

PR-454: 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, Jaspair 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,
and Brianna Zitkovich

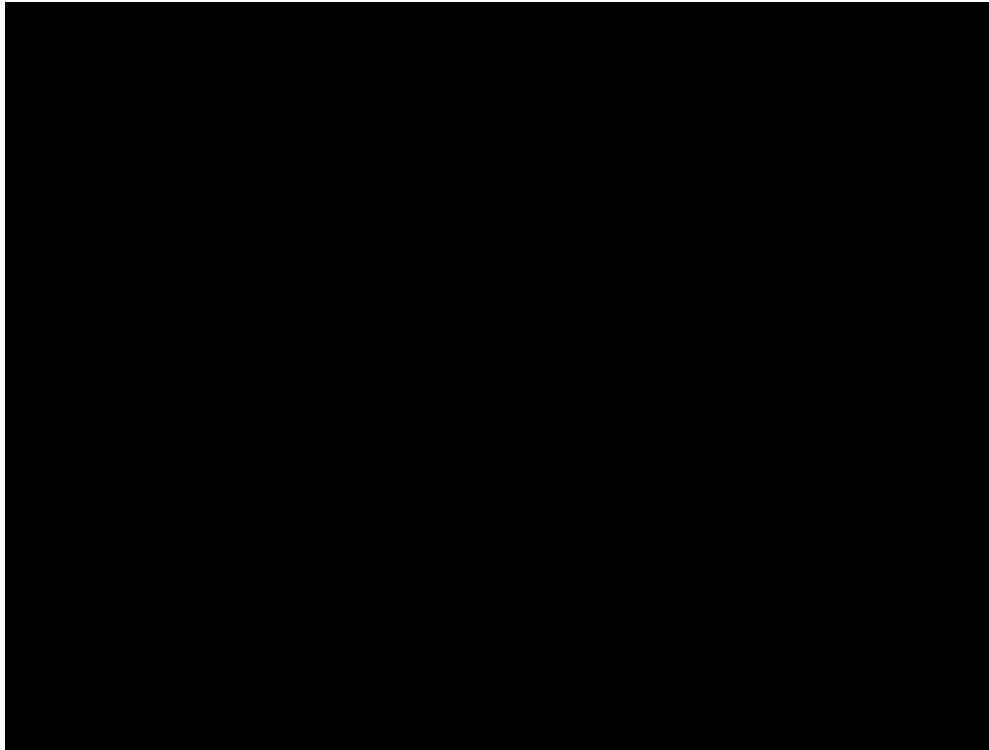
Google DeepMind. Authors listed in alphabetical order, with contributions listed in Appendix A.

PR by Yunsung Lee

Contents

- Overview
 - Related Work
 - RT-2: Vision-Language-Action
 - Experiments
 - Generalizable?
 - Emergent capabilities?
 - Vary parameter count and design?
 - Chain-of-thought reasoning?
-

Overview



Videos: <https://robotics-transformer2.github.io/#videos>

Overview

Internet-Scale VQA + Robot Action Data



Q: What is happening in the image?

A: 311 423 170 55 244

A grey donkey walks down the street.

Q: Que puis-je faire avec ces objets?

A: 3455 1144 189 25673



Faire cuire un gâteau.



Q: What should the robot do to <task>?

