

Robot Manipulation and Generative AI

- We're seeing **astounding 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? What model structures can we train on such data?

1

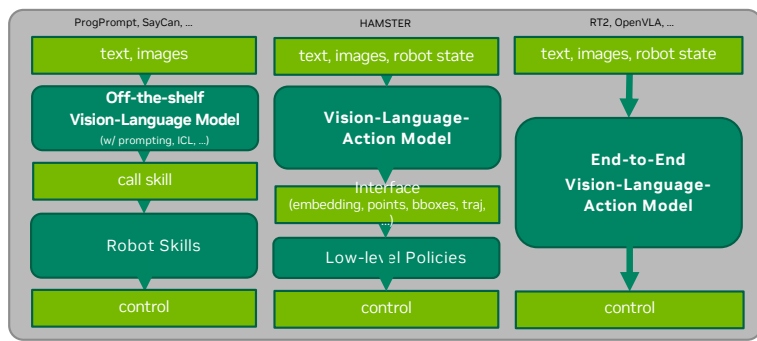
Training Data for Behavior Cloning

| Videos | Real World | Simulation |
|---|---|--|
| Youtube, HowtoVideos  | ArmFarm, RT1, RT2 [Google, Deepmind]  | ManiSkill, Orbit, Isaac Sim [UCSD, NVIDIA]  |
| Ego4D, R3M, MVP, Voltron [Meta AI, Stanford, Berkeley]  | Open-X Embodiment / DROID [21/50 institutions]  | RoboTurk, MimicGen [NVIDIA, Stanford]  |
| <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

Large Language Models

How Should Robotics Leverage LLMs / VLMs / VLAs?

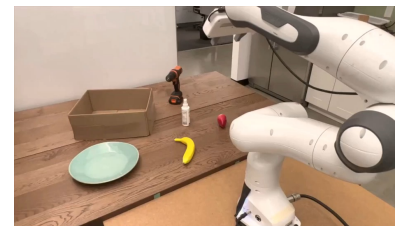


3

ProgPrompt: Leveraging LLMs for Reasoning

Generating Grounded Plans for Manipulation

Sort the fruits on the plate and the bottles in the box



[Singh-Bhalla-Mousavian-Goyal-Xu-Tremblay-Fox-Thomason-Garg: ICRA-23]

```

from actions import grab_and_putin, switchon, switchoff, open, close, ...

def throw_away_banana():
    objects = ['banana', 'garbage can', ...]
    # 1: put banana in garbage can
    grab_and_putin('banana', 'garbagecan')
    ...

def sort_fruits_on_plate_and_bottles_in_box():
    objects = ['banana', 'bottle', 'box', 'plate', 'table', 'drill', 'strawberry']
    ...

# 1: put banana on plate
grab_and_putin('banana', 'plate')
# 2: put strawberry on plate
grab_and_putin('strawberry', 'plate')
# 3: put bottle in box
grab_and_putin('bottle', 'box')
# 4: Done
  
```

Examples

Task

Detected Objects

Generated Plan

4

Google DeepMind

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, Kanishk 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



5

Google DeepMind

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?



6

Google DeepMind

Our journey

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

7

Google DeepMind

Our journey

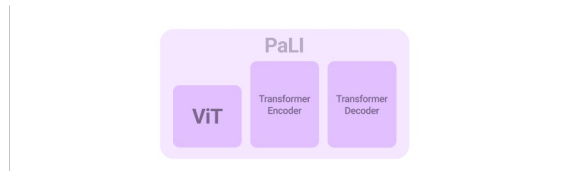


Let's dive into RT-2!

8

Vision-Language Models

Google DeepMind

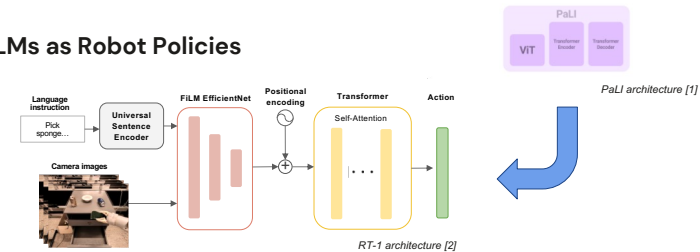


- VLMs encompass both **visual** and **semantic** understanding of the world
- In Robotics we have to deal a lot with **both** of these
- How do we leverage all of this knowledge?

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

9

VLMs as Robot Policies



- 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.

