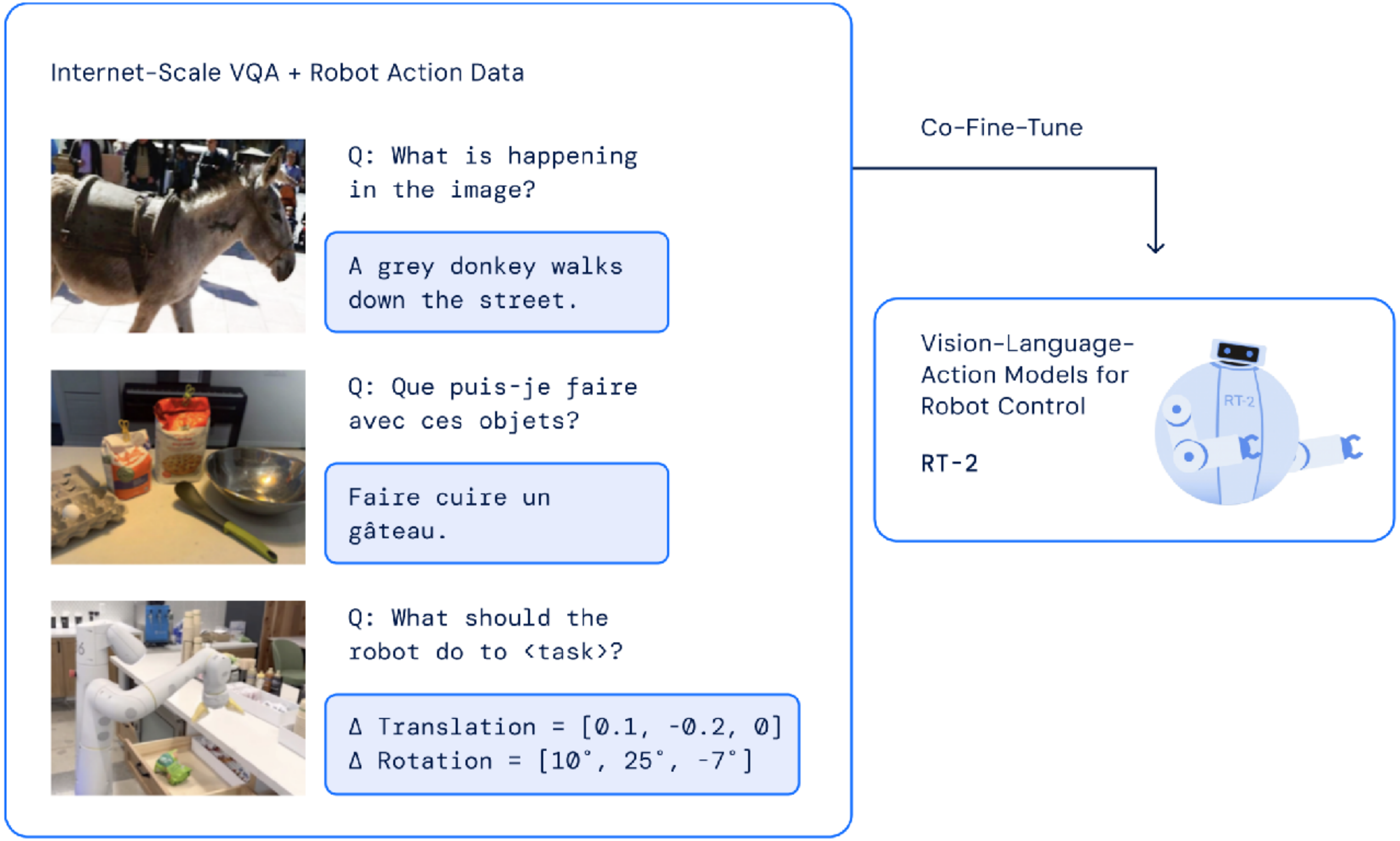
Training data and underlying models

Models

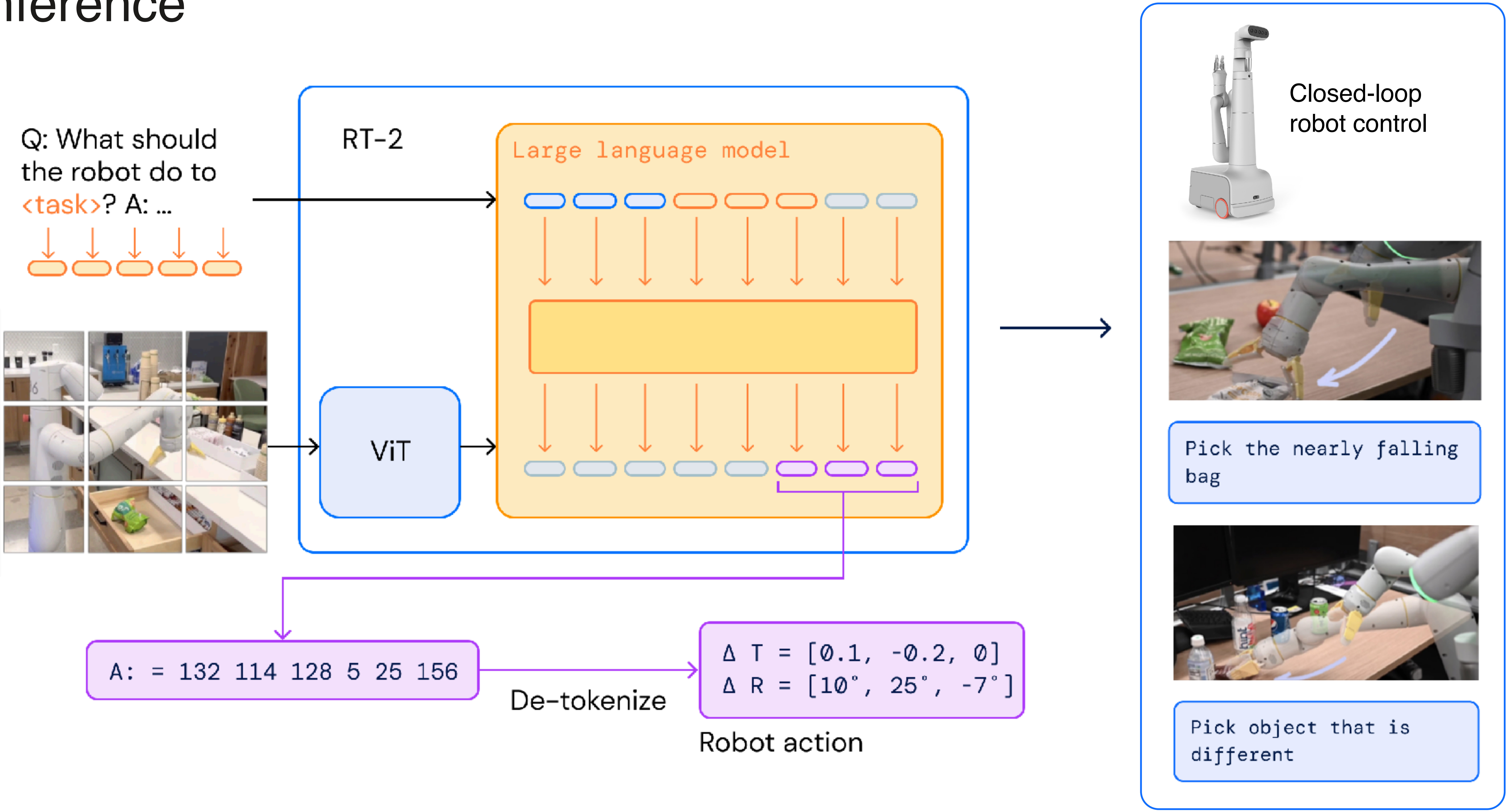
- PaLI-X (5B, 55B)
- PaLM-E (12B)

Data

- Pretraining: Web-data
- Robot data
  - RT-1 data
  - Human demos
  - 13 robots
  - 17 months



Inference



Q: What should the robot do to <task>? A: ...
RT-2
Large language model
ViT
A: = 132 114 128 5 25 156
De-tokenize
Δ T = [0.1, -0.2, 0]
Δ R = [10°, 25°, -7°]
Robot action
Closed-loop robot control
Pick the nearly falling bag
Pick object that is different

