Applying Lessons from AI to Robot Learning

Karol Hausman, Quan Vuong

Google DeepMind

Stanford University

LLMs!

vs


CHALLENGES
OF ROBOTICS

Guidance along the way

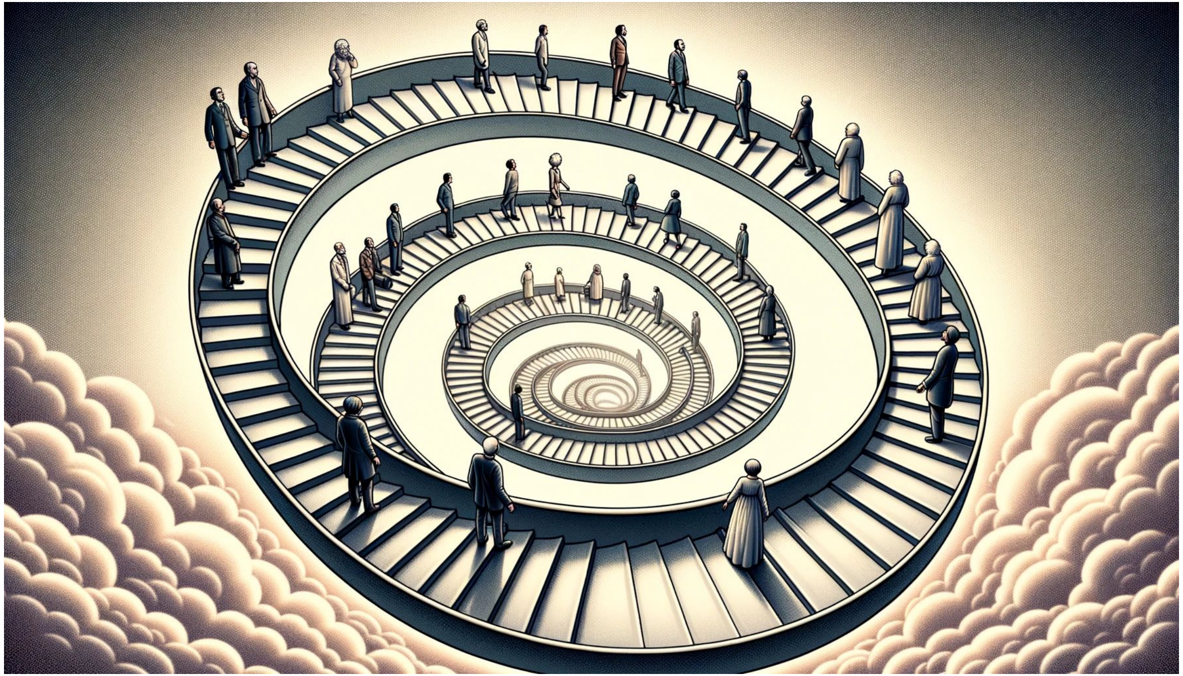


Illustration of a spiral staircase where each step represents a significant technological milestone in computer science and AI. As the staircase ascends, we see figures from the past observing and interacting with the figures from the present, symbolizing the repetitive and interconnected nature of history and innovation.

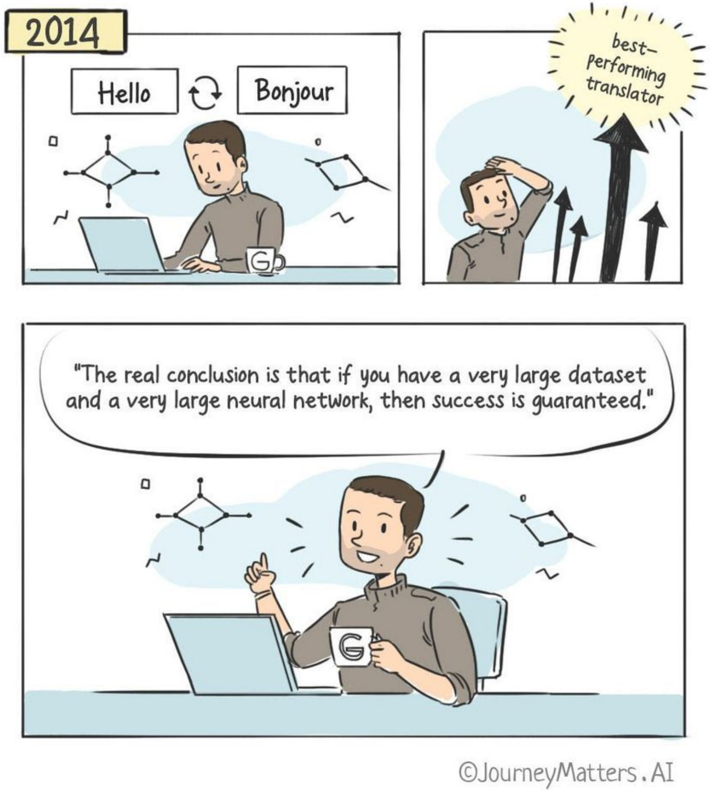
AI Lessons

5 lessons:

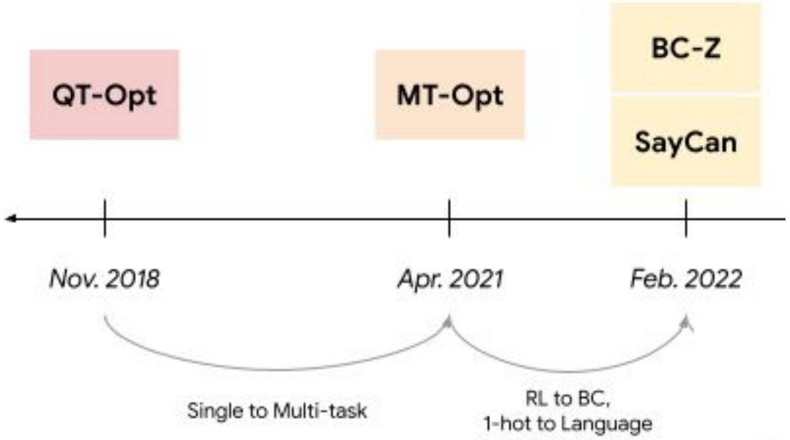
1.
2.
3.
4.
5.

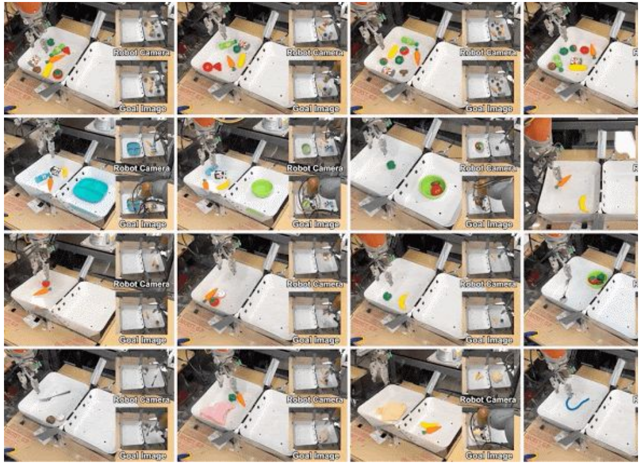
Lesson 1: Success is guaranteed

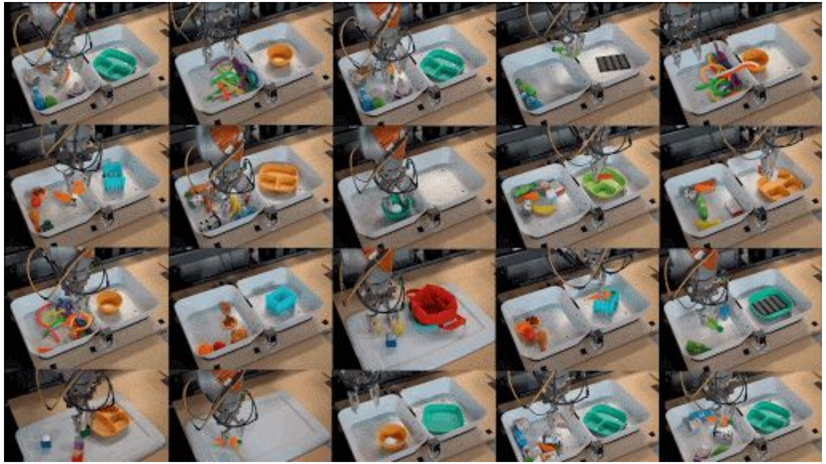
"The real conclusion is that if you have a very large dataset and a very large neural network, then success is guaranteed."



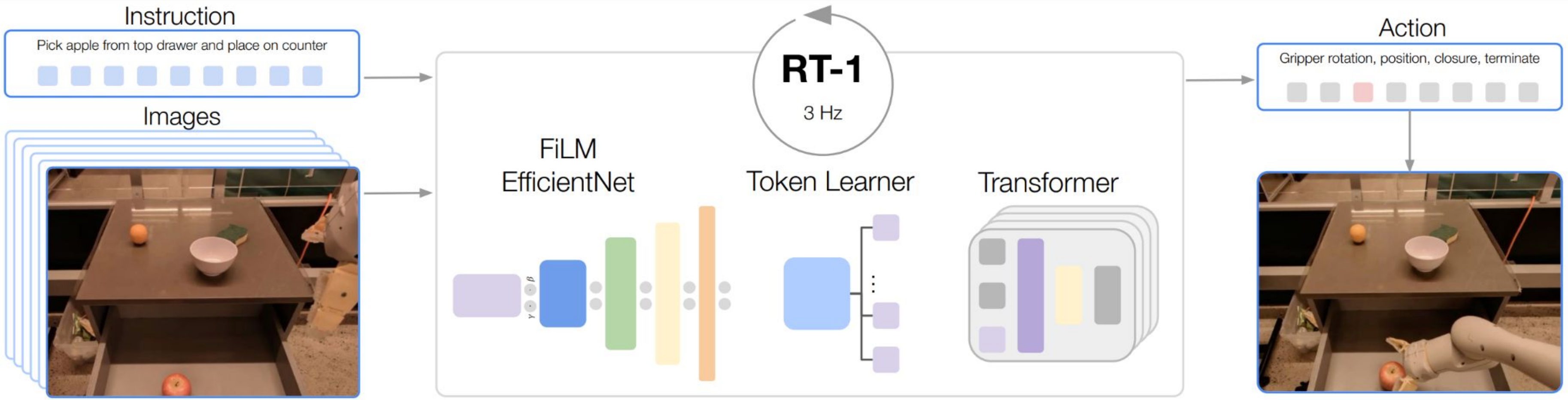
Difficult beginnings





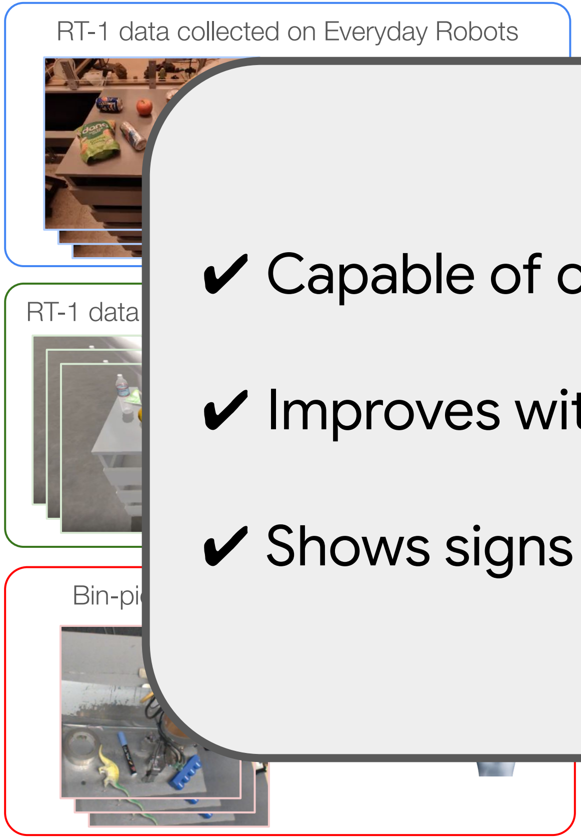


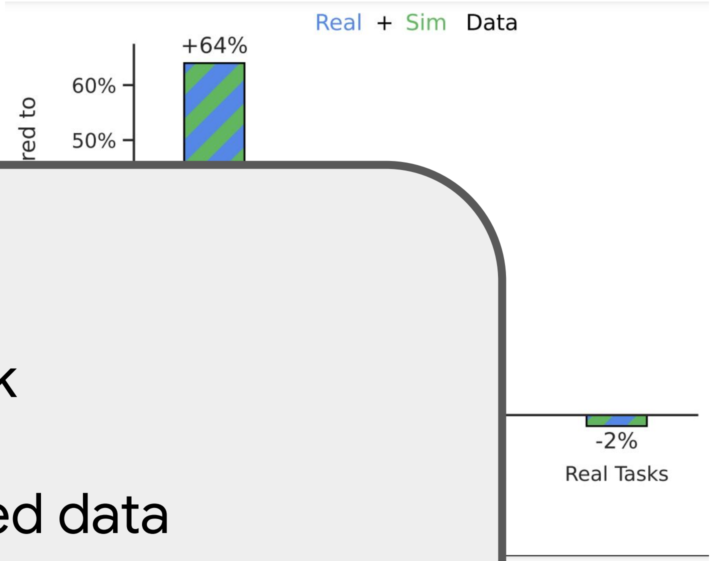
Robotics Transformer



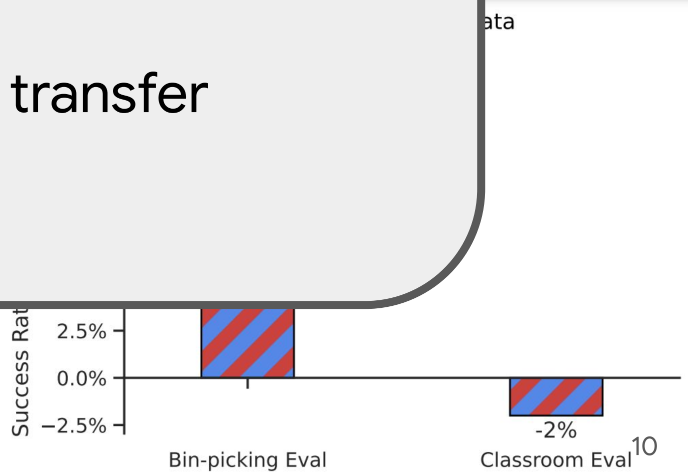
- 130000 episodes, 13 robots, over 17 months, 700 tasks
- Tokenized input and outputs
- Decoder only transformer, sparse categorical entropy objective
- Image tokenizer: Pre-trained film efficient net backbone
- Token learner for compression/ faster inference