10

Representing Actions in VLMs

Google DeepMind



- **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) model!**

11

Training data and underlying models

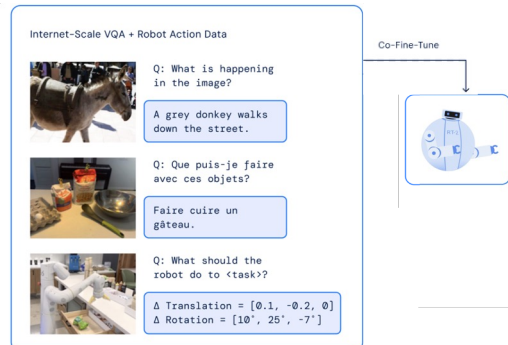
Google DeepMind

Models

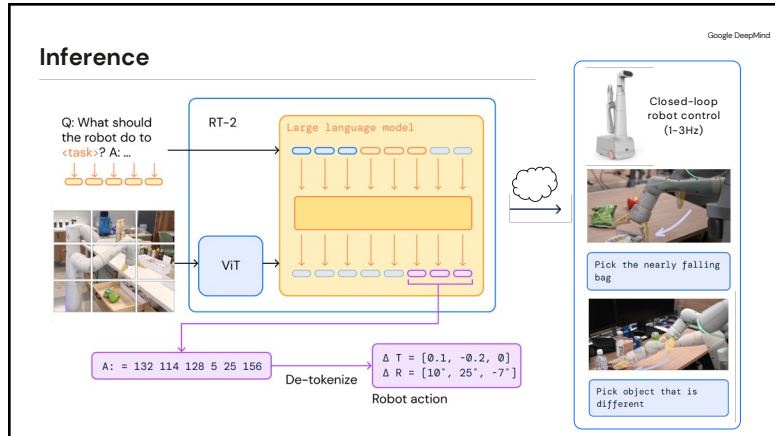
- PaLI-X (5B, 55B)
- PaLM-E (12B)

Data

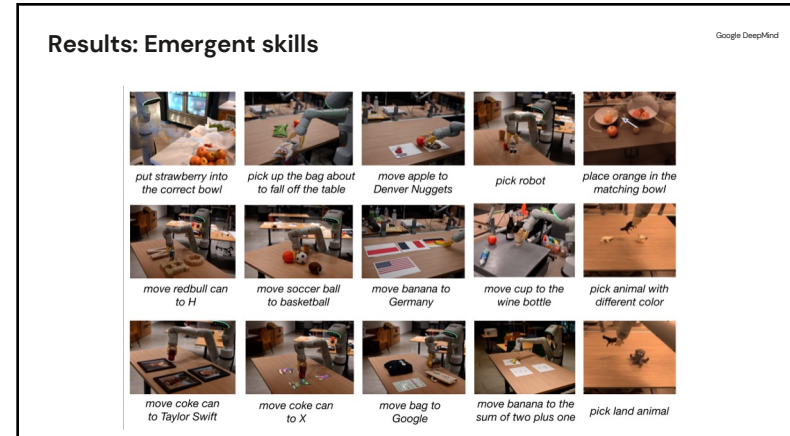
- Pretraining: Web-data
- Robot data
 - RT-1 data
 - 13 robots
 - 17 months
 - 130k demos