A: 132 114 128 5 25 156

$\Delta T = [0.1, -0.2, 0]$
 $\Delta R = [10^\circ, 25^\circ, -7^\circ]$

Vision-Language-Action Models for Robot Control

Q: What should the robot do to <task>? A: ...
↓ ↓ ↓ ↓ ↓



RT-2

ViT

Large Language Model

A: 132 114 128 5 25 156

De-Tokenize

$\Delta T = [0.1, -0.2, 0]$
 $\Delta R = [10^\circ, 25^\circ, -7^\circ]$

Robot Action

Co-Fine-Tune

Deploy

Closed-Loop Robot Control



Put the strawberry into the correct bowl



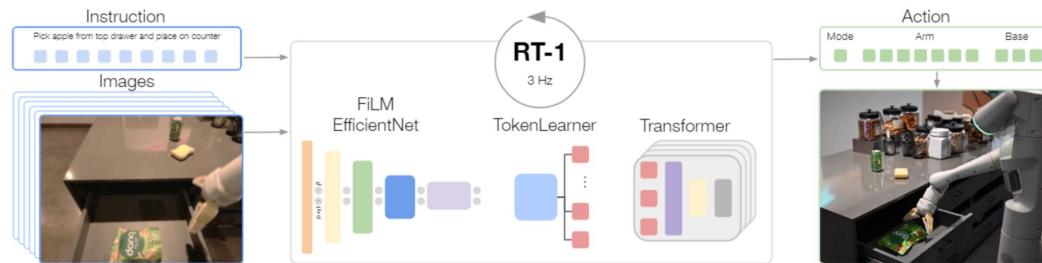
Pick the nearly falling bag



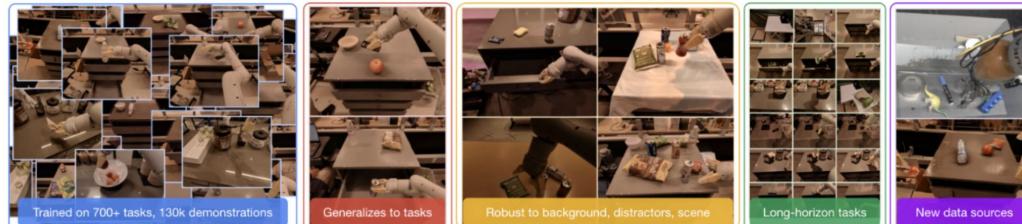
Pick object that is different

represent robot actions as another language,
which can be cast into text tokens and trained together with Internet-scale vision-language datasets.

Related Work - RT-1



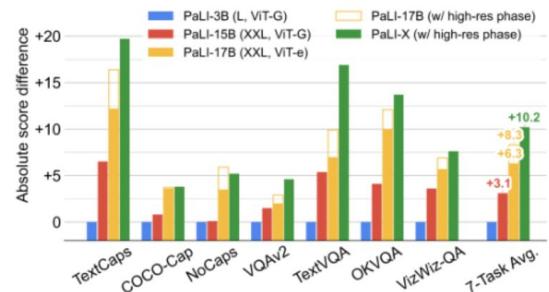
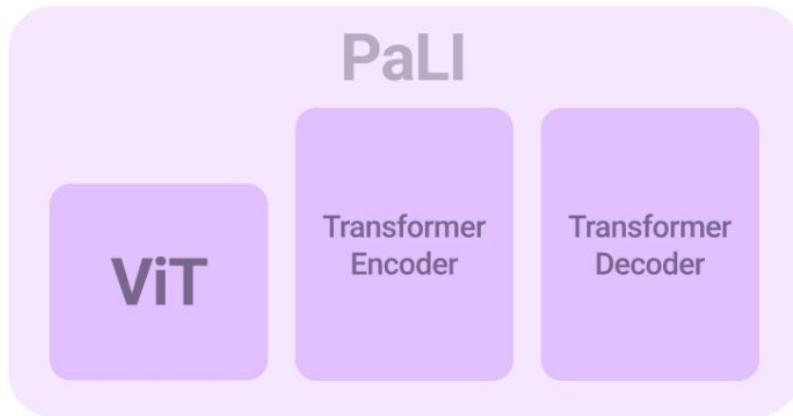
(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).



(b) RT-1's large-scale, real-world training (130k demonstrations) and evaluation (3000 real-world trials) show impressive generalization, robustness, and ability to learn from diverse data.

Figure 1: A high-level overview of RT-1's architecture, dataset, and evaluation.

Related Work - PaLI-X



PaLI-X: 55B

Related Work - PaLM-E

Mobile Manipulation



Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see . 3. Pick the green rice chip bag from the drawer and place it on the counter.

Visual Q&A, Captioning ...



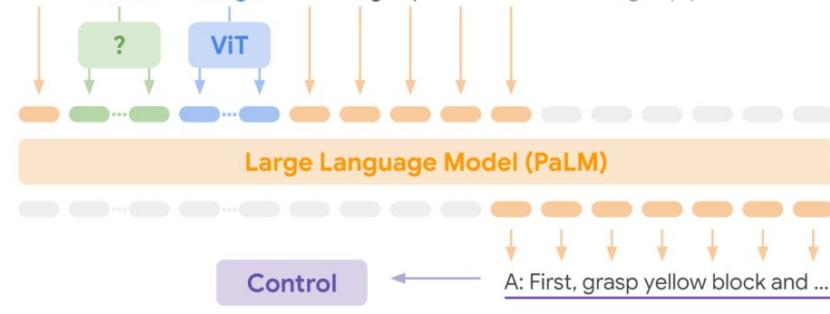
Given . Q: What's in the image? Answer in emojis.
A: 🍎🍌🍇🍐🍊🍒



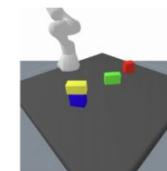
Describe the following :
A dog jumping over a hurdle at a dog show.

