

Robot Manipulation and Generative AI

- We're seeing **astonishing progress** in capabilities of Gen-AI models to **generate and reason about language, images, tasks, and videos**. Main contributors to this progress are:
 - Availability of **vast amounts of suitable training data** (trillions of tokens for LLMs) such that open-world reasoning becomes in-distribution
 - Very large models that can **digest this data** (100s of billions to trillions parameters)
 - Mainly **behavior cloning** for training (w/ careful data curation, RLHF for fine-tuning)
- Gen-AI doesn't readily provide **broadly applicable manipulation skills** for next gen robots
 - **Moravec's paradox**: "the hard problems are easy, and the easy problems are hard." [Pinker-94]
 - **Data**: we don't have the vast amounts of demonstration data needed to train a RobotGPT model
- **Hypothesis**: If we can generate **very large data sets demonstrating robot tasks**, then Gen-AI models with BC can **greatly elevate robot manipulation capabilities**
- **Question**: Where do we get sufficient high-quality data that covers the vast space of manipulation tasks?

1


1

Training Data for Behavior Cloning

Videos	Real World	Simulation
<p>Youtube Howto Videos</p>  <p>FoodID, B3M, MVP, Volttron [Meta AI, Stanford, Berkeley]</p> 	<p>ArmFarm, RT1, RT2, [Google, Deepmind]</p>  <p>Open-X Embodiment / DROID [21/50 institutions]</p> 	<p>ManiSkill, Orbit, Isaac Sim [UCSD, NVIDIA]</p>  <p>RoboTurk, MimicGen, [NVIDIA, Stanford]</p> 
<ul style="list-style-type: none"> • Provides strong, robust priors for visual data • Large gap between human and robot hands • Not accurate enough to provide fine-grained guidance 	<ul style="list-style-type: none"> • Excellent for pre-training / behavior cloning • Very significant effort, limited variability 	<ul style="list-style-type: none"> • Low-cost, scalable, reproducible • Sim2Real gap • Asset generation


2

2




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, Jaspier Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, Brianna Zitkovich



3



Robotics and Large Models


Large Models are changing the world

- Large Language Models (LLMs)
- Vision-Language Models (VLMs)

Robotics traditionally difficult

- Real-world data collection difficult to scale
- Human demonstrations is a bottleneck
- Autonomous collection requires bootstrapping
- Sim-to-real does not offer real world diversity

How do we get Large Model knowledge into Robotics?



4

Our journey

Google DeepMind

<p>"I spilled my drink, can you help?"</p> <p>LARGE LANGUAGE MODEL</p> <p>SayCan</p> <p>LLM as a planner Q-function as an affordance model Grounded planning</p>	<p>RT-1</p> <p>Scalable Transformer robot policy Many more tasks Compatible with SayCan</p>	<p>PaLM-E</p> <p>Vision Language Model (VLM) Trained on Web and embodied data Better planning than LLM-only</p>	<p>RT-2</p> <p>Unified web-scale VLM as robot policy Generalization to new tasks and situations Chain-of-thought reasoning possible</p>
---	--	--	--

5

Our journey

Google DeepMind

Let's dive into RT-2!

6

Vision-Language Models

Google DeepMind

- 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.

7

VLMs as Robot Policies

Google DeepMind

- **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.

8


Representing Actions in VLMs

“ Terminate or continue Δ Pos X Δ Pos Y Δ Pos Z Δ Rot X Δ Rot Y Δ Rot Z Gripper ”

Positional change Rotational change

- 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
 - Extra-IDs, least used language tokens
 - Human language (left, right etc.) - can't be directly executed on a robot

→ Vision-Language-Action (VLA) mode!!