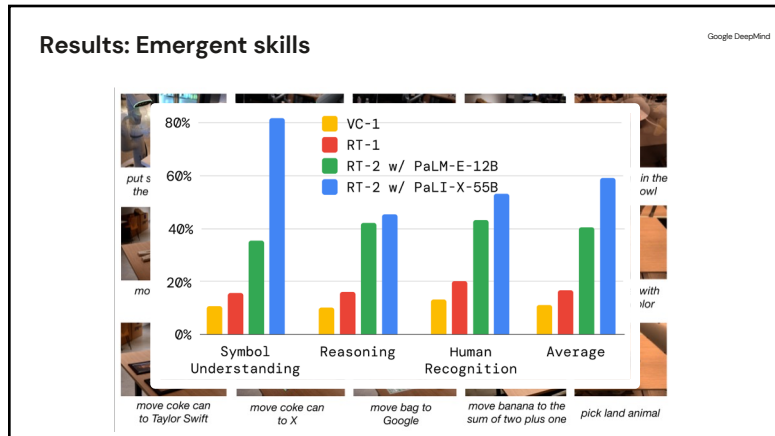
12



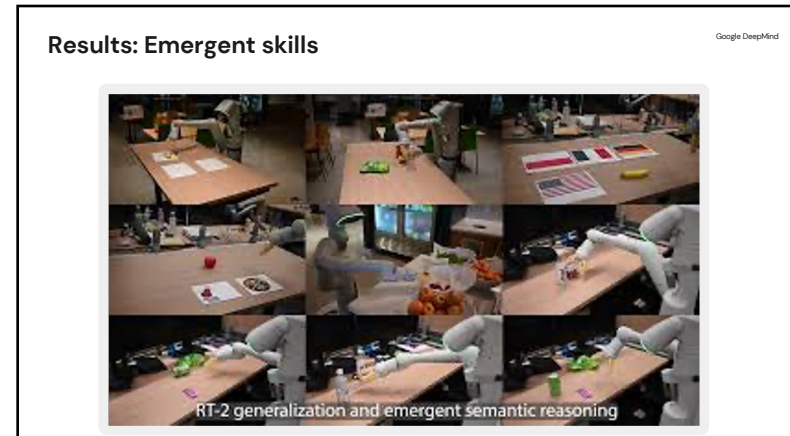
13



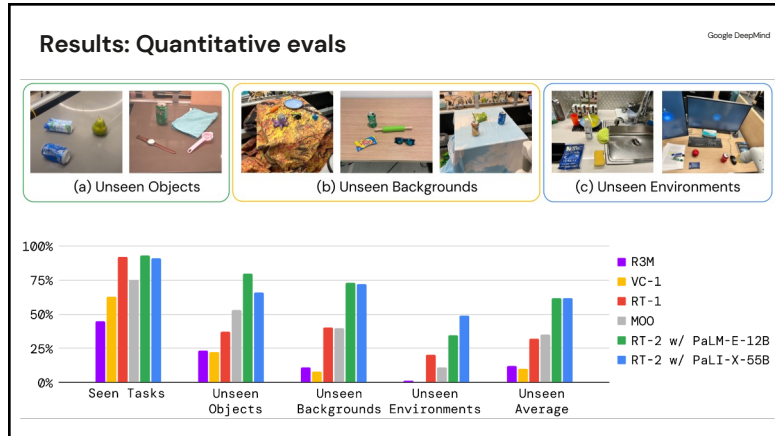
14



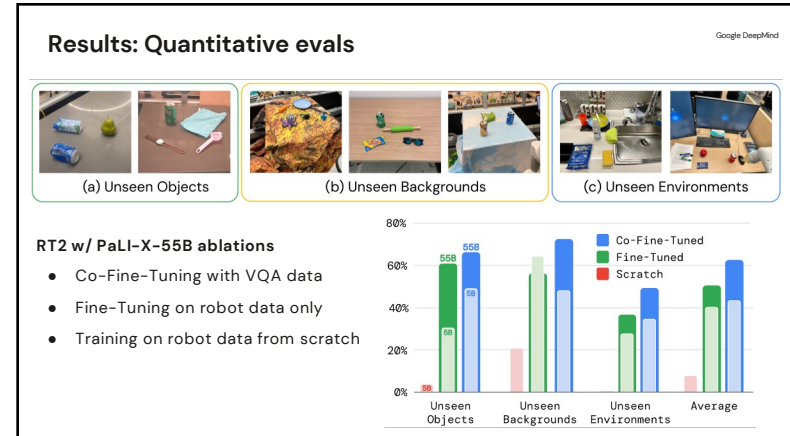
15



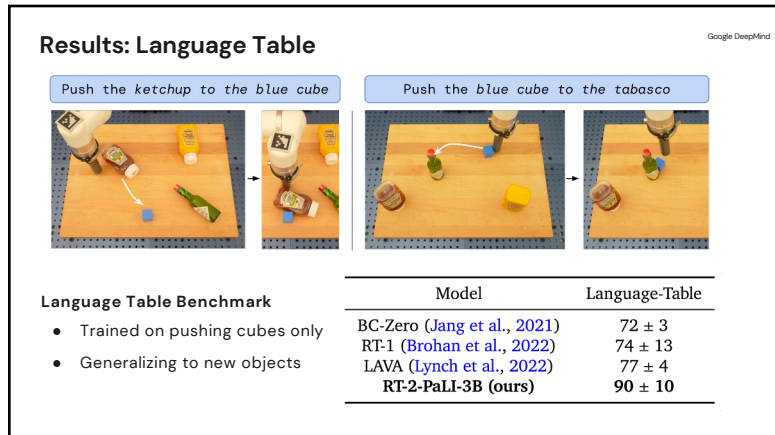
16



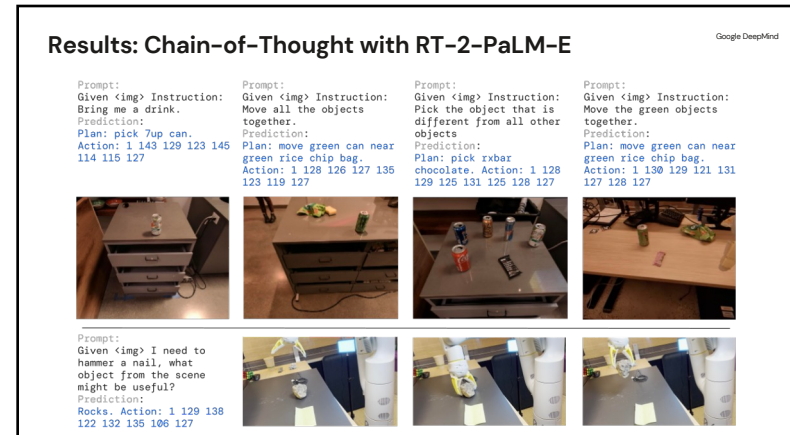
17



18



19



20

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

Google DeepMind



21

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!



22

HAMSTER Hierarchical Action Models for Open-World Robot Manipulation

HiRobot, GeminiRobotics, GR00T-N1, HAMSTER, ...