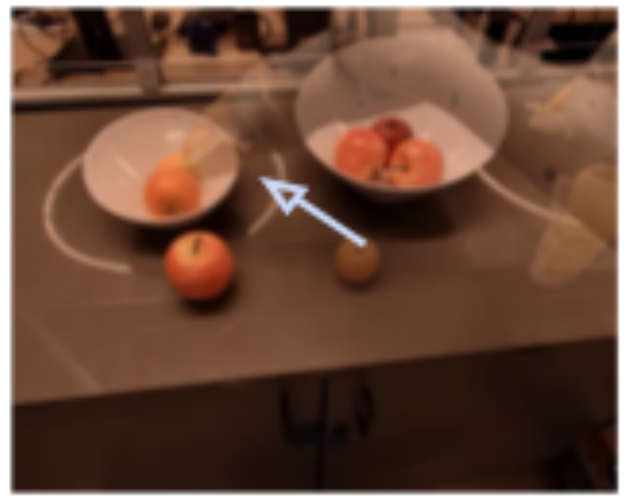
Results: Emergent skills

put strawberry into the correct bowl

pick up the bag about to fall off the table

move apple to Denver Nuggets

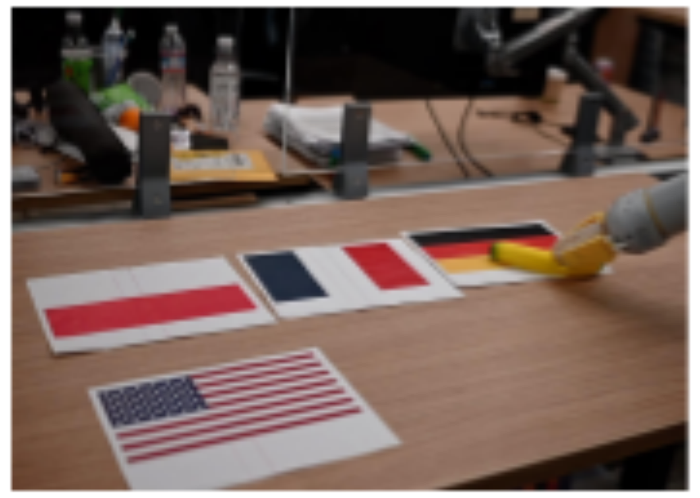
pick robot



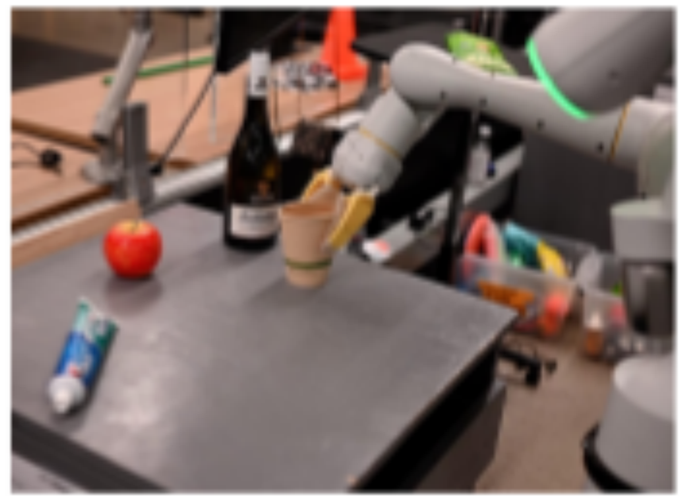
place orange in the matching bowl

move redbull can to H

move soccer ball to basketball