PaLM-E: An Embodied Multimodal Language Model

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



Task and Motion Planning



Given **<emb>** Q: How to grasp blue block?
A: First grasp yellow block and place it on the table, then grasp the blue block.

Tabletop Manipulation

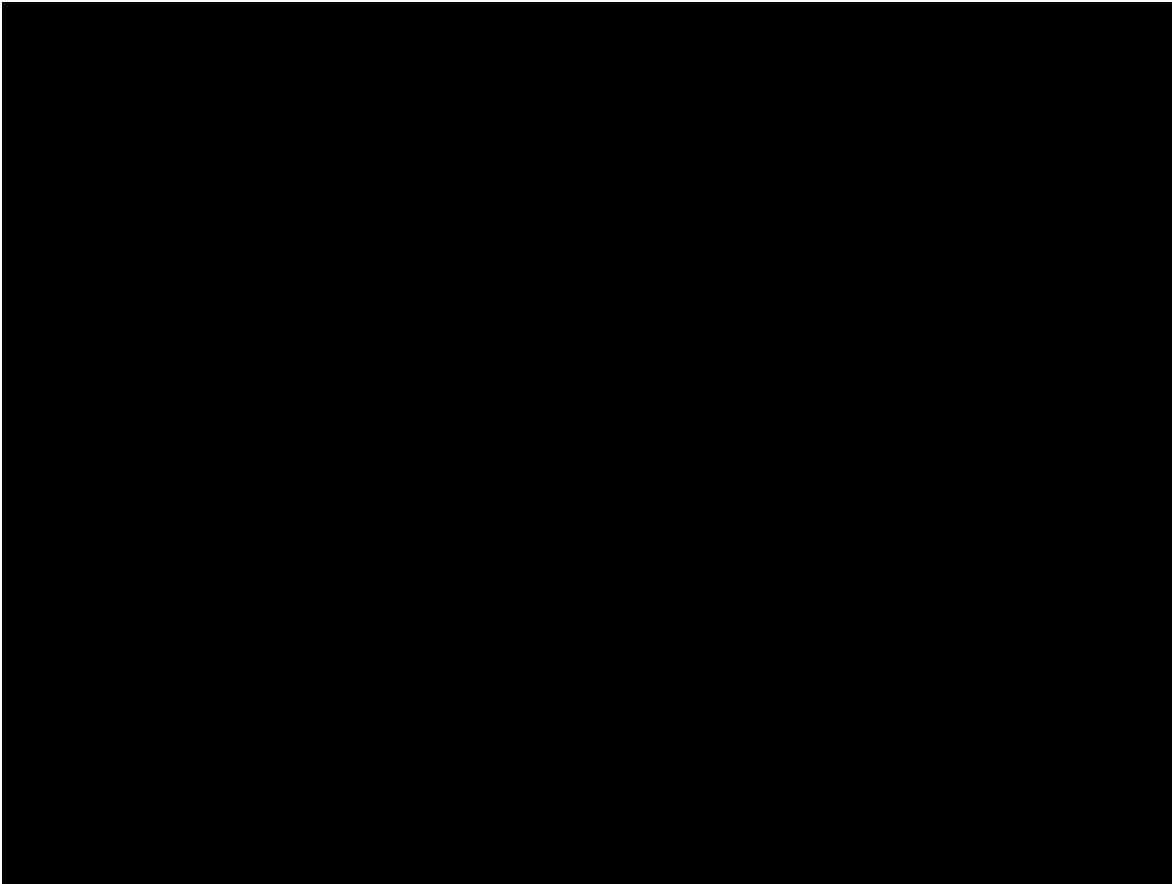


Given Task: Sort colors into corners.
Step 1. Push the green star to the bottom left.
Step 2. Push the green circle to the green star.