9

Training data and underlying models

Internet-Scale VQA + Robot Action Data

Co-Fine-Tune

Models

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

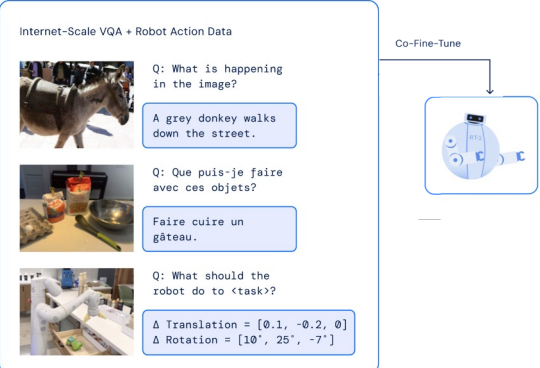
Data

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

Q: What is happening in the image?
A grey donkey walks down the street.

Q: Que puis-je faire avec ces objets?
Faire cuire un gâteau.

Q: What should the robot do to <task>?
Δ Translation = [0.1, -0.2, 0]
Δ Rotation = [10°, 25°, -7°]



10

Inference

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

RT-2

Large language model

ViT

Robot action

Δ T = [0.1, -0.2, 0]
Δ R = [10°, 25°, -7°]

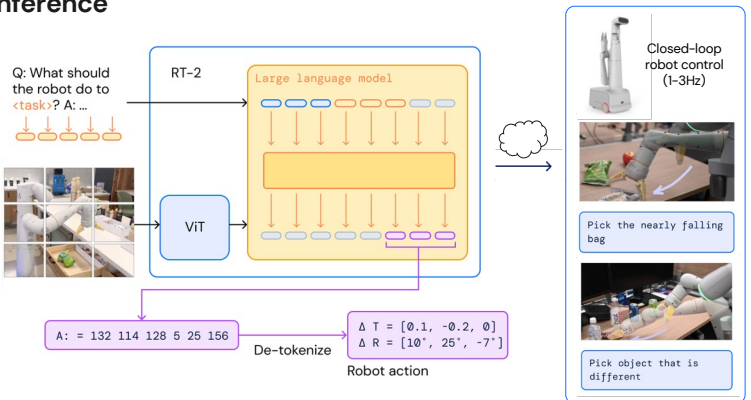
A: = 132 114 128 5 25 156

De-tokenize

Closed-loop robot control (1-3Hz)