move banana to Germany



move cup to the wine bottle

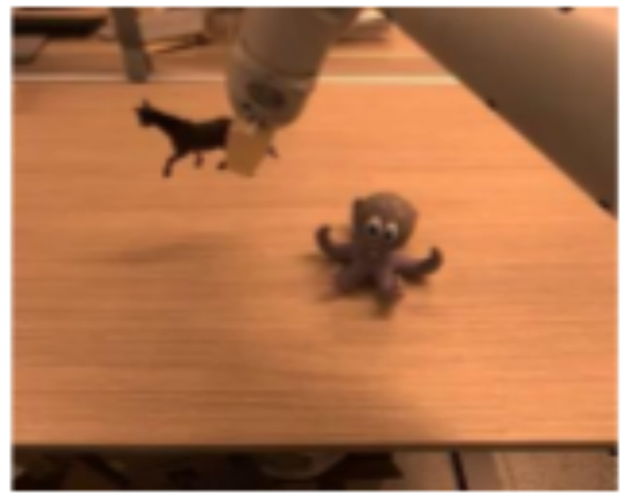
pick animal with different color

move coke can to Taylor Swift

move coke can to X

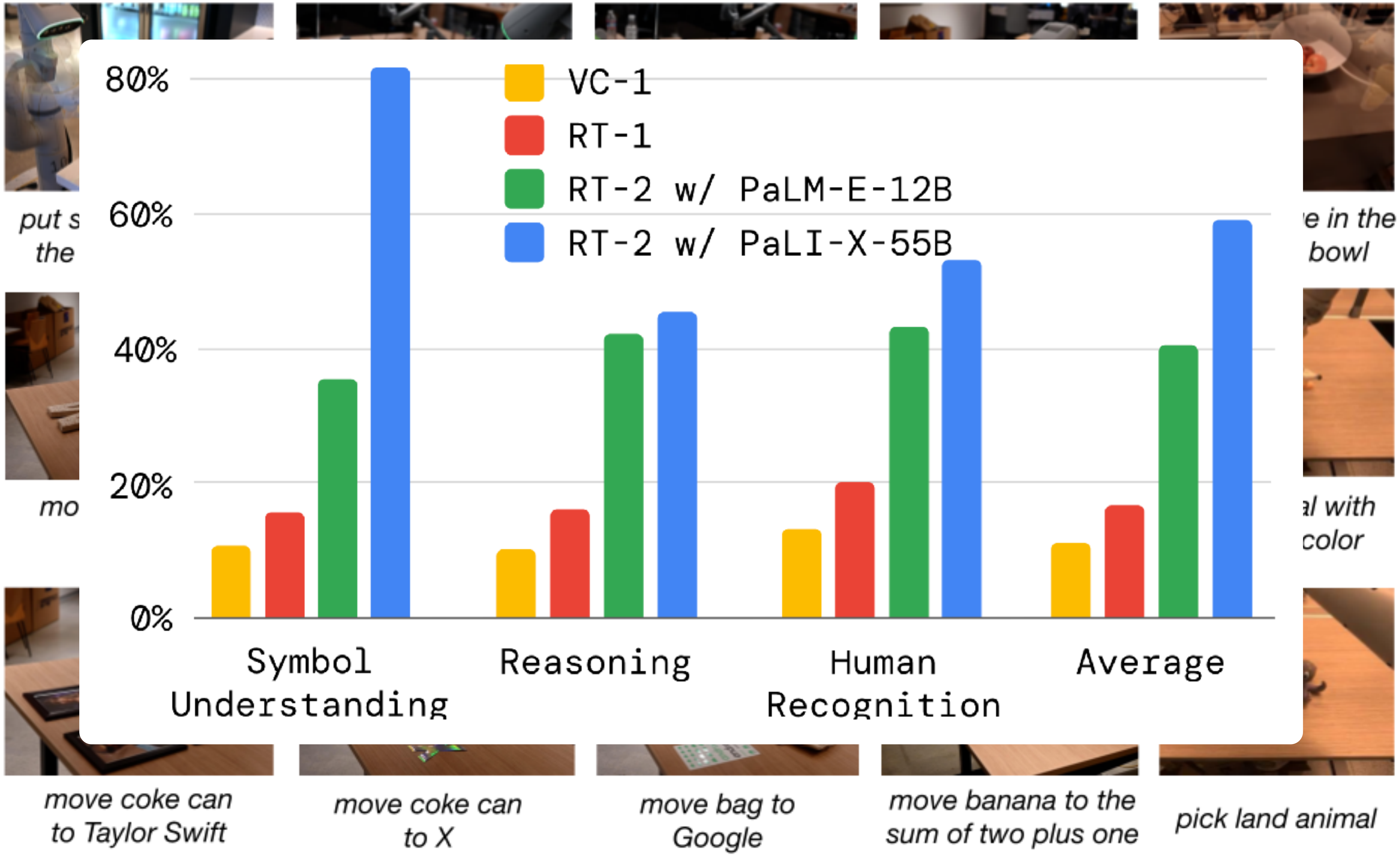
move bag to Google

move banana to the sum of two plus one



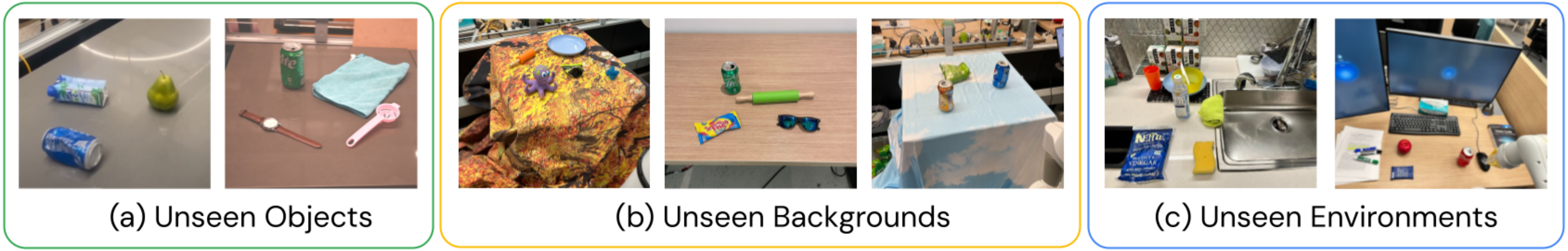
pick land animal