Performance of Robotics Transformer



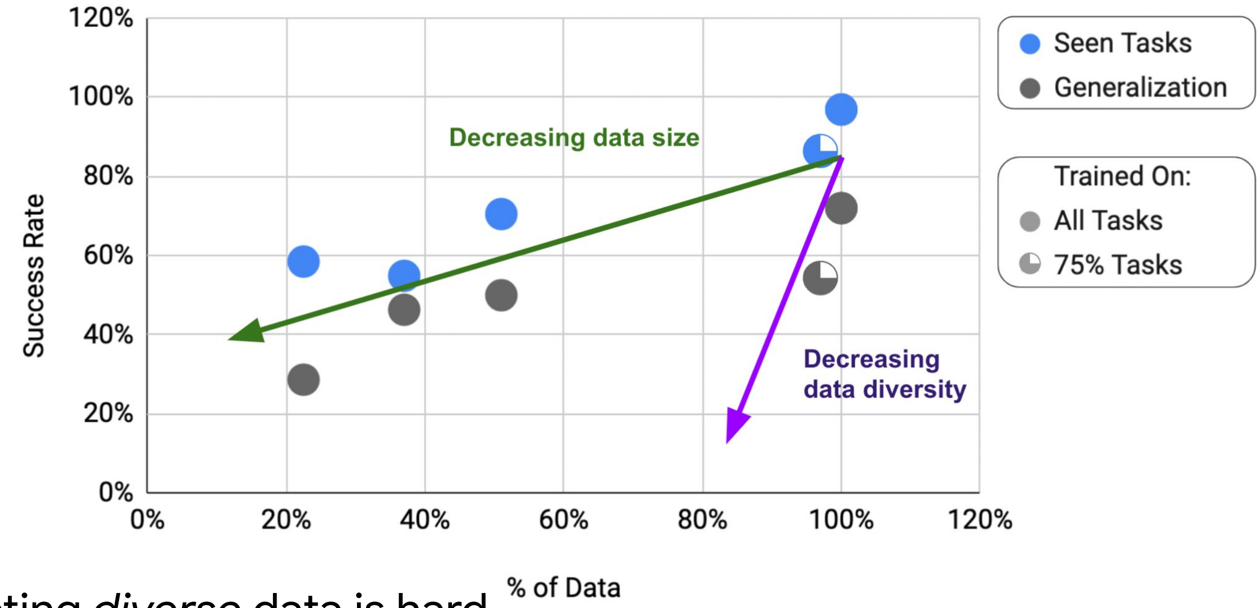


- ✔ Capable of overfitting training task
- ✔ Improves with addition of simulated data
- ✔ Shows signs of cross-embodiment transfer



The challenge with RT-1-like approaches

- Performance is strongly predicated on *training data diversity*.



- Collecting *diverse* data is hard.
- Collecting data for objects and environments that are functionally identical is a waste of time and resource.

ROSIE: **Ro**bot Learning with **S**emantically **I**magined **E**xperience

https://diffusion-rosie.github.io

AI Lessons

5 lessons:

1. Success is guaranteed
2.
3.
4.
5.

Lesson 2: The great AI consolidation

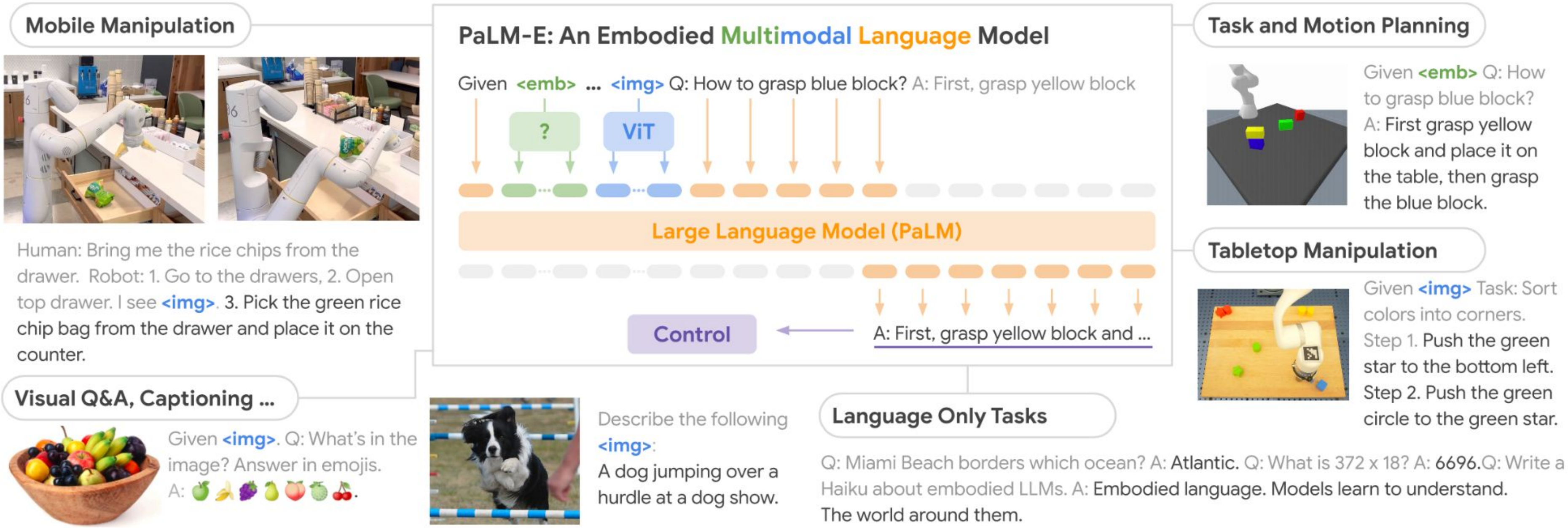
“Everyone is working with essentially the same model, so most improvements and ideas can "copy paste" rapidly across all of AI.”

1. *Approach*
2. *Architecture*
3. *Data format*
4. *Weights*

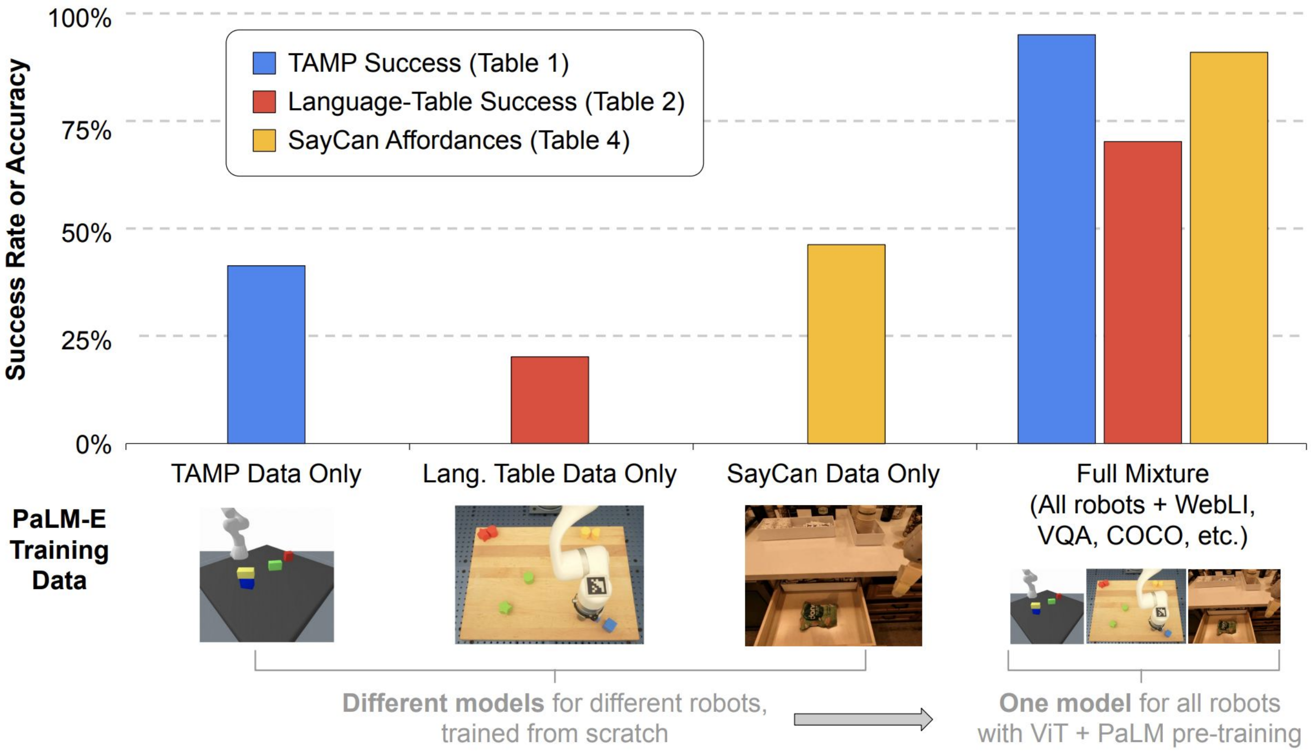
PaLM-E: An Embodied Multimodal Language Model

PaLM-E: An Embodied Multimodal Language Model

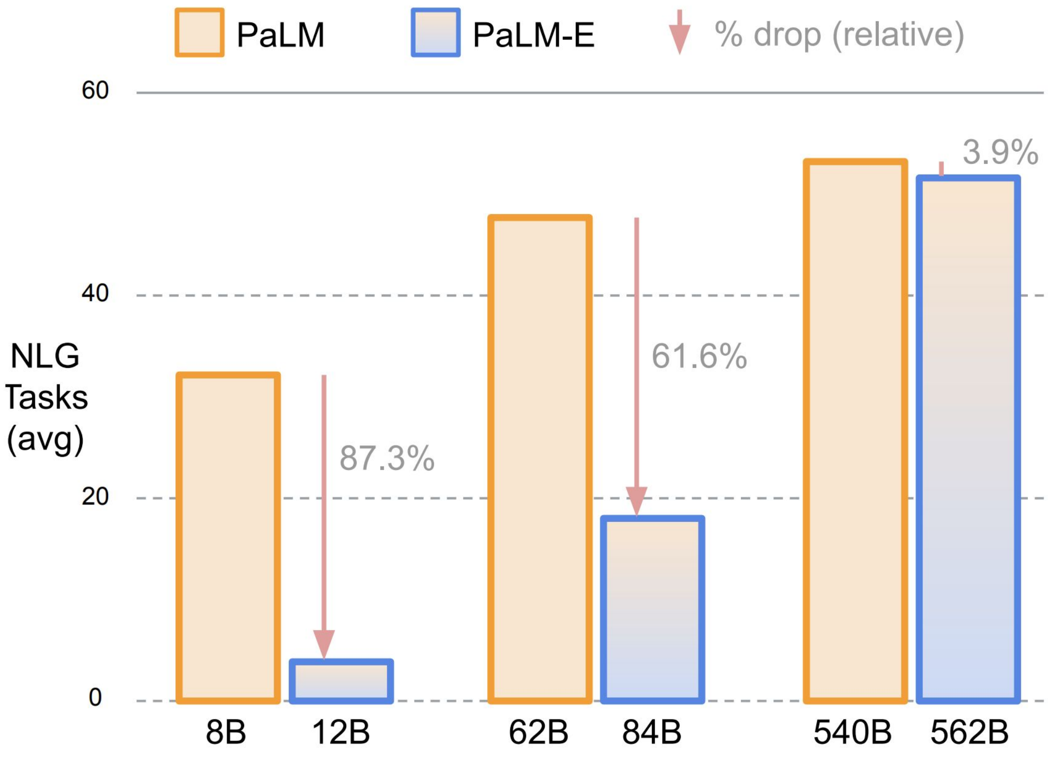
palm-e.github.io



Positive transfer



Catastrophic forgetting significantly reduced with scale



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.

Given **<img>**. Q: Who are the two teams playing in this photo, and which was the last to win a championship, and which year did they win, and who was their star player that year? A: Let's think step by step. 1. team in white is new york knick. 2. team in green is boston celtic. 3. last to win championship was boston celtic. 4. year was 2008. 5. star player was paul pierce. 6. i am not sure if this is correct. i am not basketball fan. i just google it.

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.

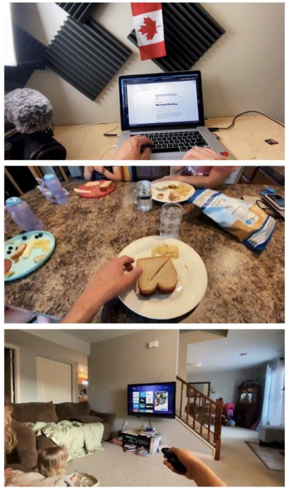
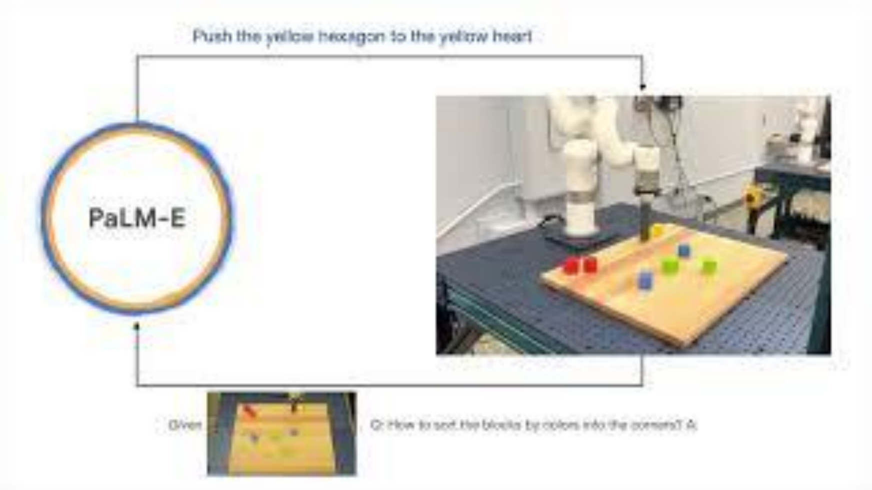


Photo 1, at 10:30 am: **<img1>**.
Photo 2, at 12:45 pm: **<img2>**.
Photo 3, at 3:45 pm: **<img3>**.
Q: I forget, what did I have for lunch, and what time was it? A: Let’s think step by step.