Pick the nearly falling bag

Pick object that is different



11

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

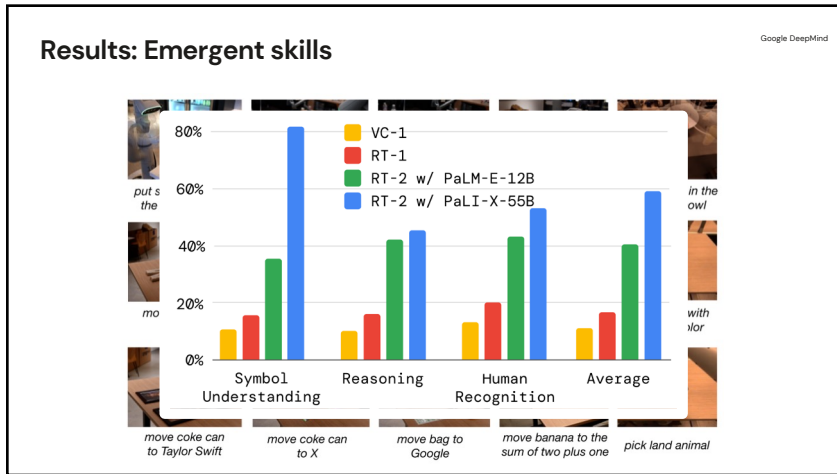
move bag to Google

move banana to the sum of two plus one

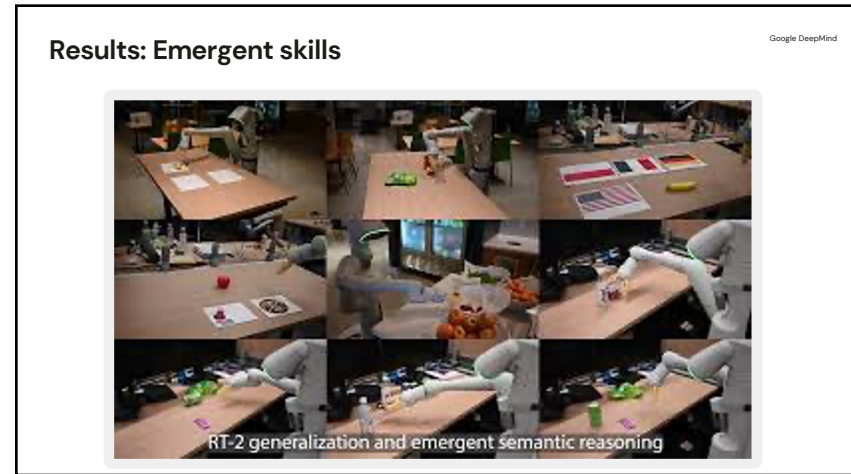
pick land animal



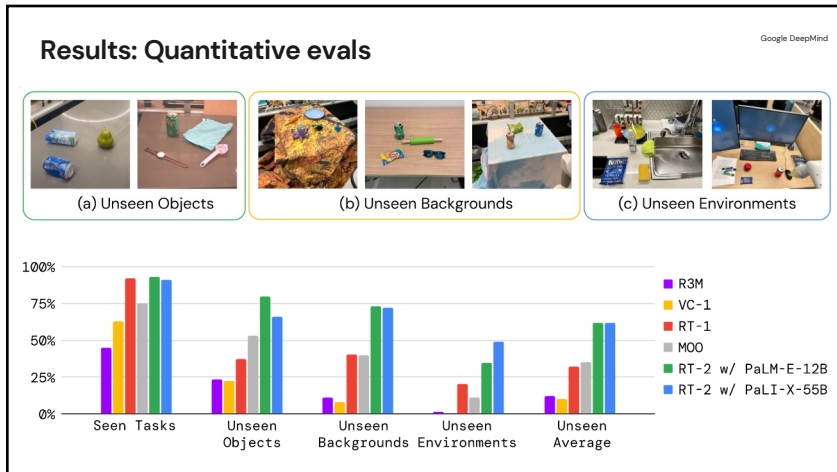
12



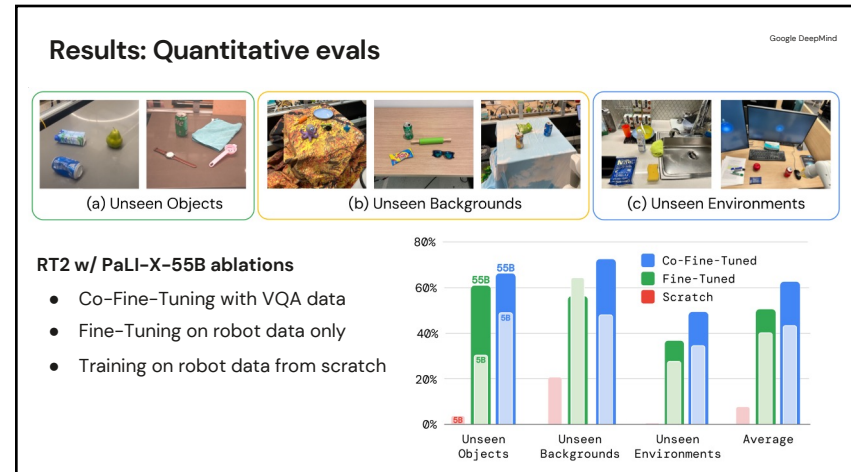
13



14



15



16

Results: Language Table

Google DeepMind

Push the ketchup to the blue cube

Push the blue cube to the tabasco

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

17

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

Google DeepMind