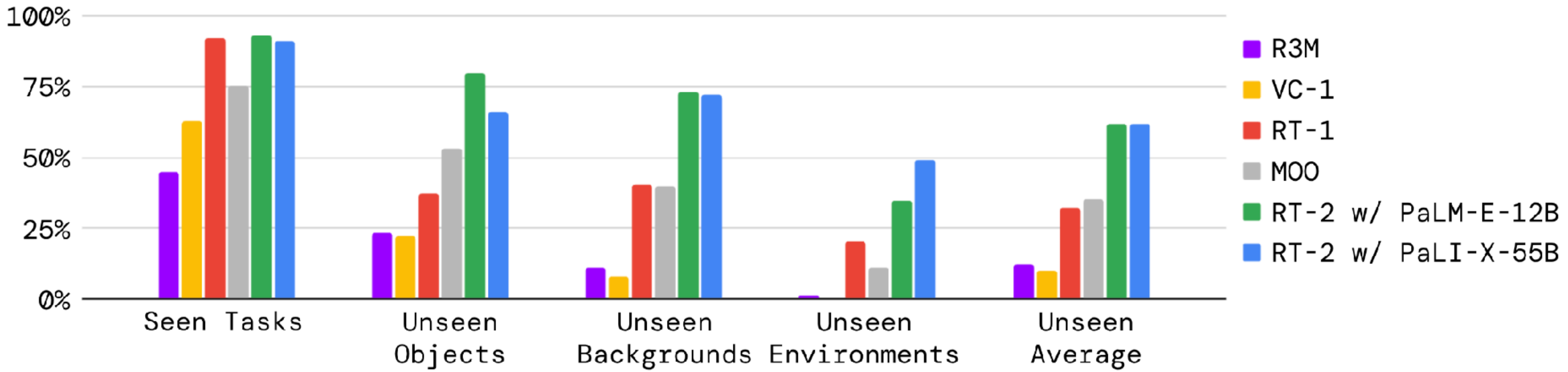
Results: Emergent skills



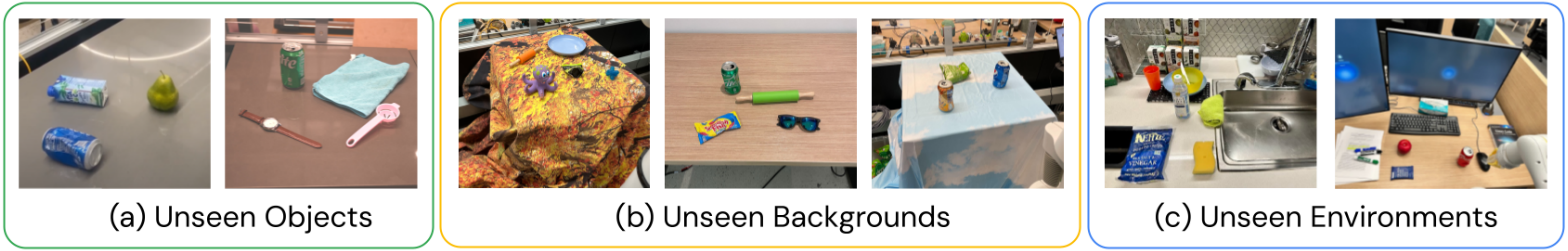
Results: Emergent skills

Results: Quantitative evals



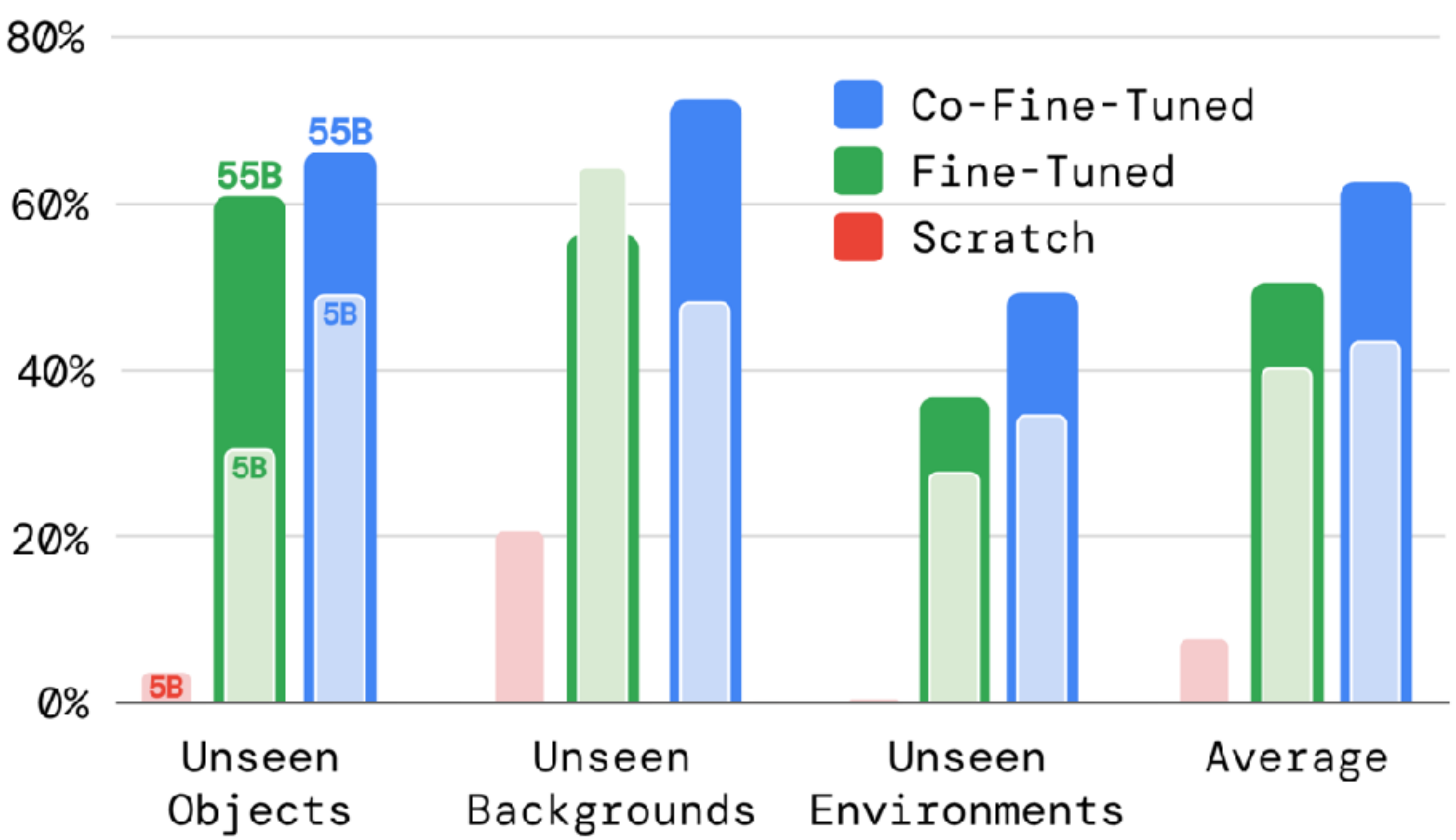


Results: Quantitative evals

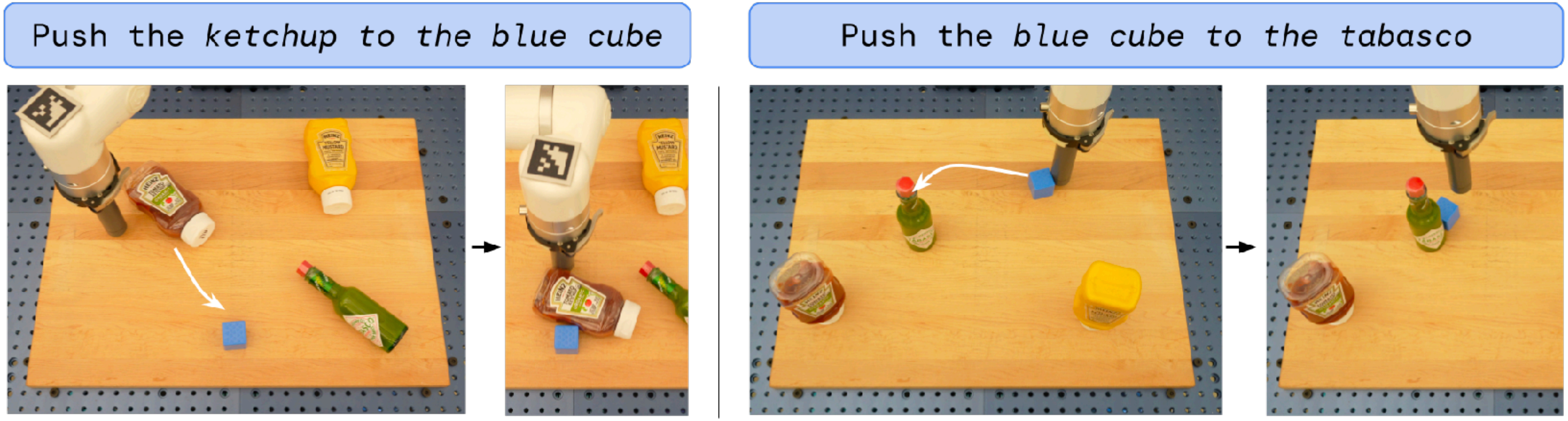


RT2 w/ PaLI-X-55B ablations

- Co-Fine-Tuning with VQA data
- Fine-Tuning on robot data only