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

Real Robot Results



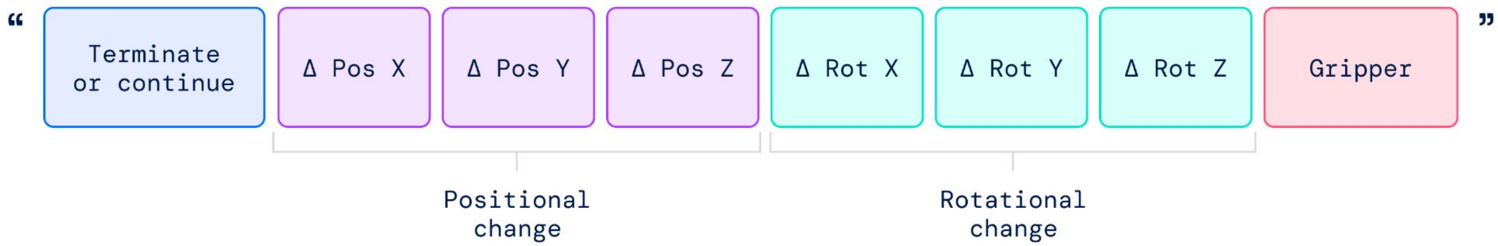
Same exact model checkpoint!
(PaLM-E can be a multi-embodiment robot brain)

The bottleneck is still on the actions.

So what about the actions?

The great AI consolidation

Representing actions in VLMs



- Actions in VLMs
  - Convert to a string of numbers
  - Example: "1 127 115 218 101 56 90 255"
  - Alternatives:
    - *Float numbers* - more tokens needed
    - *Extra-IDs, least used* language tokens
    - *Human language (left, right etc.)* - can't be directly executed on a robot

⟶ Vision-Language-Action (VLA) model!

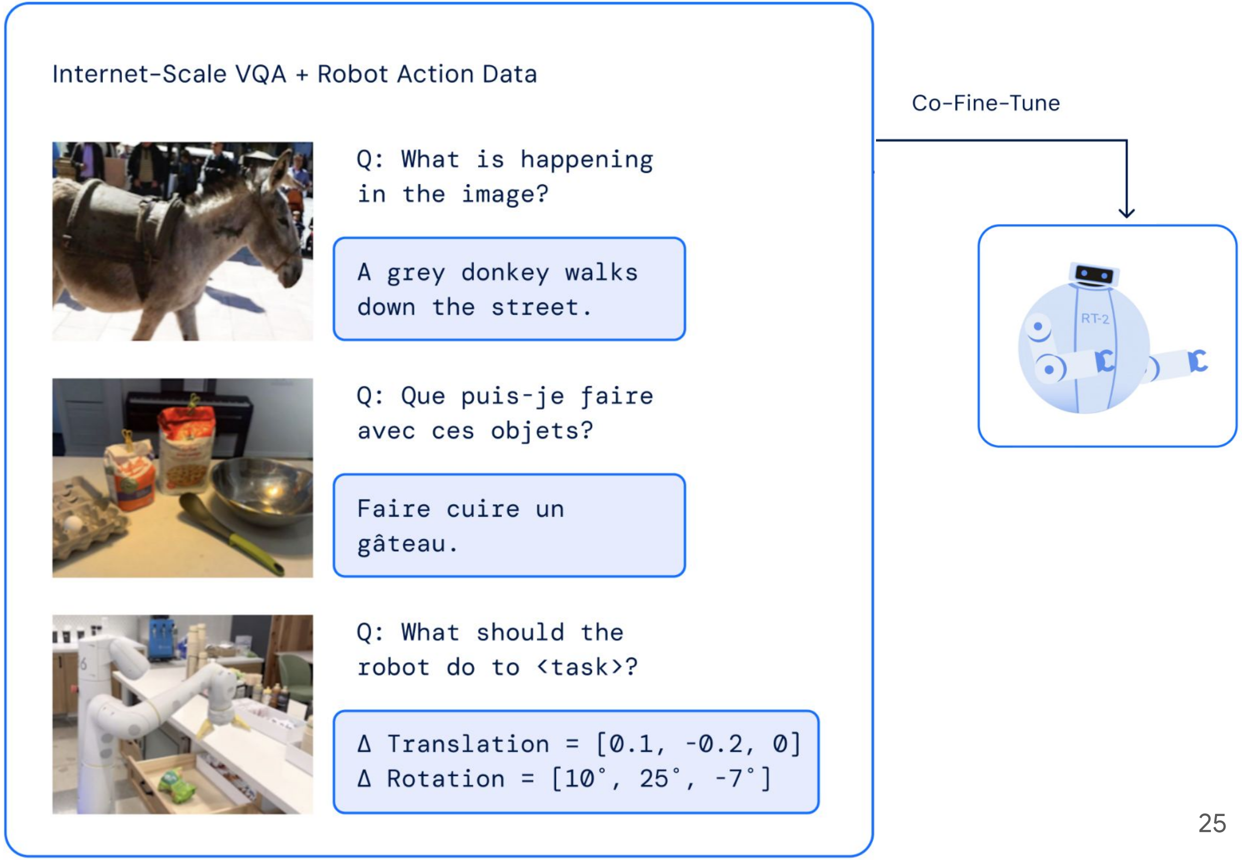
Training data and underlying models

Models

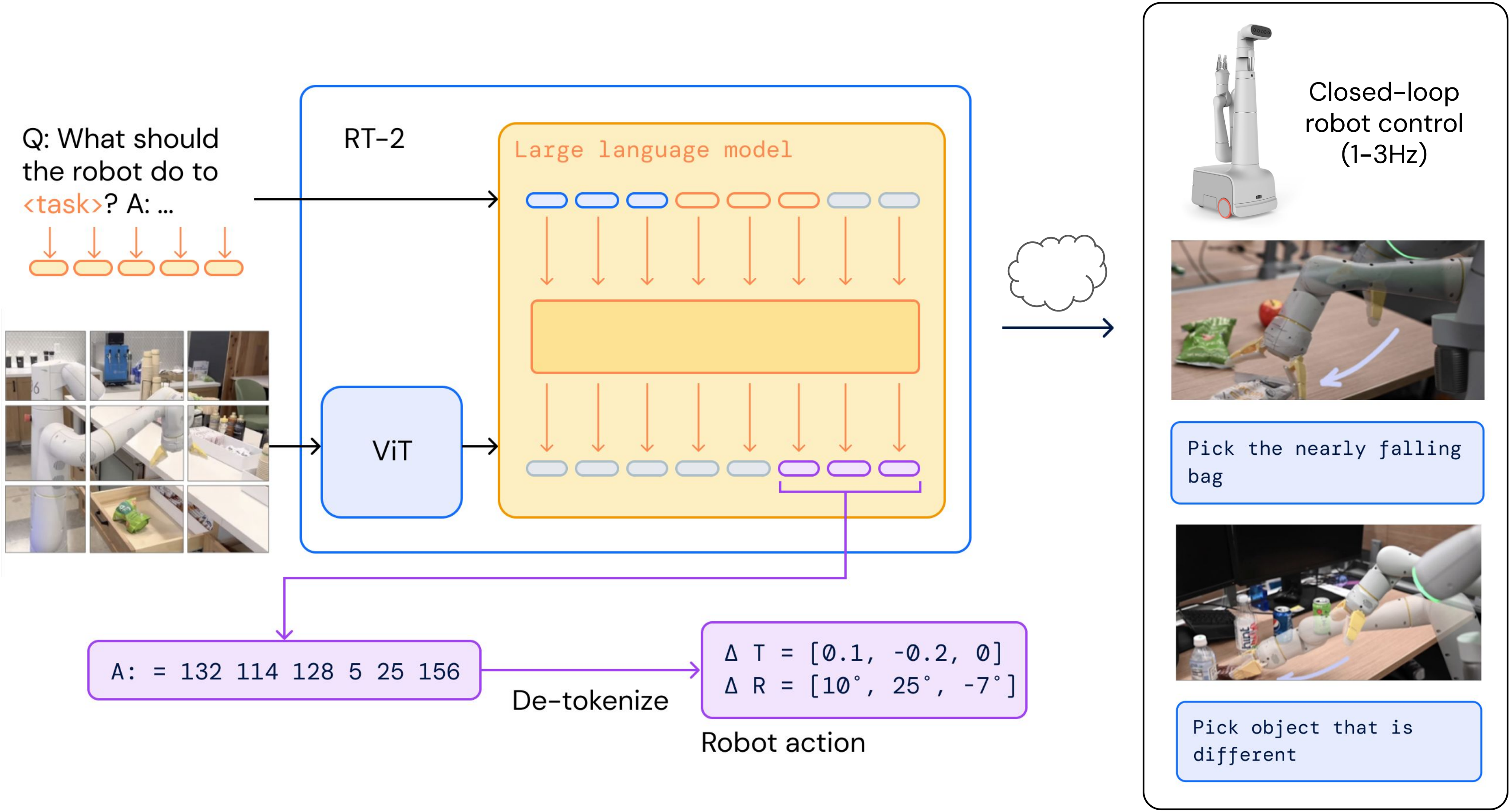
- PaLI-X (5B, 55B)
- PaLM-E (12B)

Data

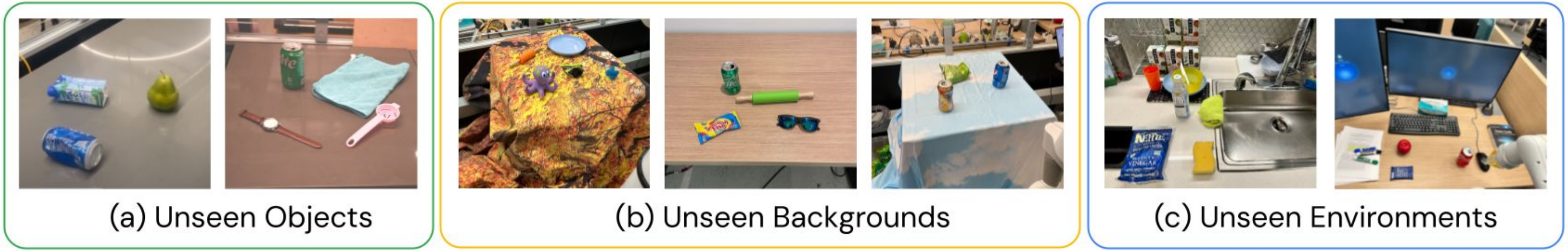
- Pretraining: Web-data
- Robot data
  - RT-1 data
    - 13 robots
    - 17 months
    - 130k demos

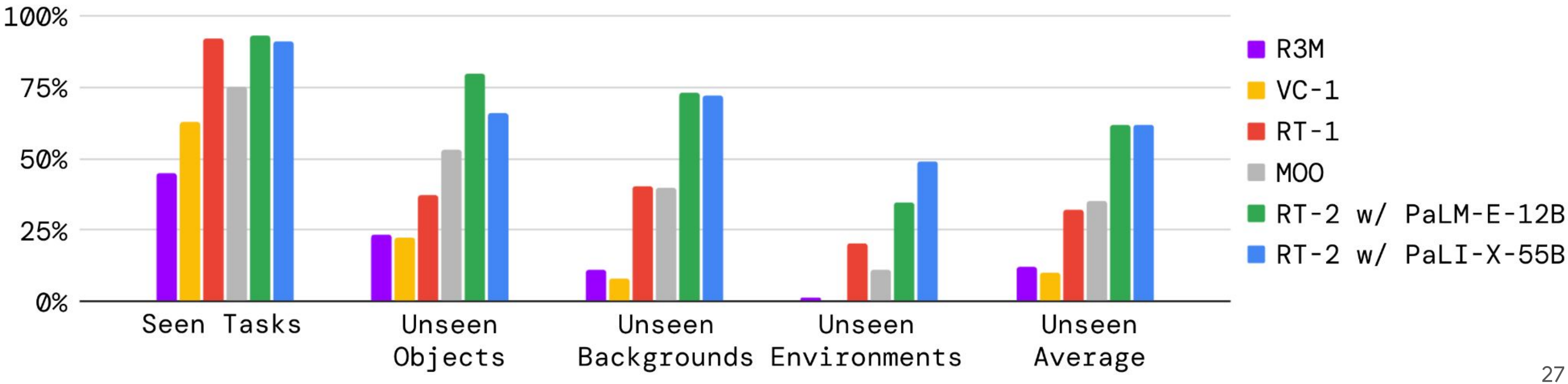


Inference

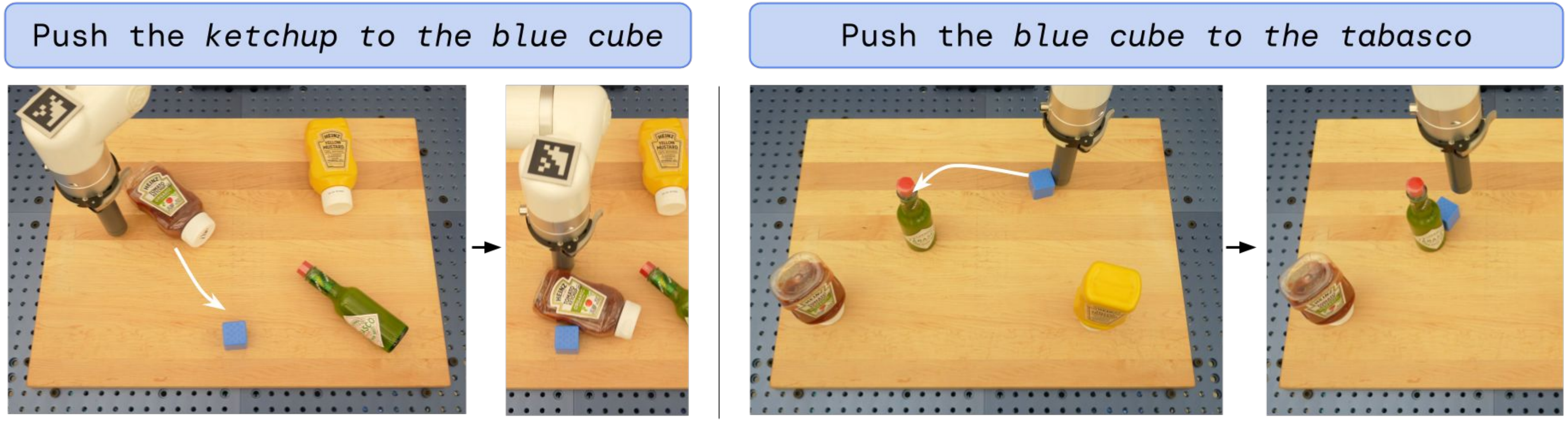

Q: What should the robot do to <task>? A: ...
RT-2
Large language model
ViT
A: = 132 114 128 5 25 156
De-tokenize
Δ T = [0.1, -0.2, 0]
Δ R = [10°, 25°, -7°]
Robot action
Closed-loop robot control (1-3Hz)
Pick the nearly falling bag
Pick object that is different