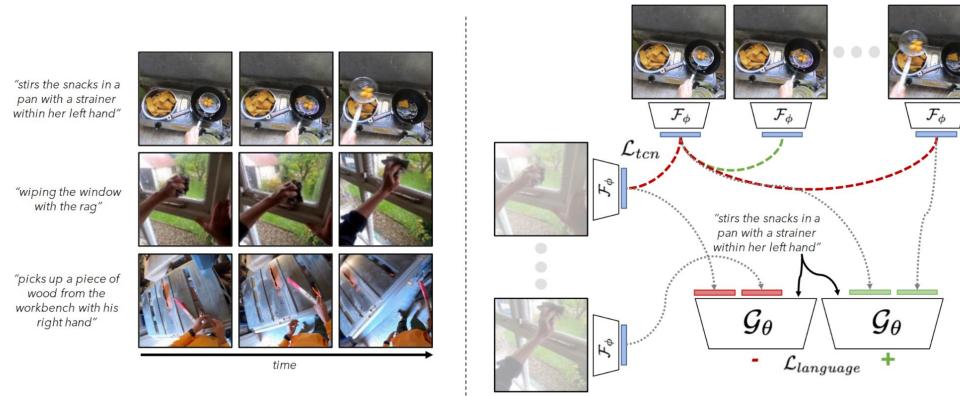
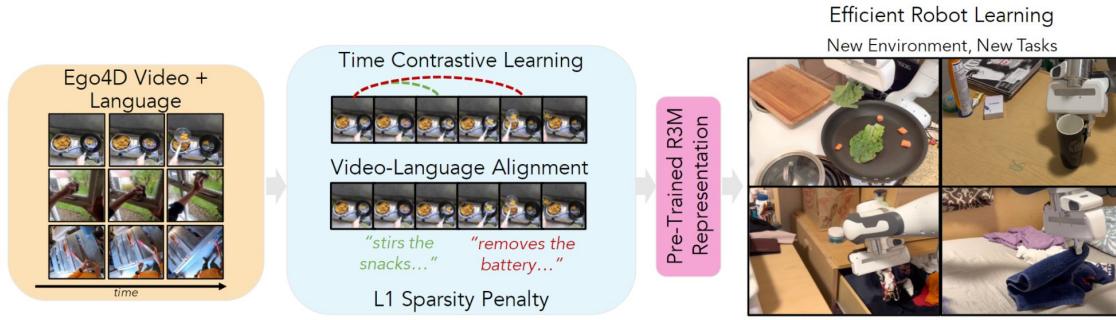
Language Only Tasks

Here is a Haiku about embodied language models:
Embodied language models are the future of natural language

Q: Miami Beach borders which ocean? A: Atlantic.
Q: What is 372×18 ? A: 6696.
Language models trained on robot sensor data can be used to guide a robot's actions.

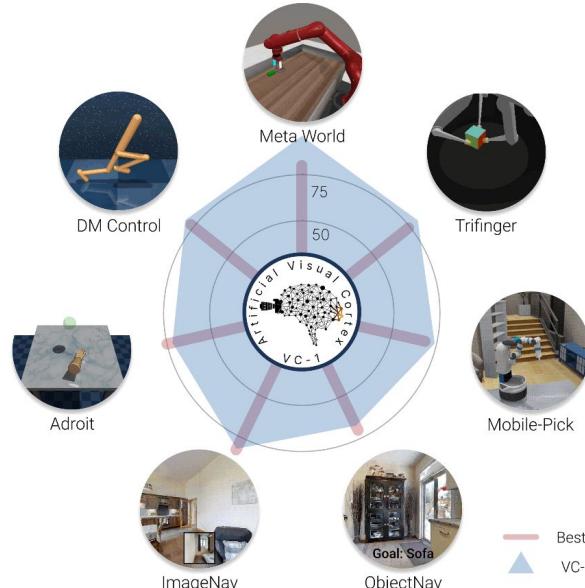


Related Work - R3M

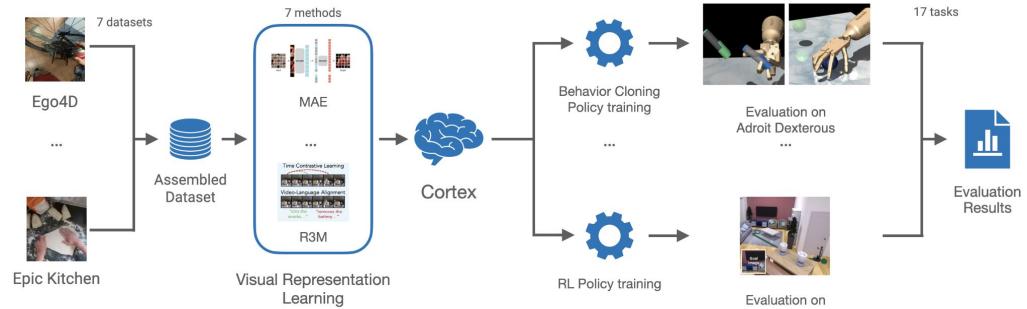


Opensource!