- Training on robot data from scratch



Results: Language Table



Language Table Benchmark

- Trained on pushing cubes only
- Generalizing to new objects

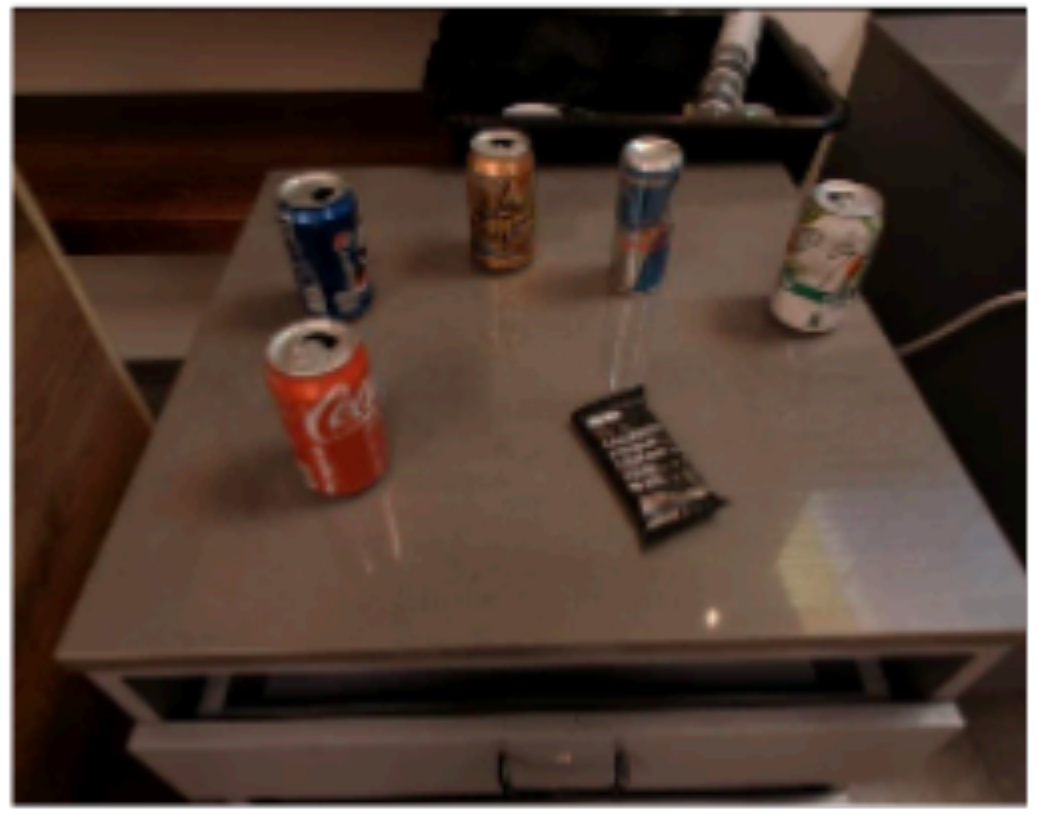
Model	Language-Table
BC-Zero (Jang et al., 2021)	72 ± 3
RT-1 (Brohan et al., 2022)	74 ± 13
LAVA (Lynch et al., 2022)	77 ± 4
RT-2-PaLI-3B (ours)	**90 ± 10**