Results: Quantitative evals





Results: Language Table

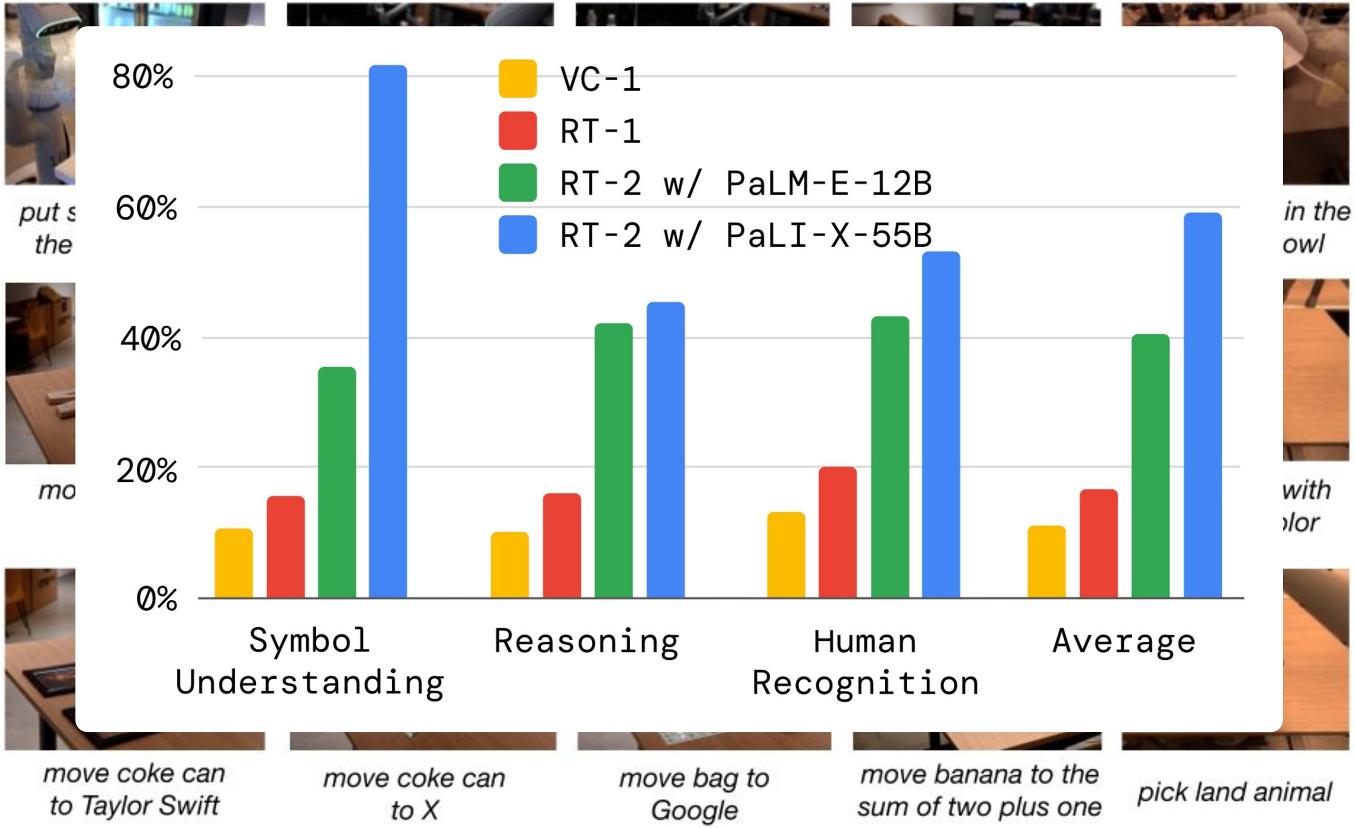


Language Table Benchmark

- 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: Emergent skills



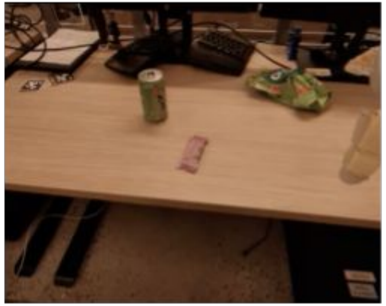
Results: Chain-of-Thought with RT-2-PaLM-E

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

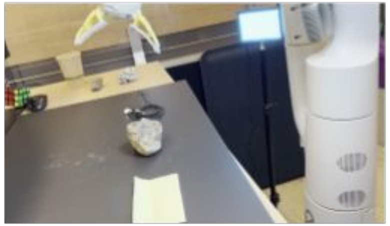
Prompt:
Given <img> Instruction:
Move all the objects together.
Prediction:
Plan: move green can near green rice chip bag.
Action: 1 128 126 127 135 123 119 127

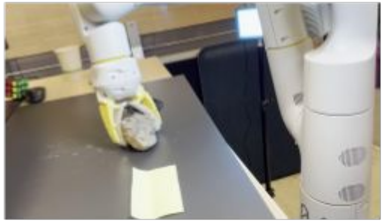
Prompt:
Given <img> Instruction:
Pick the object that is different from all other objects
Prediction:
Plan: pick rxbar chocolate. Action: 1 128 129 125 131 125 128 127

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

AI Lessons

5 lessons:

1. Success is guaranteed
2. The great AI consolidation
3.
4.
5.

Lesson 3: The power of community

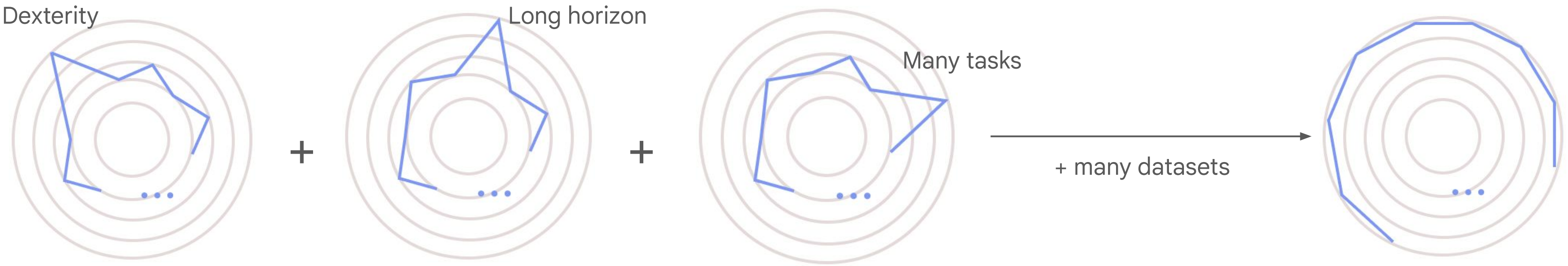
Common Crawl

Wikipedia The Free Encyclopedia

- Data sources are <u>naturally occuring</u>
- Large scale datasets unlock new capabilities in CV and NLP research

?

Robotics research



Specialist models trained on specialized datasets

Generalist models

The Open X-Embodiment Dataset



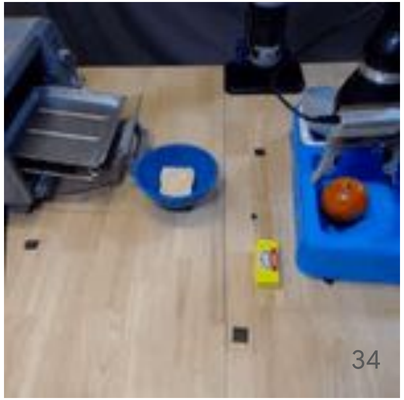
1M+ Real Robot Episodes

22 Robot Embodiments

34 Research Labs

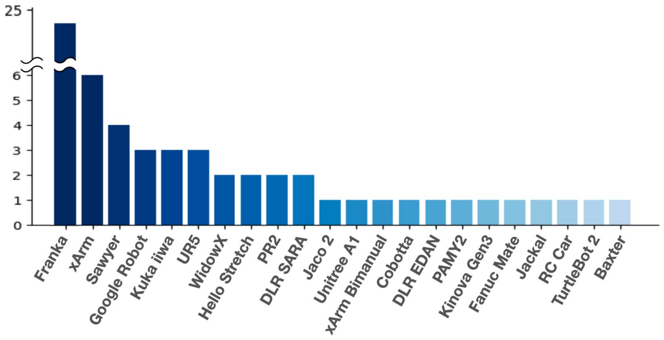
300+ Scenes



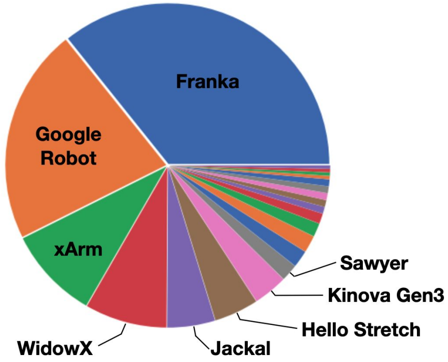





The Open X-Embodiment Dataset



Datasets per Robot Embodiment



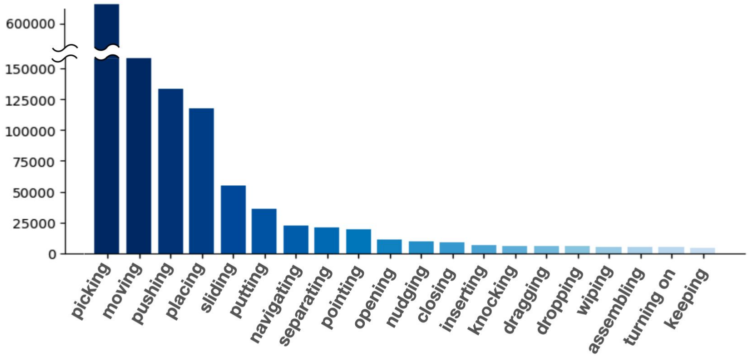
Scenes per Embodiment

Many Robots

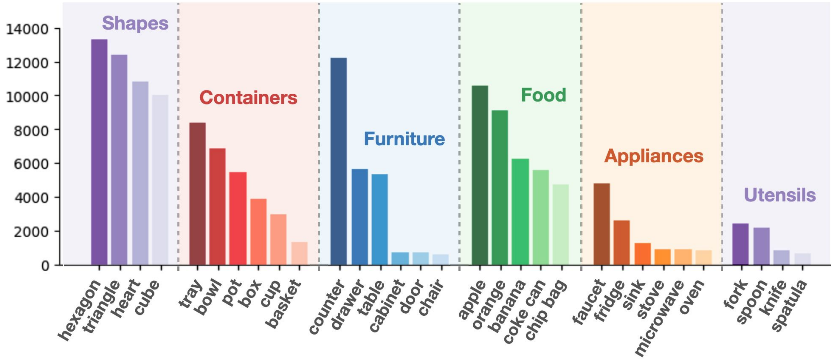
Many Scenes

Many Skills

Many Objects



Common Dataset Skills



Common Dataset Objects

Using the Open X-Embodiment Data

All data is stored in efficient tfrecord files
Download & sampling is straightforward:

```
import tensorflow_datasets as tfds

ds = tfds.load('bridge', split='train')
for episode in ds:
  for step in episode['steps']:
    image = step['observation']['image_0']
    action = step['action']
    ...
```

(see example Colab)

For details, download + Colabs, see:
robotics-transformer-x.github.com

Provide easy workflow for data filtering:

Open X-Embodiment Dataset Overview

Please check out the full data release README and Colab here: LINK

To filter the datasets based on attributes of your choice:

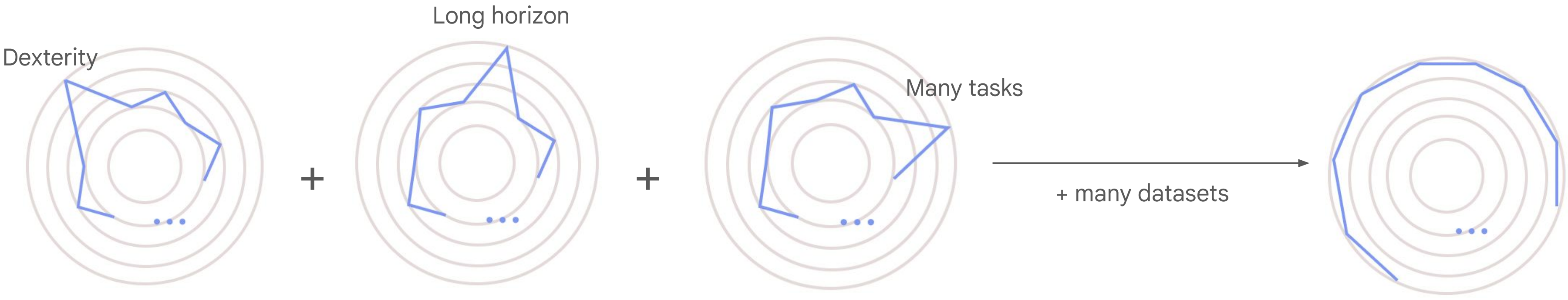
(1) select the columns you want to filter by (select row 15 and below)
(2) select "Data" --> "Filter views" --> "Create New Filter View" (you can also use one of the views we prepared)
(3) click the little filter symbol next to the column you want to filter by and choose your filtering condition (you can check the example filter views if you're unsure)
(4) once you selected your filters and are happy with the remaining datasets, copy the list of dataset names below (it auto-updates to reflect the remaining datasets) and paste it into th
(5) for convenience, we also provide a correponsing list of citations that you can directly copy into your bib file and latex document to appropriately credit the used datasets

 # Total Episodes: 1,402,930
Current Download Size (GB): 3605.41
Dataset Download List: ['fractal20220817_data', 'kuka', 'bridge', 'taco_play', 'jaco_play', 'berkeley_cable_routing', 'roboturk', 'nyu_door_opening_surprising_effectiveness', 'viola', 'berkeley_
Citation List (copy into bib file): @article{brohan2022rt,
Cite cmd (copy into Latex file): \citep{brohan2022rt, kalashnikov2018qt, walke2023bridgedata, rosete2022tacorl, mees23hulc2, dass2023jacoplay, luo2023multistage, mandlekar2019scaling, pari

Dataset	Examples	Robot	Gripper	# Episodes	File Size (GB)	Robot Morphology	Action Space	# RGB Cams	# Depth Cams	# Wrist Cams	Language Annotatio
RT-1 Robot Action		Google Robot	Default	73,499	111.06	Mobile Manipulator	EEF Position	1	1	0	Templated
QT-Opt		Kuka iiwa	Default	580,392	778.02	Single Arm	EEF Position	1	0	0	None
Berkeley Bridge		WidowX	Default	25,460	387.49	Single Arm	EEF Position	4	1	1	Natural
Freiburg Franka Play		Franka	Custom 3D print	3,242	47.77	Single Arm	EEF Position	2	2	2	Templated
USC Jaco Play		Jaco 2	Default	976	9.24	Single Arm	EEF Position	2	0	1	Templated
Berkeley Cable Routing		Franka	Default	1,482	4.67	Single Arm	EEF velocity	3	0	2	None
Roboturk		Sawyer	Default	2,144	45.39	Single Arm	EEF Position	2	1	0	Templated
NYU VINN		Hello Stretch	Default	435	7.12	Mobile Manipulator	EEF Position	1	0	1	None
Austin VIOLA		Franka	Default	135	10.4	Single Arm	EEF Position	2	0	1	Templated
Berkeley Autolab UR5		UR5	Robotiq 2F-85	896	76.39	Single Arm	EEF Position	2	1	1	None
TOTO Benchmark		Franka	Default	901	127.66	Single Arm	Joint position	1	0	0	None
Language Table		xArm	Stick for pushing	442,226	399.22	Single Arm	EEF Position	1	0	0	Natural
Columbia PushT Dataset		UR5	3D printed stick	122	2.8	Single Arm	EEF Position	5	0	1	None
Stanford Kuka Multimodal		Kuka iiwa		3,000	31.98	Single Arm	EEF Position	1	0	0	None
NYU ROT		xArm	Default	14	0.01	Single Arm	EEF Position	1	0	0	Templated
Stanford HYDRA		Franka	Default	550	72.48	Single Arm	EEF Position	2	0	1	Templated

E.g. based on robot type, # cameras, collection type, ...