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

18

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

Google DeepMind

19

Conclusions

Google DeepMind

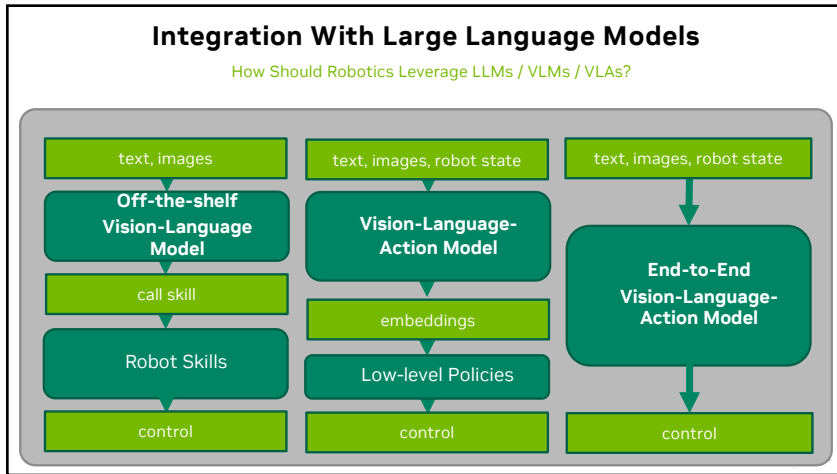
Vision-Language-Action models

- Improved generalization
- New tasks and objects
- Chain-of-Thought (CoT) reasoning
- Improving underlying VLM model can improve robot control

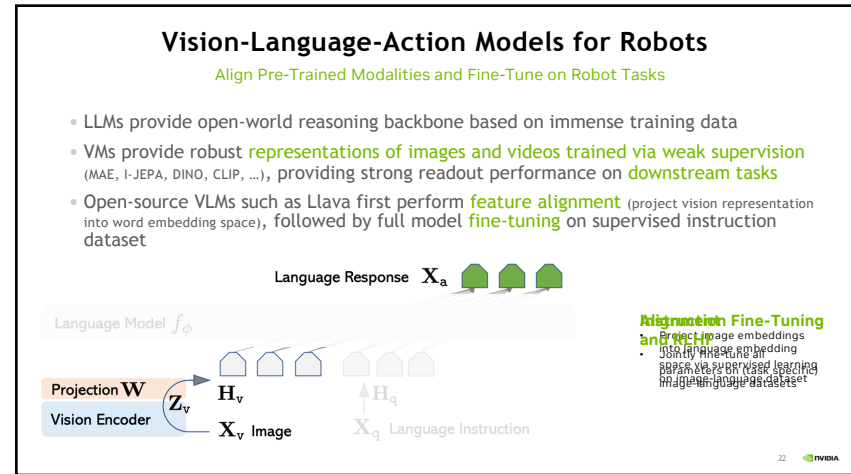
Future

- Increasing motion diversity
- Extending on CoT capabilities
- Performing RL with VLAs
- Many more!

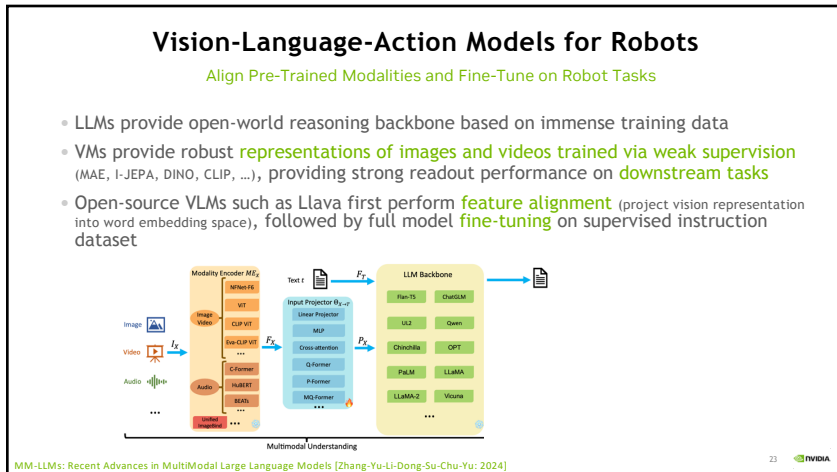
20



21



22



MM-LLMs: Recent Advances in MultiModal Large Language Models [Zhang-Yu-Li-Dong-Su-Chu-Yu: 2024]

23