Results: Chain-of-Thought with RT-2-PaLM-E

Prompt:
Given <img> Instruction:
Bring me a drink.
Prediction:
Plan: pick 7up can.
Action: 1 143 129 123 145 114 115 127

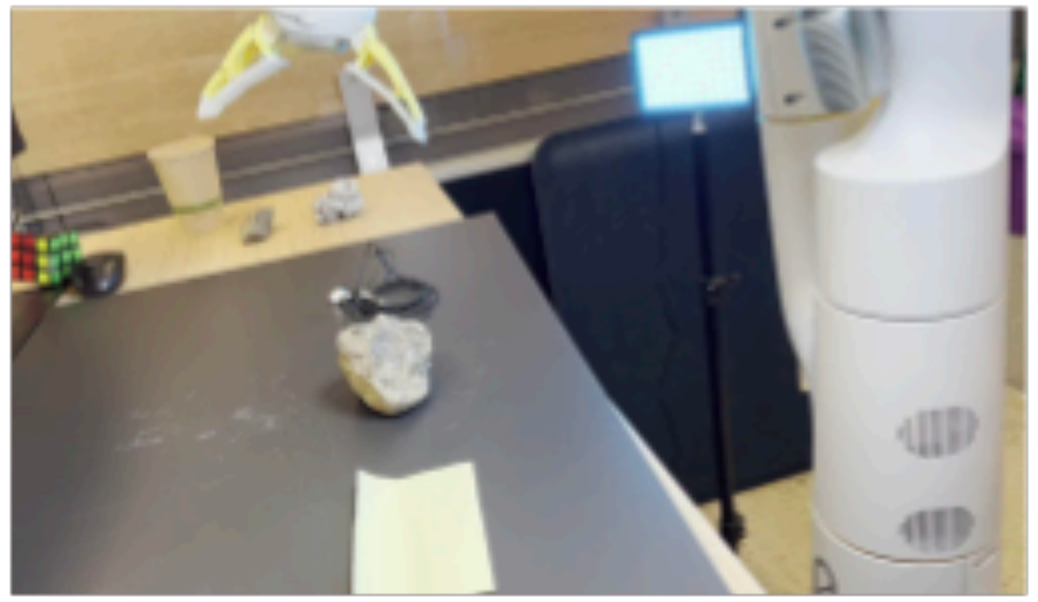
Prompt:
Given <img> Instruction:
Move all the objects together.
Prediction:
Plan: move green can near green rice chip bag.
Action: 1 128 126 127 135 123 119 127

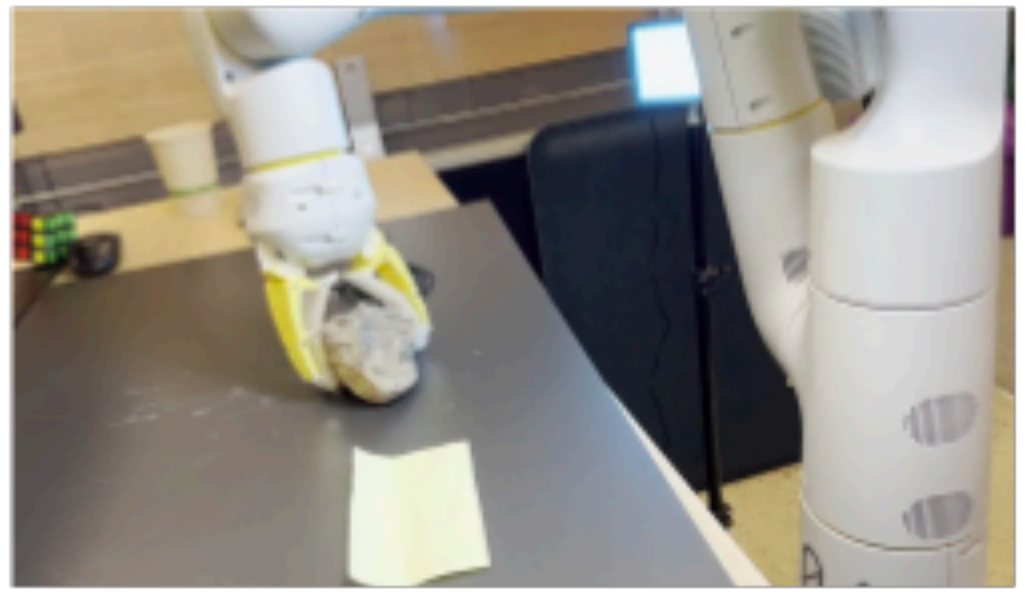
Prompt:
Given <img> Instruction:
Pick the object that is different from all other objects
Prediction:
Plan: pick rxbar chocolate. Action: 1 128 129 125 131 125 128 127