Time to be speculative!



Specialist models trained on specialized datasets

Generalist models

Analogy applies to other areas of robotics:

- simulation
- evaluation
- others ?

Key Research Questions

Generalist models >> specialist models ?

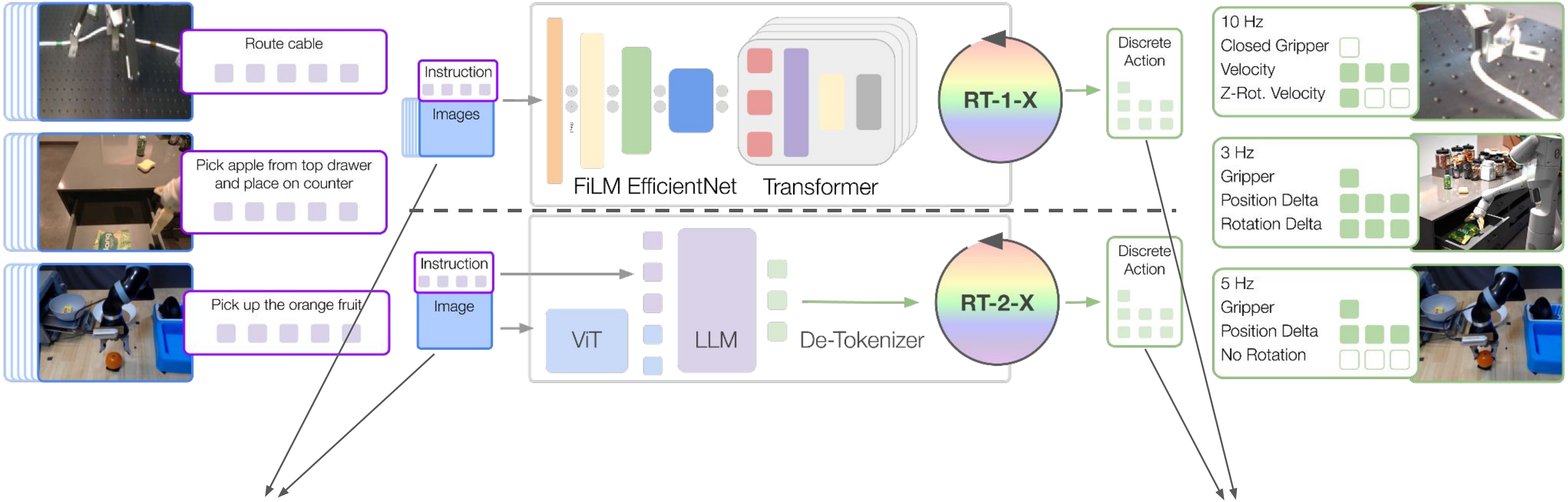
High-capacity architectures

- Absorbs knowledge from large and diverse datasets

Make minimal modifications to existing architectures

- Impacts of data scaling

Model architectures



images and a text instruction as input

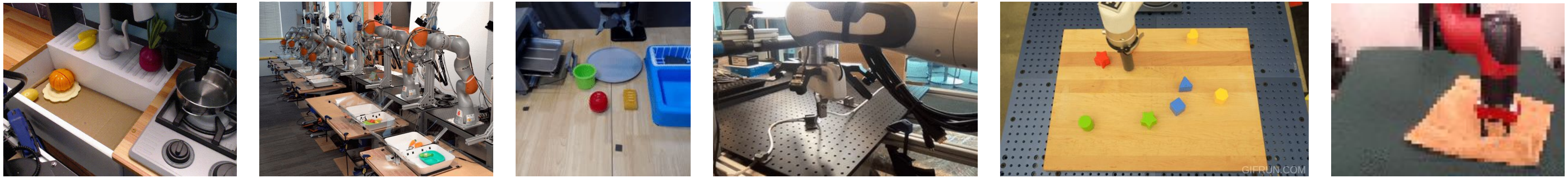
discretized actions as output

Current modeling assumptions for now

Single Arm

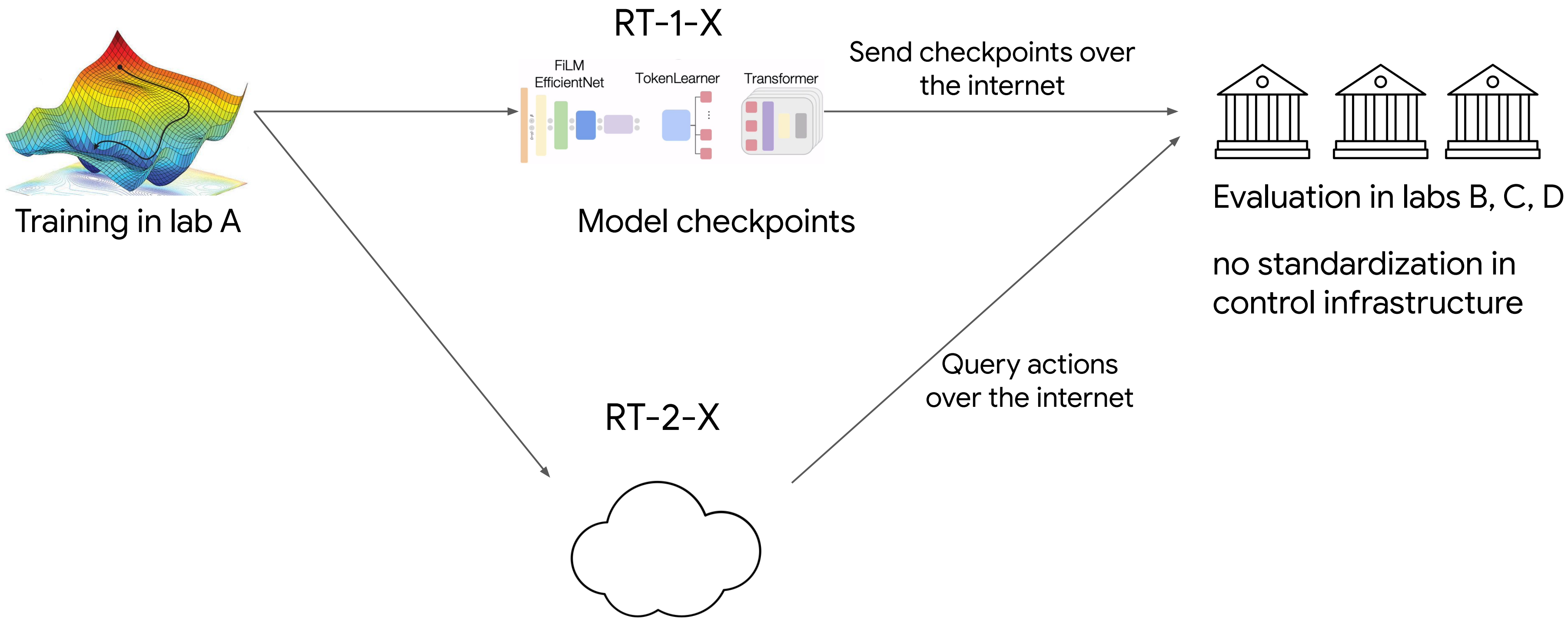
2-fingers, mostly parallel yaw

Still interesting diversity!

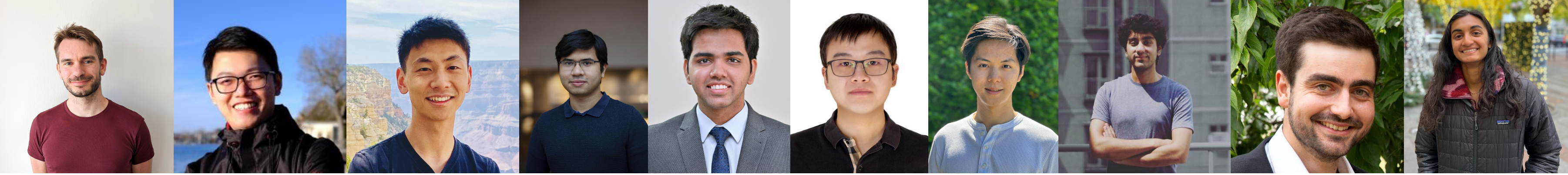


Subset of datasets with single arm

Evaluation methodologies

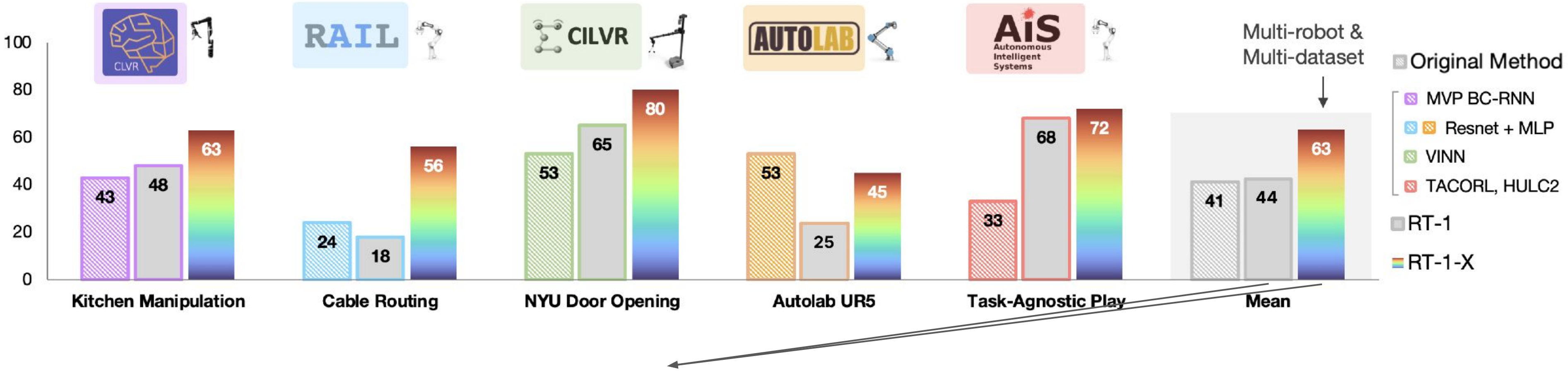


in total 3600 evaluation trials across 6 different robots



Oier Mees | Chenguang Huang | Lawrence Yunliang Chen | Nur Muhammad Mahi Shafiullah | Anant Rai | Jianlan Luo | Charles Xu | Shivin Dass | Karl Persch | Priya Sundaresan

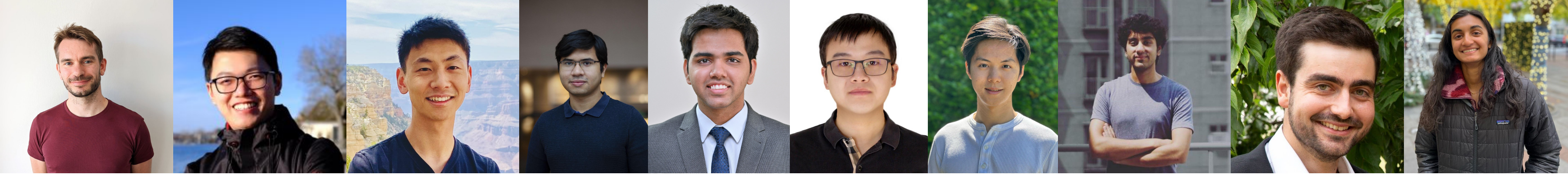
Summary



RT-1 vs RT-1-X

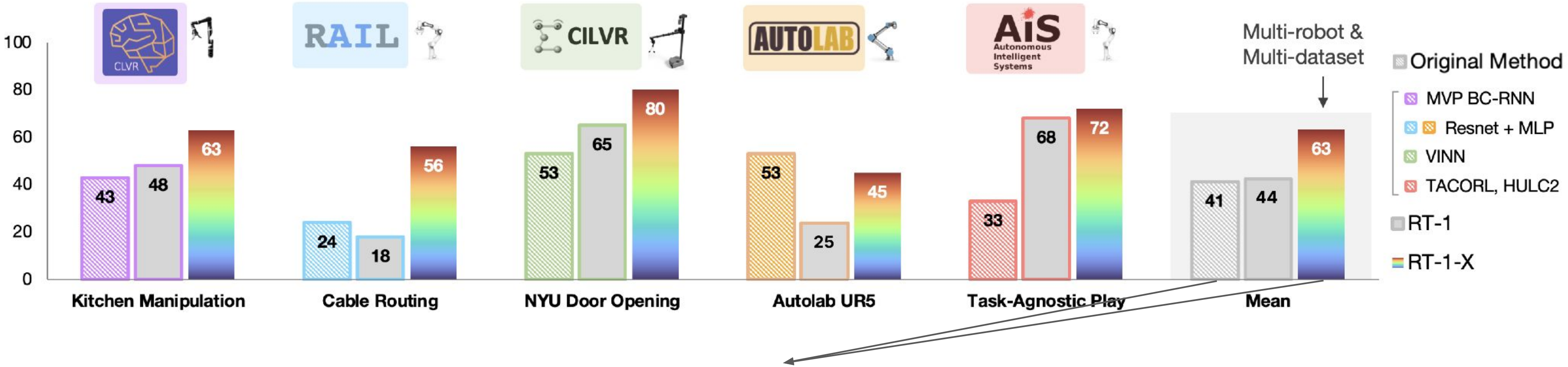
- Does training on X-Embodiment datasets improves in-distribution performance?
- Yes!