Related Work - VC-1



CortexBench



Opensource!

Related Work - RoboSet

Data						
Search by task name						
Activity	Task	Scene	Number of demonstrations	Example	Download	
Baking Prep	Slide-Open Drawer	Scene 1	250 Demonstrations		Download now	
Baking Prep	Slide-Open Drawer	Scene 4	250 Demonstrations		Download now	
Baking Prep	Pick Butter	Scene 1	250 Demonstrations		Download now	
Baking Prep	Pick Butter	Scene 4	250 Demonstrations		Download now	
Baking Prep	Place Butter	Scene 1	250 Demonstrations		Download now	

a large-scale real-world multi-task dataset collected across a range of everyday household activities in kitchen scenes

RT-2: Vision-Language-Action Models

Internet-Scale VQA + Robot Action Data



Q: What is happening in the image?
A: 311 423 170 55 244

A grey donkey walks down the street.

Q: Que puis-je faire avec ces objets?

A: 3455 1144 189 25673



Q: What should the robot do to <task>?

A: 132 114 128 5 25 156

$\Delta T = [0.1, -0.2, 0]$
 $\Delta R = [10^\circ, 25^\circ, -7^\circ]$

Vision-Language-Action Models for Robot Control

Q: What should the robot do to <task>? A: ...
↓ ↓ ↓ ↓ ↓



RT-2

ViT

Large Language Model

A: 132 114 128 5 25 156

De-Tokenize

$\Delta T = [0.1, -0.2, 0]$
 $\Delta R = [10^\circ, 25^\circ, -7^\circ]$

Robot Action

Co-Fine-Tune

Deploy

Closed-Loop Robot Control



Put the strawberry into the correct bowl



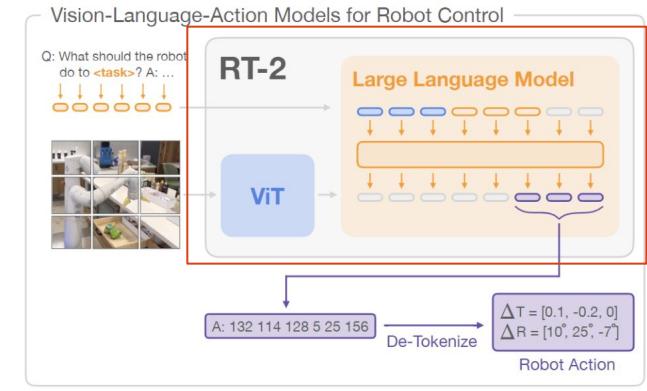
Pick the nearly falling bag



Pick object that is different

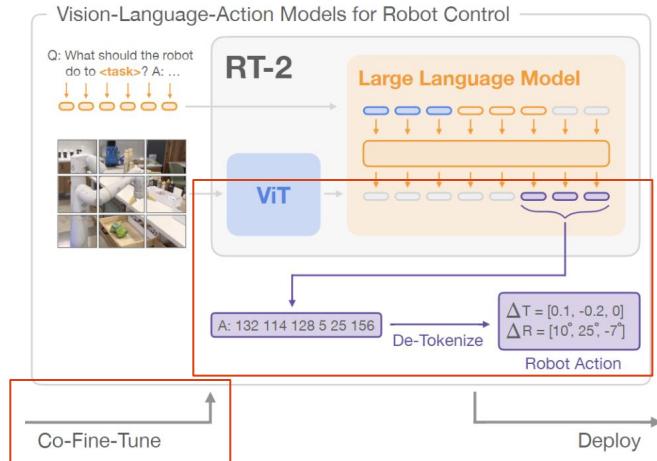
RT-2: Vision-Language Models

- Adapt two previously proposed VLMs
 - PaLI-X and PaLM-E
- Range in size from billions to tens of billions
 - PaLI-X 5B & 55B
 - PaLM-E 12B