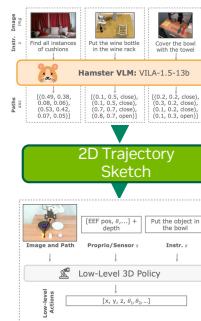
text, images, robot state

Vision-Language Model

Context information
(points, bboxes, traj, tokens, ...)

Low-level Policies

control



- VILA-1.5-13b open model
- Pre-trained on internet-scale data for open-world visual reasoning
- Fine-tuned on sim+real robot tasks

[Lin-Yin-Ping-Molchanov-Shoeybi-Han: CVPR-24]

- 3DDA or RVT-2 motion policy
- Trained on less, real robot demos

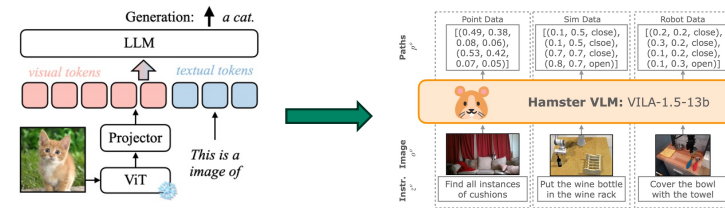
3DDA [Ke-Gkanatsios-Fragkiadaki: CoRL-24]
RVT-2 [Goyal-Blukis-Xu-Guo-Chao-Fox: CoRL-24]