Prompt:
Given <img> Instruction:
Move the green objects together.
Prediction:
Plan: move green can near green rice chip bag.
Action: 1 130 129 121 131 127 128 127

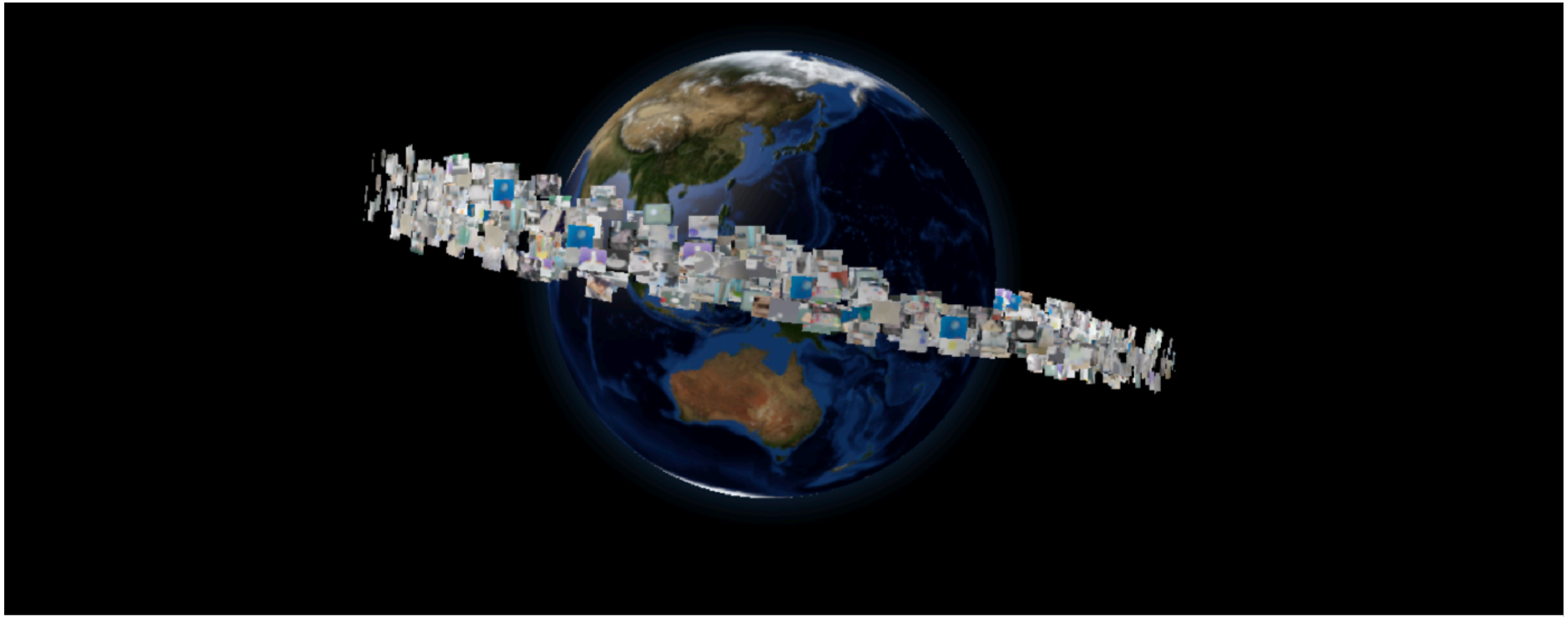
Prompt:
Given <img> I need to hammer a nail, what object from the scene might be useful?
Prediction:
Rocks. Action: 1 129 138 122 132 135 106 127





Results: Chain-of-Thought with RT-2-PaLM-E

Project idea: Open-source VLMs on RT-X data



Open X-Embodiment: Robotic Learning Datasets and RT-X Models

Open X-Embodiment Collaboration

Embodied Reasoning Through Planning with Language and Vision Foundation Models

Georgia Tech CS 7643/4644: Deep Learning
Fei Xia, Google DeepMind
11/7/2023

Google DeepMind