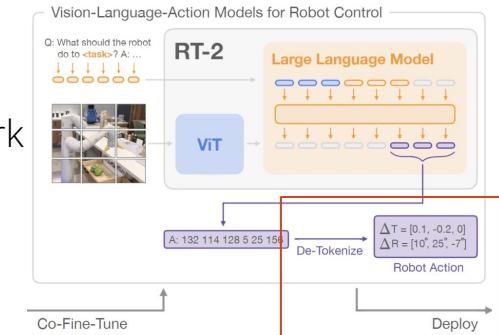
RT-2: Robot-Action Fine-tuning

- To enable VLMs to control a robot
 - must be trained to output actions
- Action encoding on the discretization by RT-1
 - 256 tokens to serve as action tokens
 - “terminate Δpos_x Δpos_y Δpos_z Δrot_x Δrot_y Δrot_z gripper_extension”.
 - 256 least frequently used tokens
 - form of symbol tuning (Wei et al., 2023)
- Co-Fine-Tuning
 - A key technical detail of training recipe
 - Co-fine-tuning robotics data with the original web data instead of naive finetuning on robot data only



RT-2: Inference

- Output Constraint
 - When model is prompted with a robot-action
 - constrain its output vocabulary via only sampling valid action tokens
- Real-Time Inference
 - Infeasible to directly run 55B models on on-robot GPUs
 - Multi-TPU cloud service and querying this service over the network
 - 55B model can run at a frequency of 1-3Hz
 - 5B model can run at a frequency of 5Hz



Experiments

4 Key Questions

1. How does RT-2 perform on seen tasks and more importantly, **generalize** over new objects, backgrounds, and environments?
2. Can we observe and measure any **emergent capabilities** of RT-2?
3. How does the generalization **vary with parameter** count and other **design** decisions?
4. Can RT-2 exhibit signs of **chain-of-thought reasoning** similarly to vision-language models?

Experiments - Generalizable?



(a) Unseen Objects



(b) Unseen Backgrounds



(c) Unseen Environments

Figure 3 | Example generalization scenarios used for evaluation in Figures 4 and 6b and Tables 4 and 6.

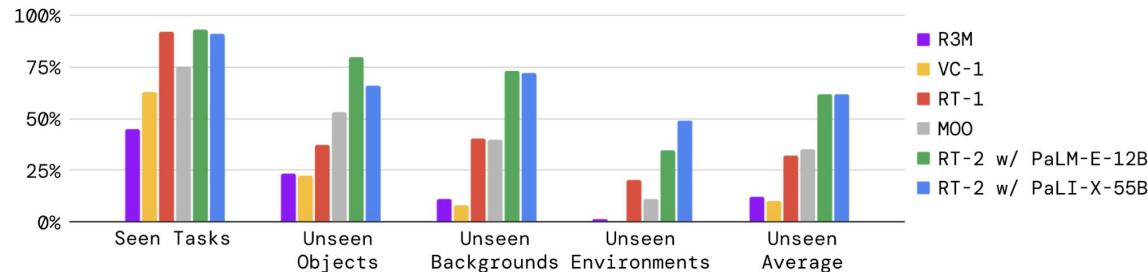
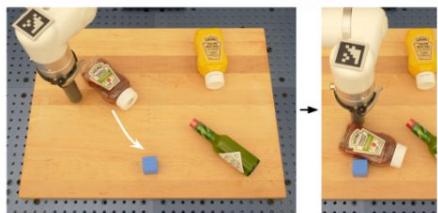


Figure 4 | Overall performance of two instantiations of RT-2 and baselines across seen training tasks as well as unseen evaluations measuring generalization to novel objects, novel backgrounds, and novel environments. Appendix Table 4 details the full results.

Experiments - Generalizable?

Push the ketchup to the blue cube



Push the blue cube to the tabasco

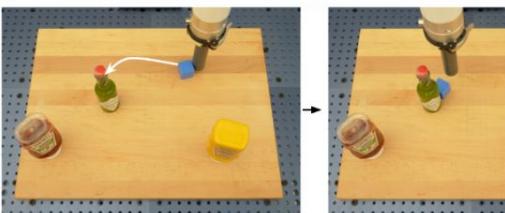


Figure 5 | Real-world out-of-distribution behaviors in the Language Table environment. Identical RT-2-PaLI-3B model checkpoint is used as in Tab. 1.