50% improvement



Oier Mees | Chenguang Huang | Lawrence Yunliang Chen | Nur Muhammad Mahi Shafiullah | Anant Rai | Jianlan Luo | Charles Xu | Shivin Dass | Karl Persch | Priya Sundaresan

Summary

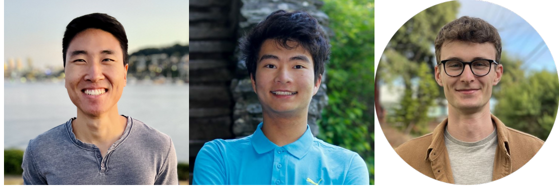


Original Methods vs RT-1-X

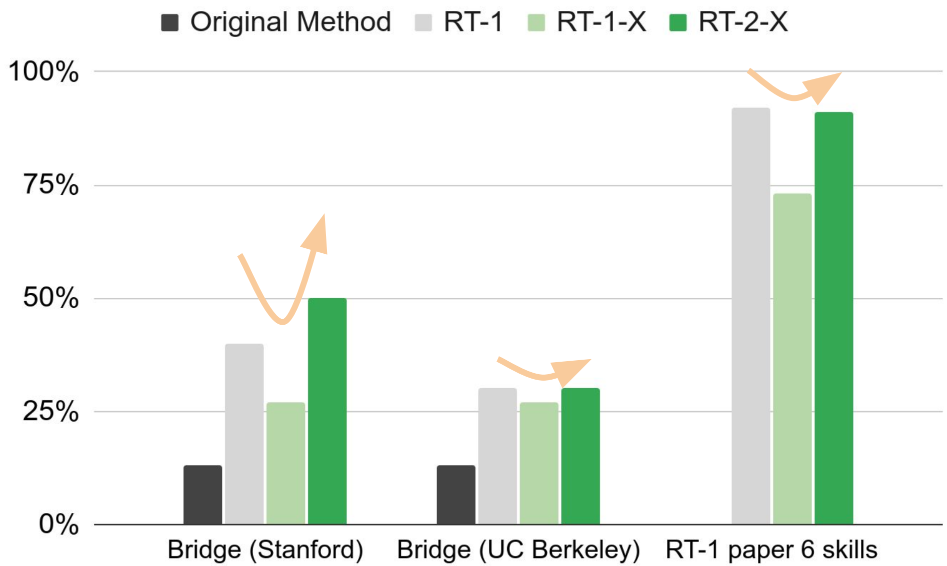
- Does generalist models outperform specialist models?
- Yes!

50% improvement

Large scale data domains



Moo Jin Kim Max Du Homer Walke

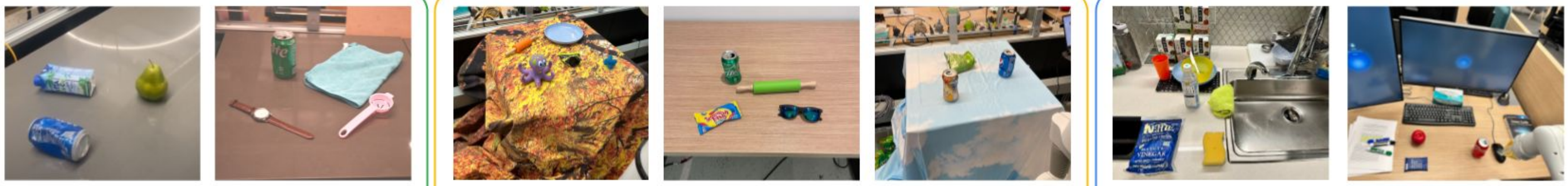


RT-1-X underfits for large datasets

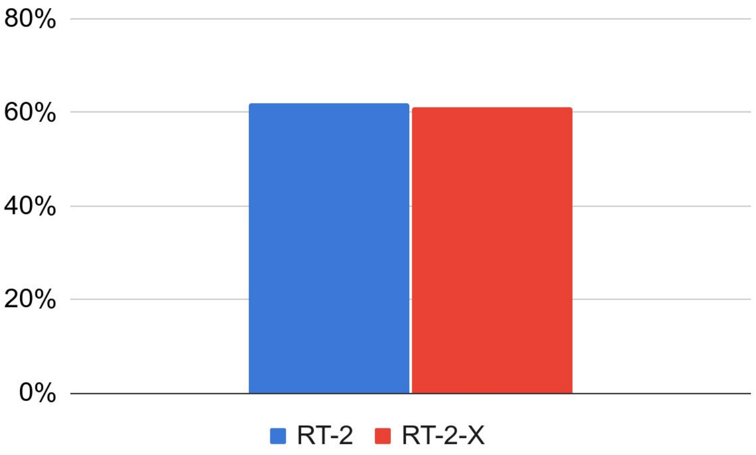
RT-2-X recovers performance

RT-2 generalization evals

More robust to distractors, on top of VLM pre-training?



(a) Unseen Objects (b) Unseen Backgrounds (c) Unseen Environments



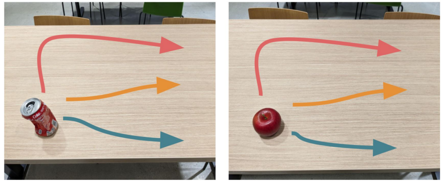
RT-2 and RT-2-X perform roughly on par

Not unexpected, since RT-2 already generalizes well along these dimensions due to its VLM backbone

Emergent skills evaluations

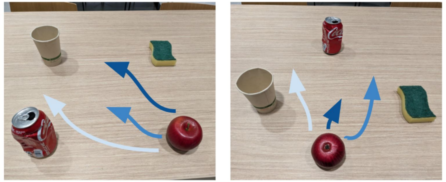
move the chip bag to the
top / bottom right of the counter

move to top right /
right / bottom right



Absolute Position Understanding

move apple between coke and cup /
coke and sponge / cup and sponge

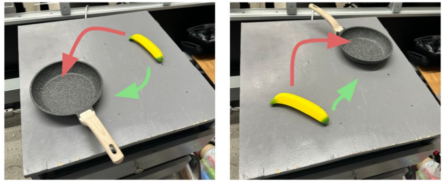


Object-Relative Position Understanding

put apple on cloth /
move apple near cloth

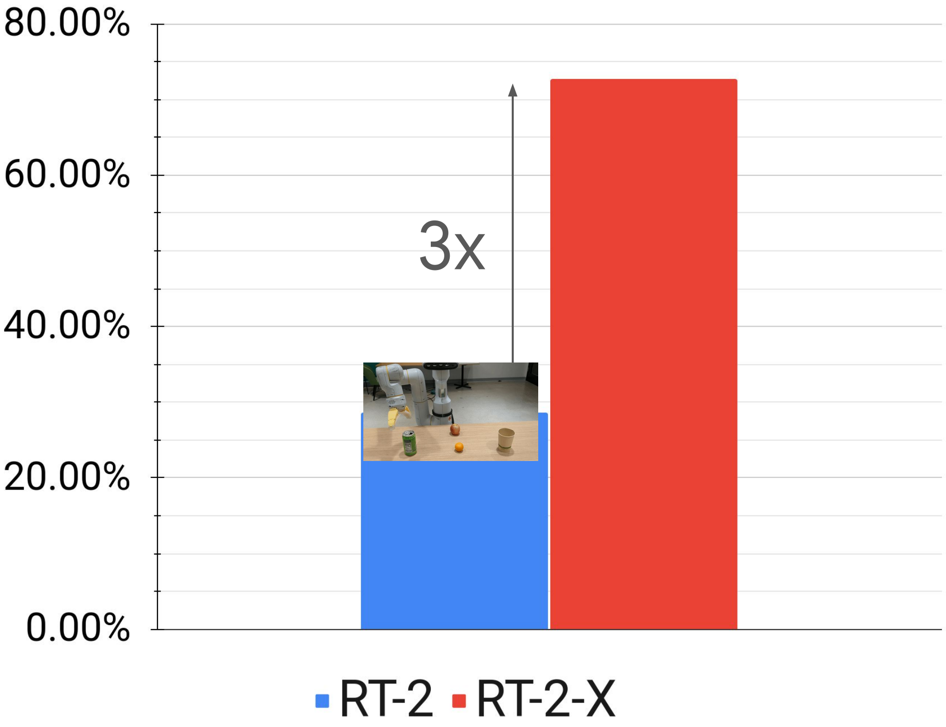
put orange into the pot /
move orange near pot

put banana on top of the pan /
move banana near pan



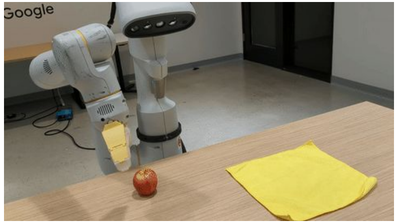
Preposition modulates low-level motion

Emergent skills evaluations

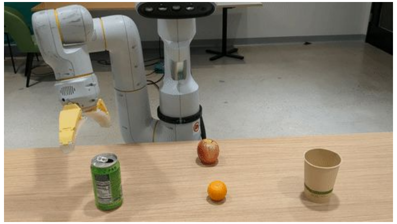


RT-2-X outperforms RT-2 by 3x in emergent skill evaluations

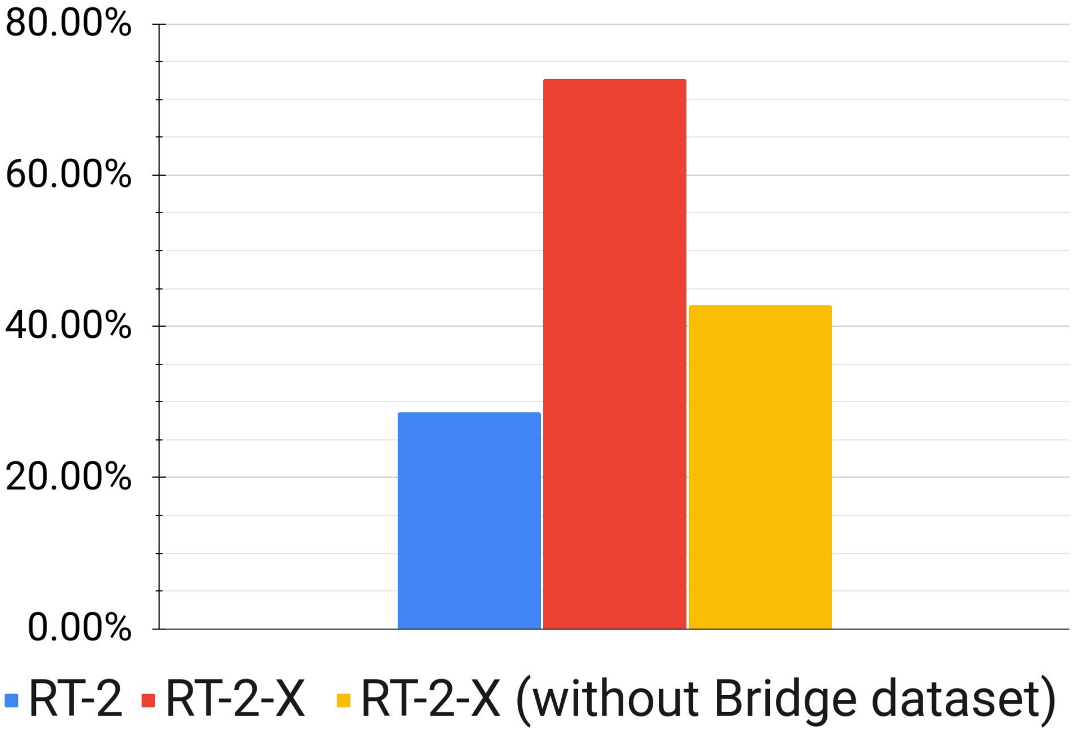
put apple on cloth



move apple near cloth



Ablations



red vs orange

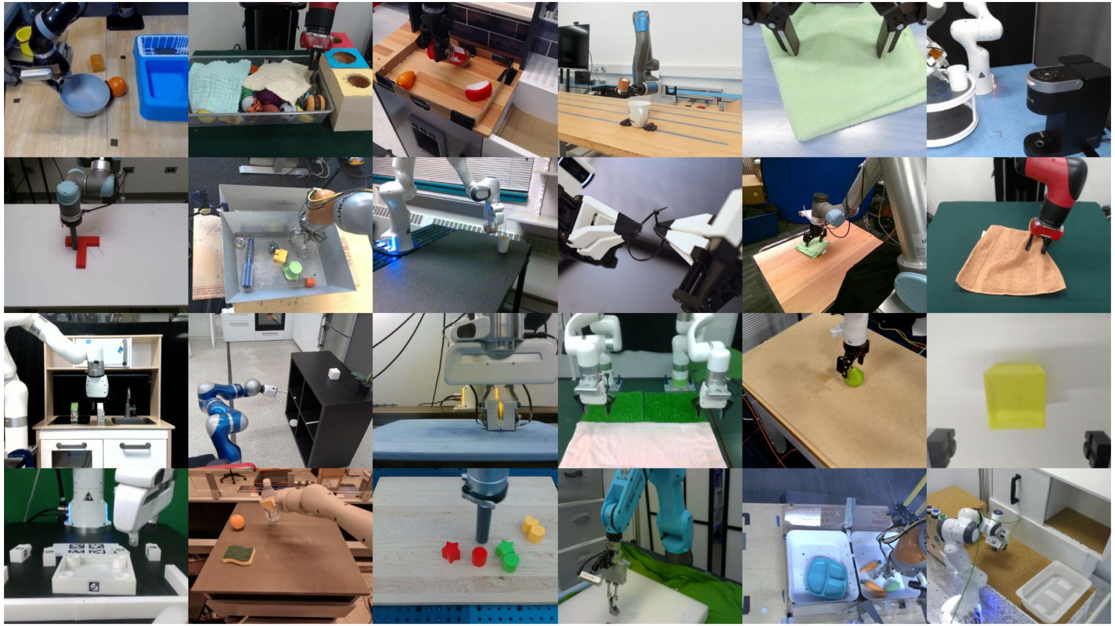
- removing Bridge dataset leads to large drop in success rate

blue vs orange

- but still almost 2x the performance
- the other datasets also help

The best is yet to come

- Further broaden participation
- Grow the initial Open X-Embodiment Dataset over time



For any inquiries, please email open-x-embodiment@googlegroups.com

AI Lessons

5 lessons:

1. Success is guaranteed
2. The great AI consolidation
3. The power of community
4.
5.

Lesson 4: Using all the data

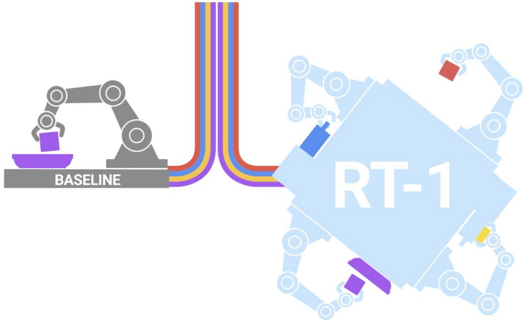
Autonomous execution

Evaluation of trained policies

RT-1 and autonomous data

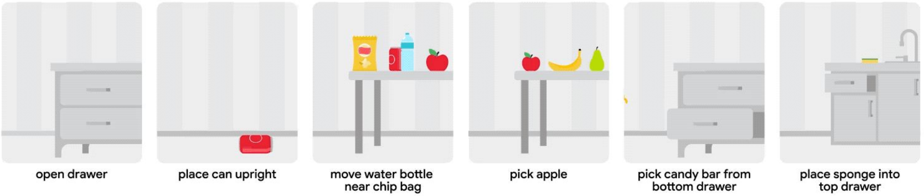
Robotics Transformer 1 (RT-1)

- Scalable model for language-conditioned tasks
- Trained with BC on demonstrations
- Autonomous / negative data unused



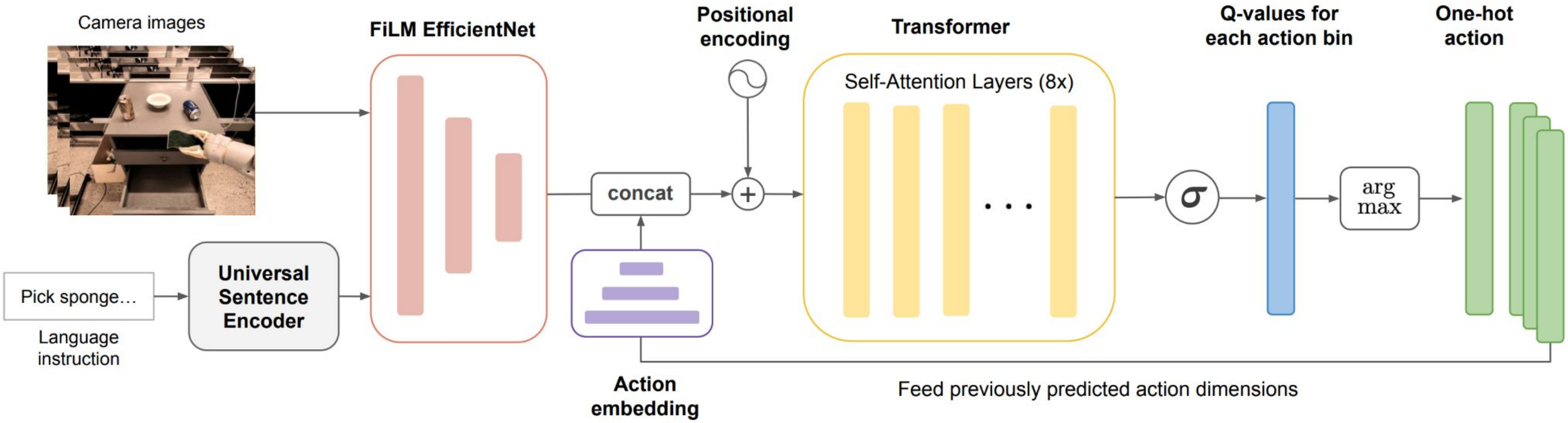
How do we get Transformer models to work with

- Introducing Q-Transformer
- Small architecture modification
- Different training procedure



RT-1 to Q-Transformer

- Same underlying architecture, add sigmoid to represent [0...1] Q-values
- Action selection the same as BC policy through argmax



Autoregressive Q-Learning

Standard Bellman Update

$$\mathcal{B}^* Q(s_t, a_t) = R(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$$

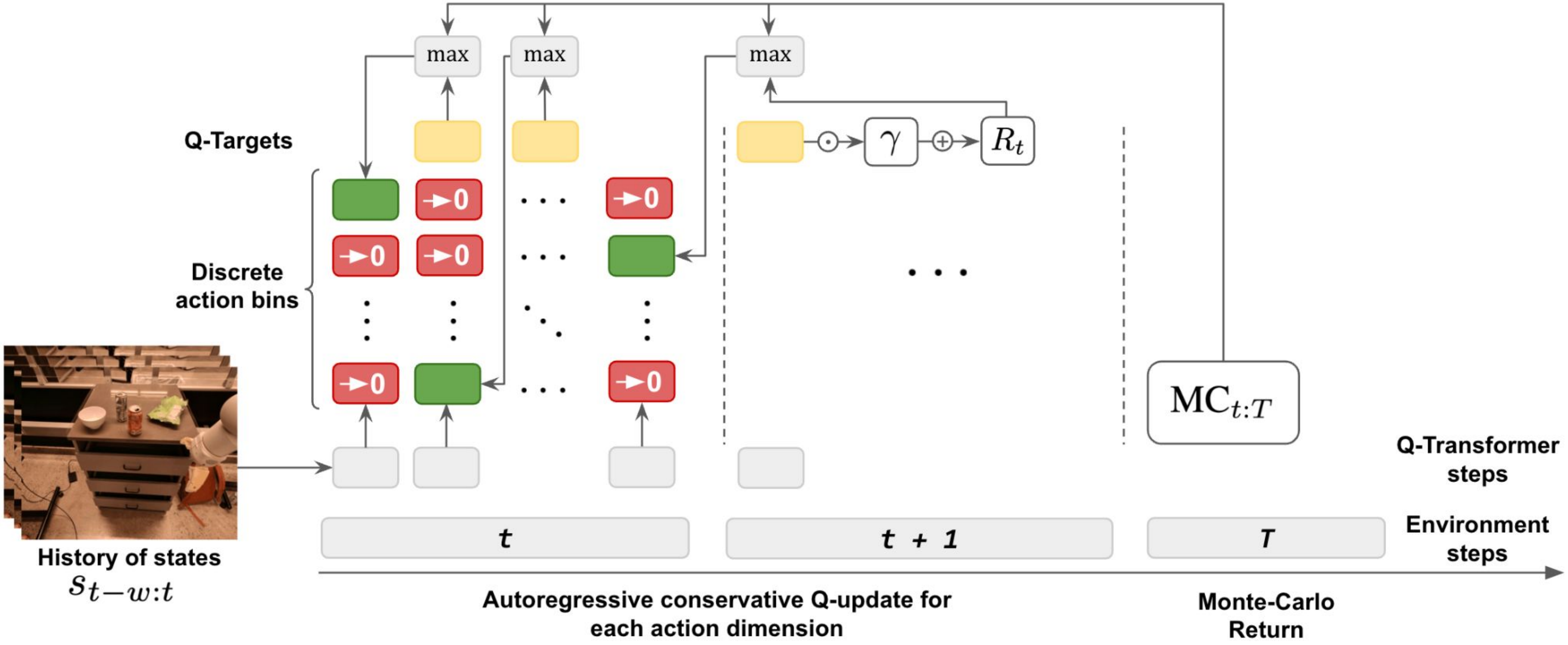
- Transformers operate on discrete tokens
- Maximization over a high number of discrete dimensions not practical

Autoregressive Bellman Update

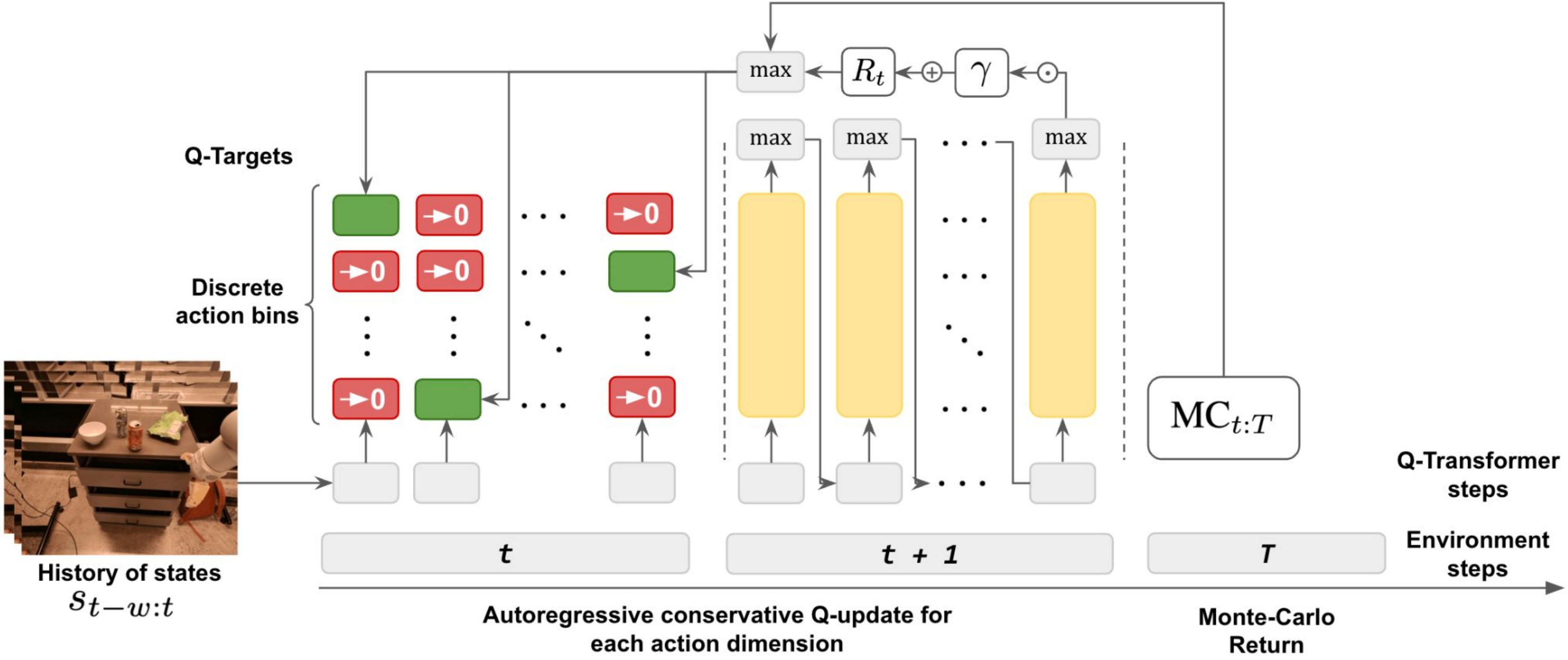
- Treat each action dimension as a separate step
- Add reward and discount next timestep after predicting the whole action

$$Q(s_{t-w:t}, a_t^{1:i-1}, a_t^i) \leftarrow \begin{cases} \max\limits_{a_t^{i+1}} \; Q(s_{t-w:t}, a_t^{1:i}, a_t^{i+1}) & \text{if } i \in \{1, \ldots, d_{\mathcal{A}} - 1\} \\ R(s_t, a_t) + \gamma \max\limits_{a_{t+1}^1} \; Q(s_{t-w+1:t+1}, a_{t+1}^1) & \text{if } i = d_{\mathcal{A}} \end{cases}$$

Autoregressive offline Q-Learning



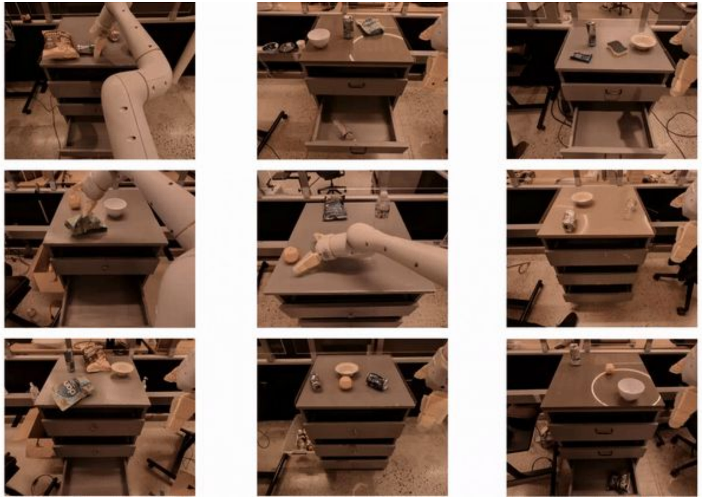
Autoregressive offline Q-Learning: “n-step” / greedy

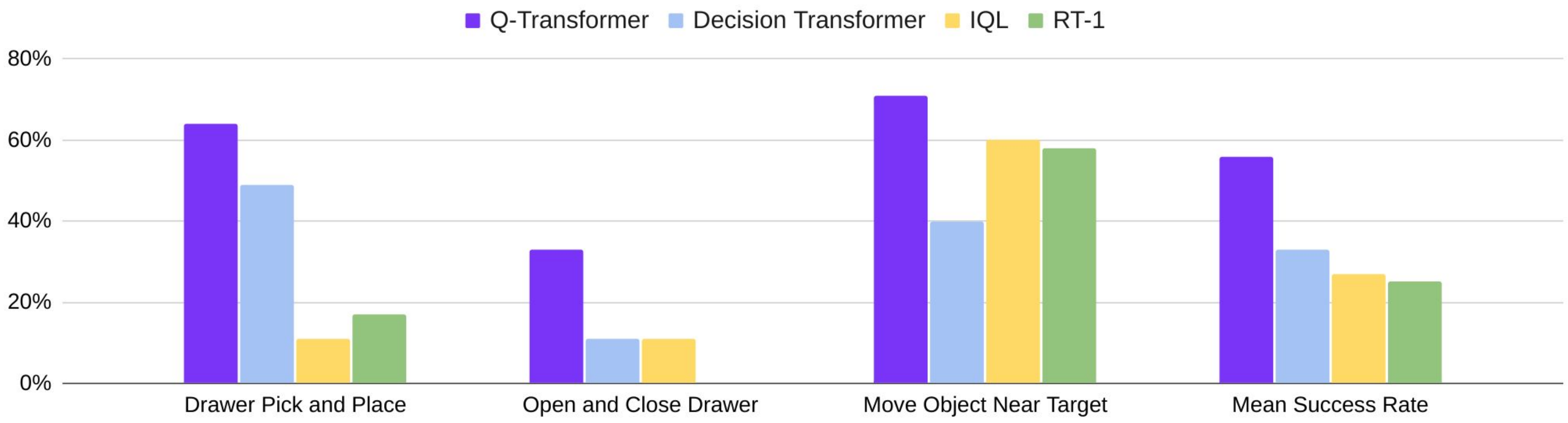


Results: Real world evaluations

Dataset

- RT-1 data limited to 100 demos / task
- Autonomous collection running BC, filtered to only contain negatives
- 38,000 demos
- 20,000 autonomous negatives