23

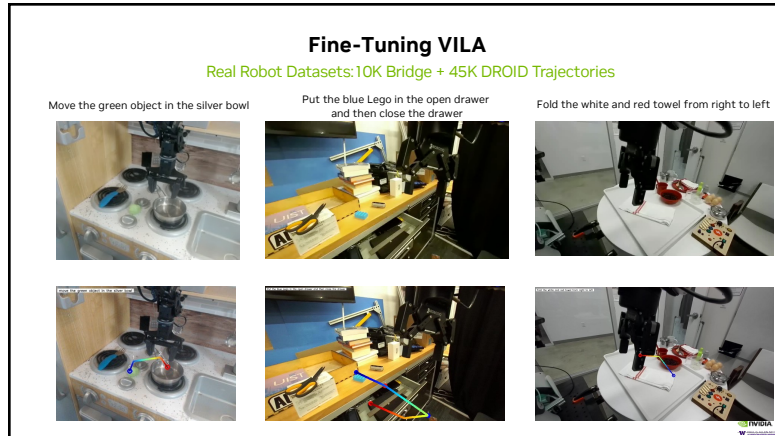
Fine-Tuning VILA

Improve VILA's Ability to Generate 2D Trajectory Output

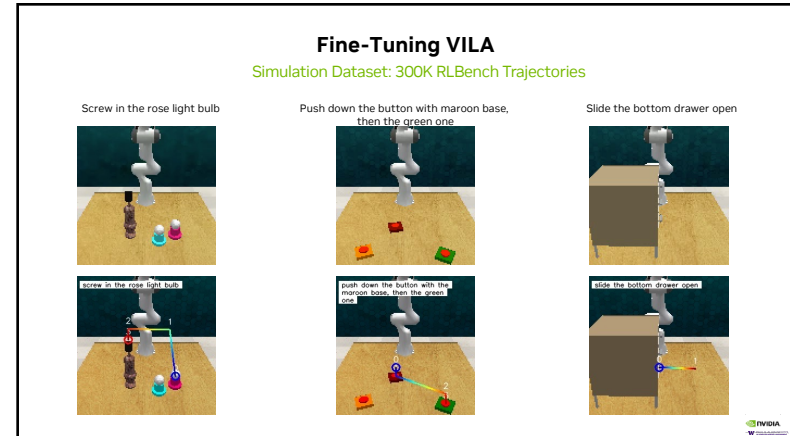
VILA [Lin-Yin-Ping-Molchanov-Shoeybi-Han: CVPR-24]