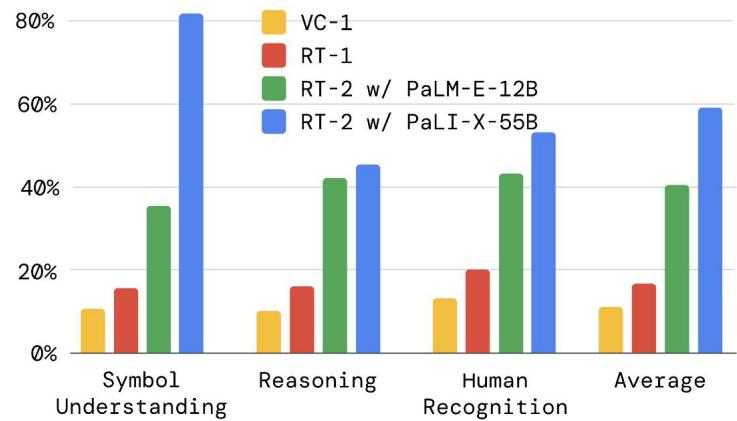
Model	Language-Table
BC-Zero (Jang et al., 2021)	72 ± 3
RT-1 (Brohan et al., 2022)	74 ± 13
LAVA (Lynch et al., 2022)	77 ± 4
RT-2-PaLI-3B (ours)	90 ± 10

Table 1 | Performance on the simulated Language-Table tasks ([Lynch and Sermanet, 2020](#)).

Experiments - Emergent capabilities?

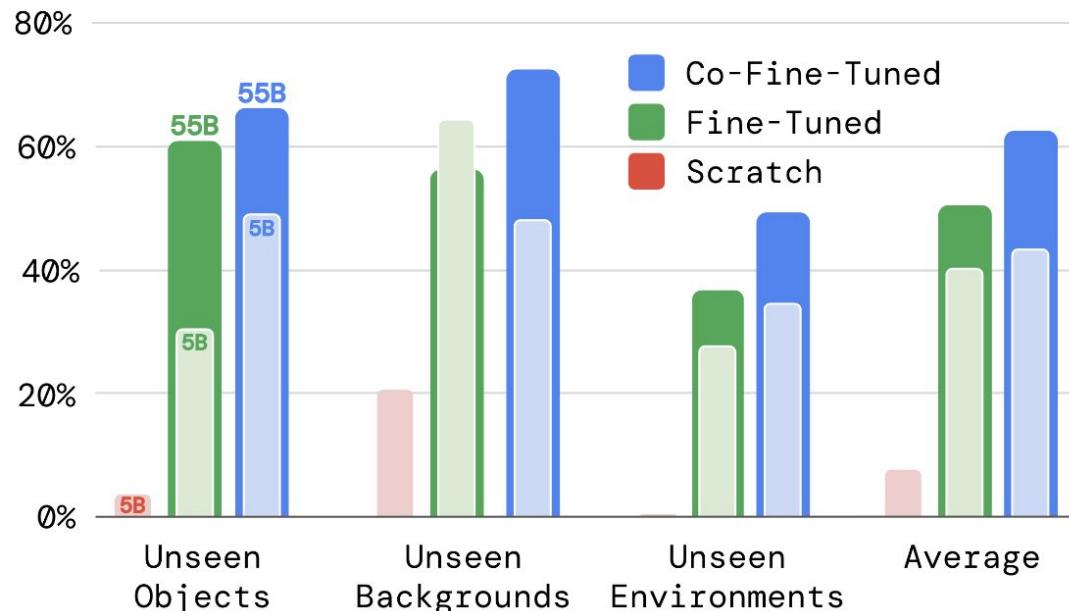


Figure 2 | RT-2 is able to generalize to a variety of real-world situations that require reasoning, symbol understanding, and human recognition. We study these challenging scenarios in detail in Section 4.



(a) Performance comparison on various emergent skill evaluations (Figure 8) between RT-2 and two baselines.

Experiments - Vary parameter count and design?



(b) Ablations of RT-2-PaLI-X showcasing the impact of parameter count and training strategy on generalization.

Experiments - Chain-of-thought reasoning?

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



Prompt:
Given 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 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 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 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

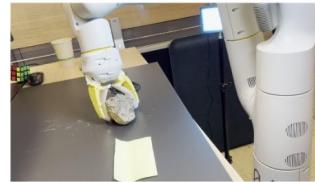


Figure 7 | Rollouts of RT-2 with chain-of-thought reasoning, where RT-2 generates both a plan and an action.