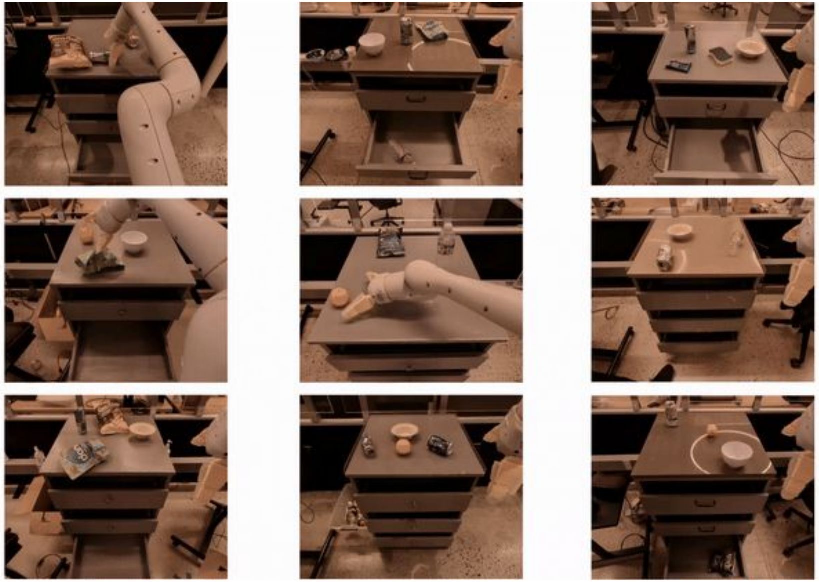




Results: Scaling up the dataset

Training on a very large dataset

- Taking all of our data
- 115,000 successes
- 185,000 negatives



Large offline dataset	Q-T	DT	RT-1
Average success rate	**88%**	78%	82%

Using value functions to grounding Language in Robotic Affordances

Using Q-values for affordances

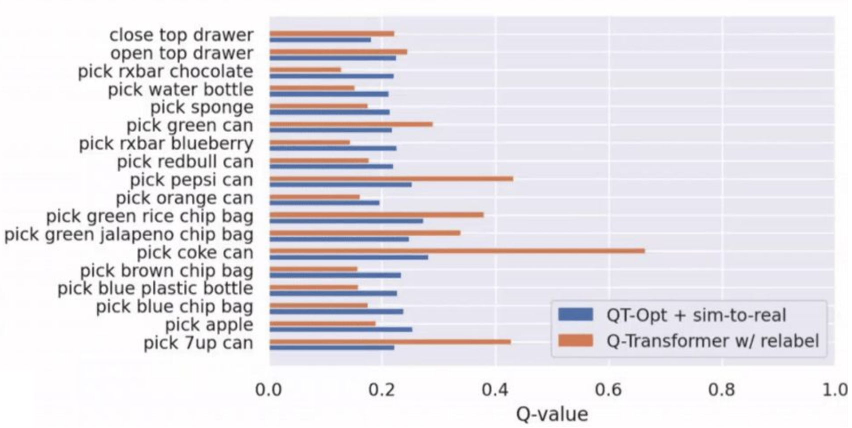
- SayCan style planning with LLMs

Relabeling

- Relabel episodes from other tasks as negatives

Model	Precision	Recall	F1
QT-Opt (sim-to-real)	0.61	0.68	0.64
Q-T w/ relabel	0.76	0.89	**0.82**
Q-T w/o relabel	0.58	0.93	0.71

Method		Success Rate	
Affordance	Execution	Planning	Execution
Q-T w/ relabel	Q-T	**93**	**93**
QT-Opt (sim-to-real)	RT-1	87	67



AI Lessons

5 lessons:

1. Success is guaranteed
2. The great AI consolidation
3. The power of community
4. Using all the data
5.

Lesson 5: The bitter lesson

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law (...)

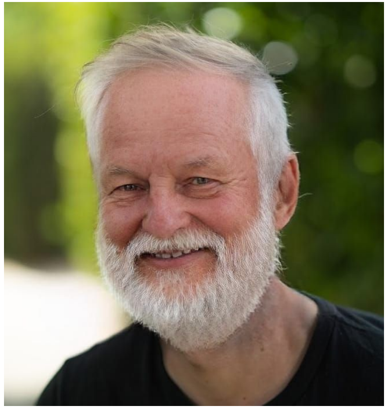
1. AI will need a lot of computation
2. Trend: computation is getting cheaper
3. Work on AI methods that leverage that trend

Bitter lesson

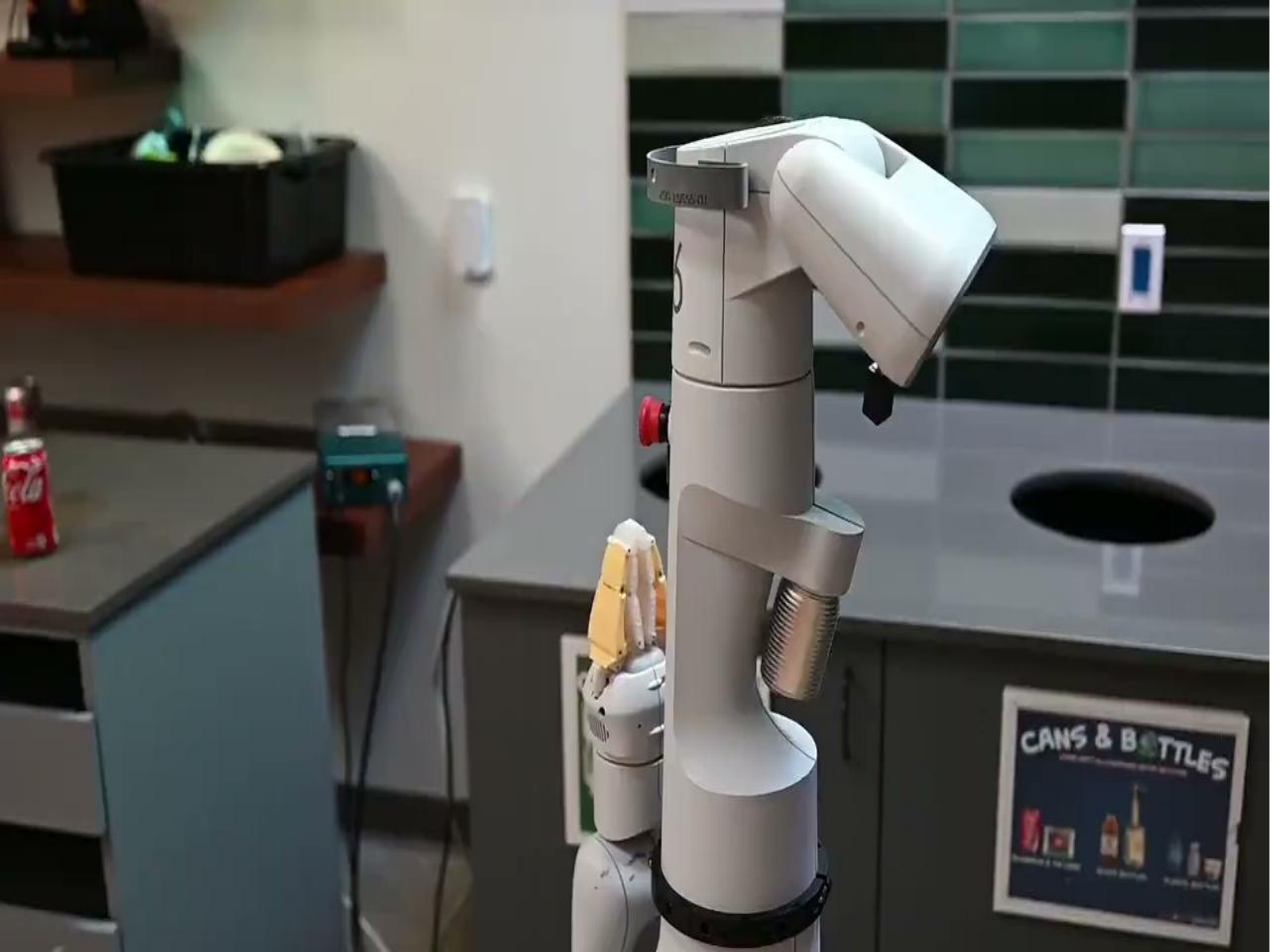
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law (...)

Bitter lesson v2

The biggest lesson that can be read from 70 years of robotics research is that general methods that leverage foundation models are ultimately the most effective.

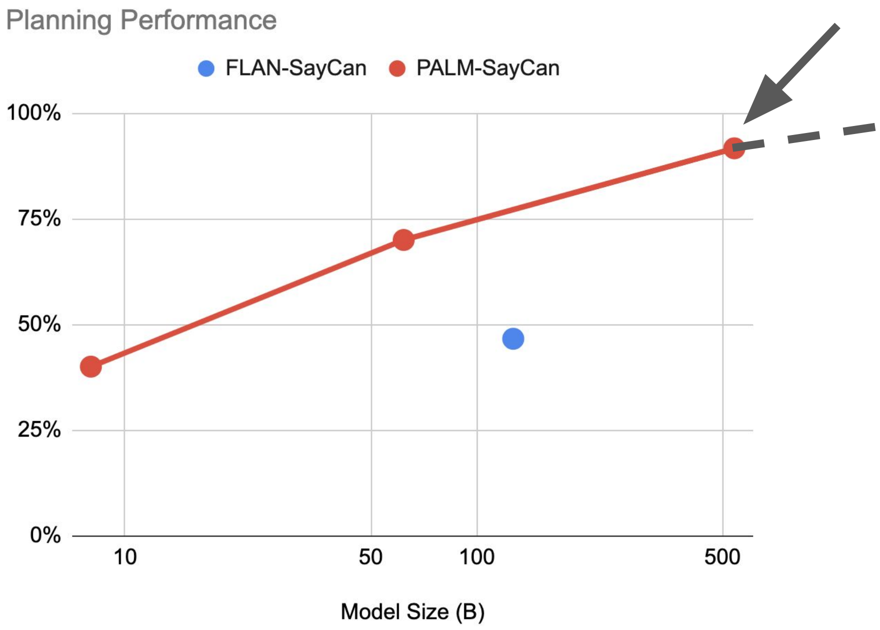


say-can.github.io



SayCan
PaLM-SayCan

Robotics performance scales with better LLMs!



PaLM-SayCan

Robotics performance scales with better LLMs!

Chain-of-thought prompting

Solves all kinds of queries:

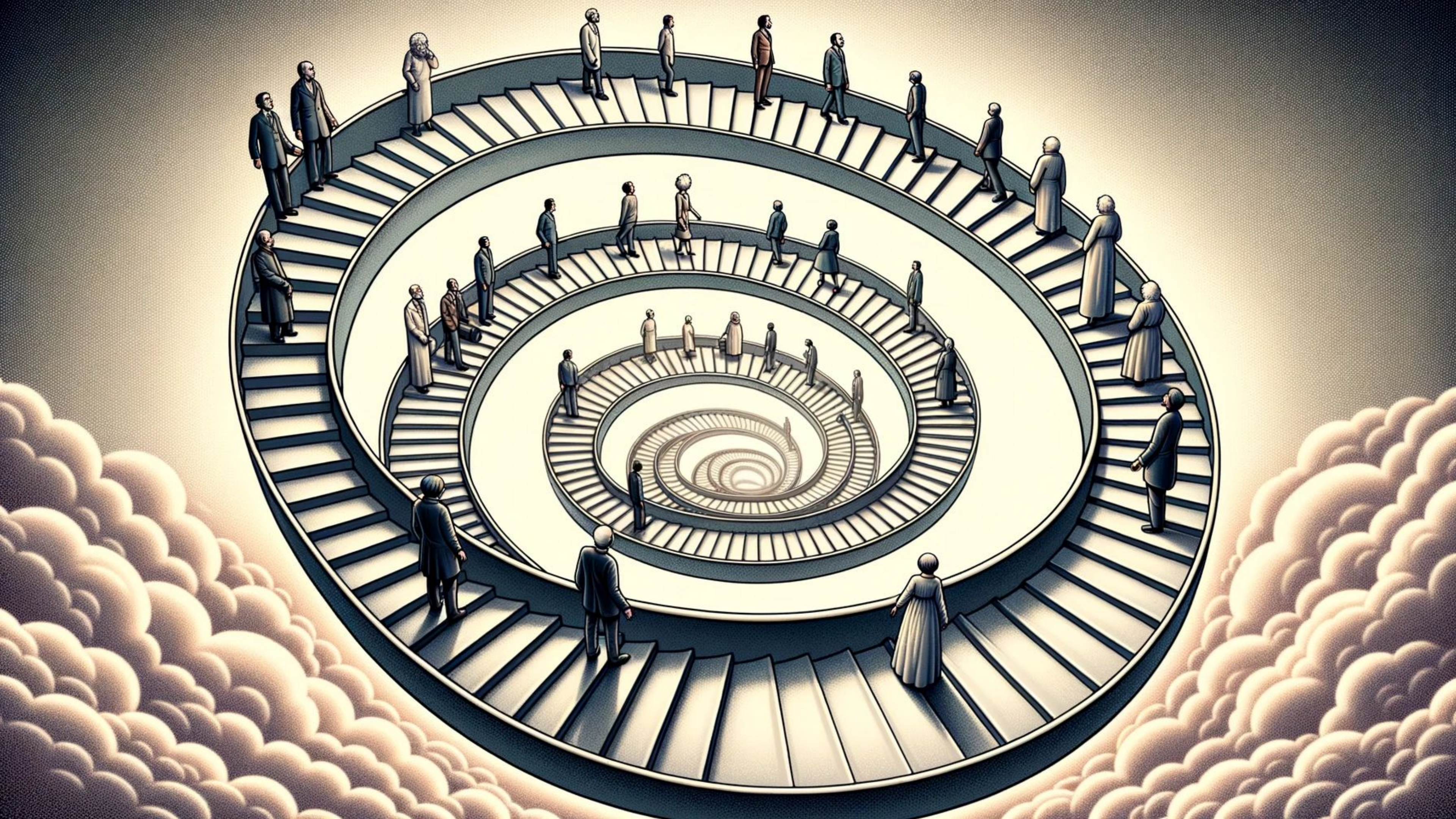
- I'm thirsty
- Bring me two different sodas
- I left out a coke, apple, and water, can you throw them away and then bring me a sponge to wipe the table?
- 你能给我拿杯可乐吗？

Human Input: How would you bring me a fruit that is not an apple?
Model Output:
Explanation: The user has requested me to bring a fruit that is not an apple. The banana is a fruit that is not an apple, I will bring the user a banana.
Robot: I would
1. find a banana
2. pick up the banana
3. go to the table
4. put down the banana
5. done.

AI Lessons

5 lessons:

1. Success is guaranteed
2. The great AI consolidation
3. The power of community
4. Using all the data
5. Bitter lesson (v2)



Thank you!

robotics-transformer1.github.io
robotics-transformer2.github.io
palm-e.github.io
robotics-transformer-x.github.io
qtransformer.github.io