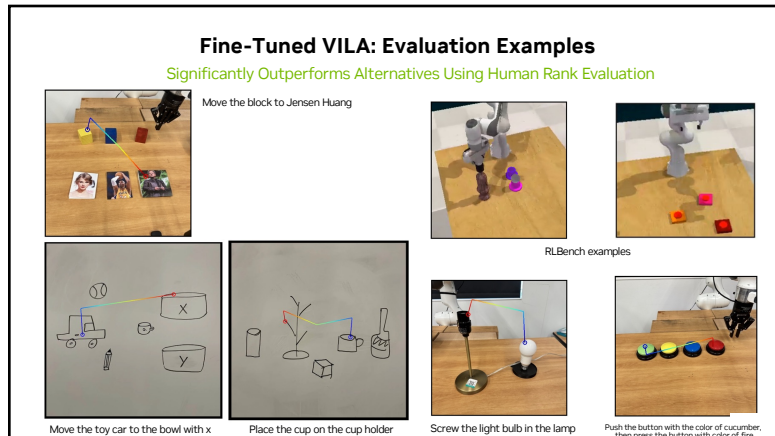
24



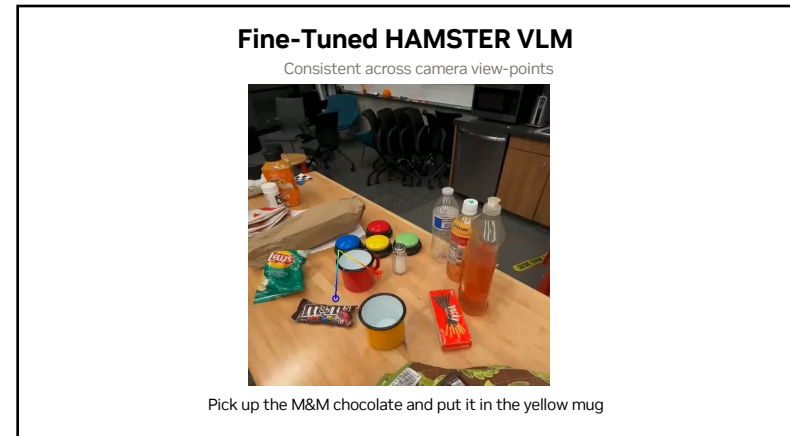
25



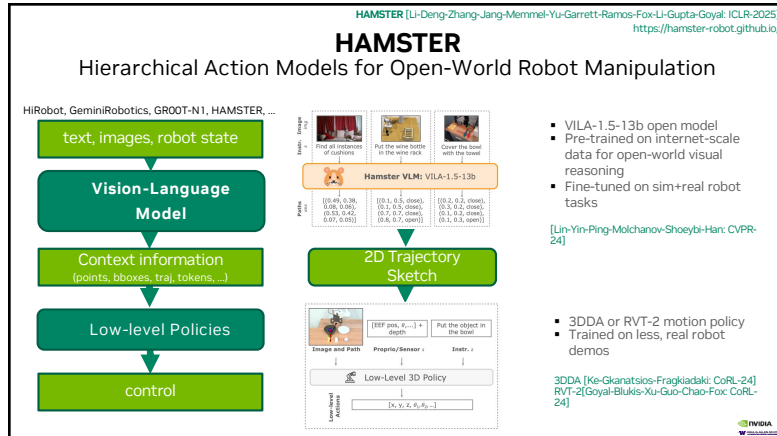
26



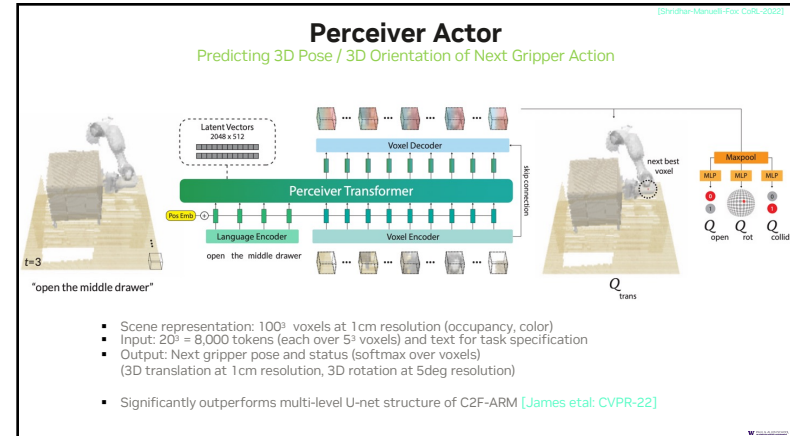
27



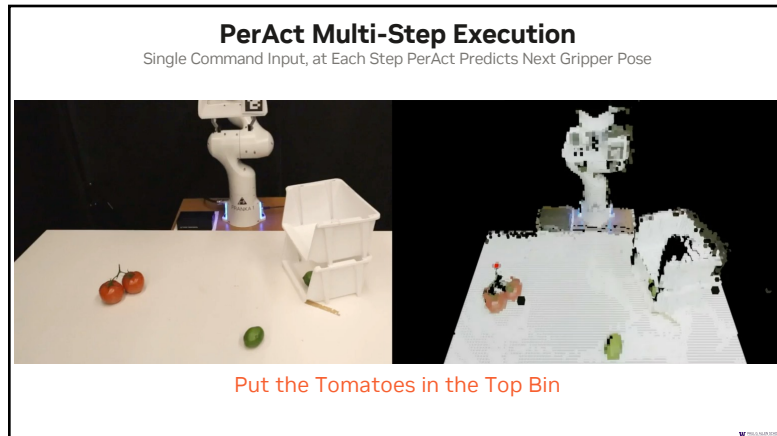
28



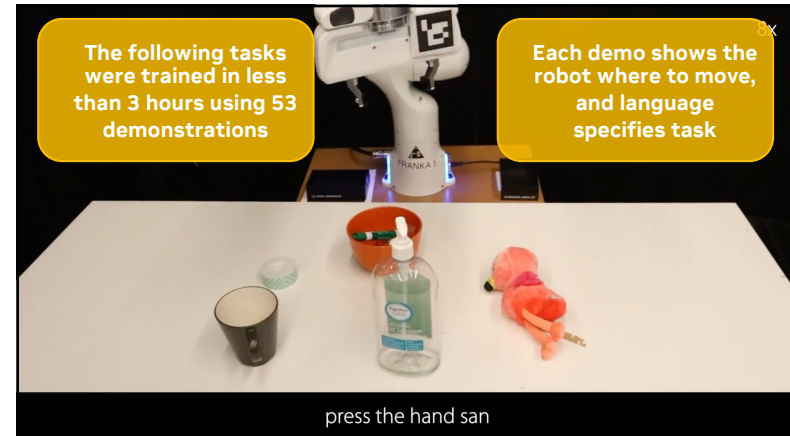
29



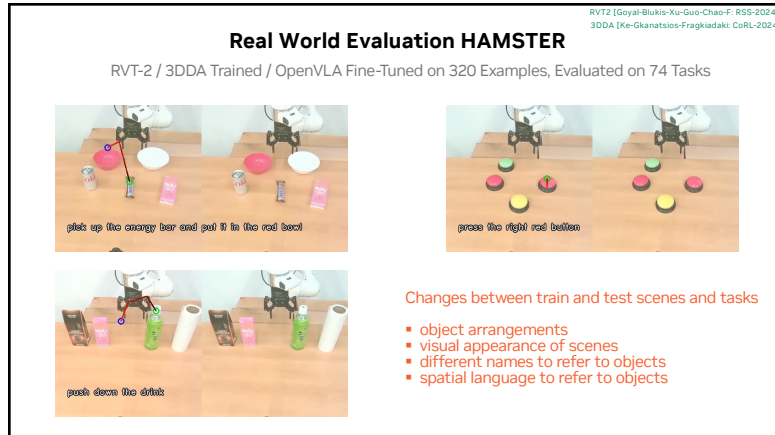
30



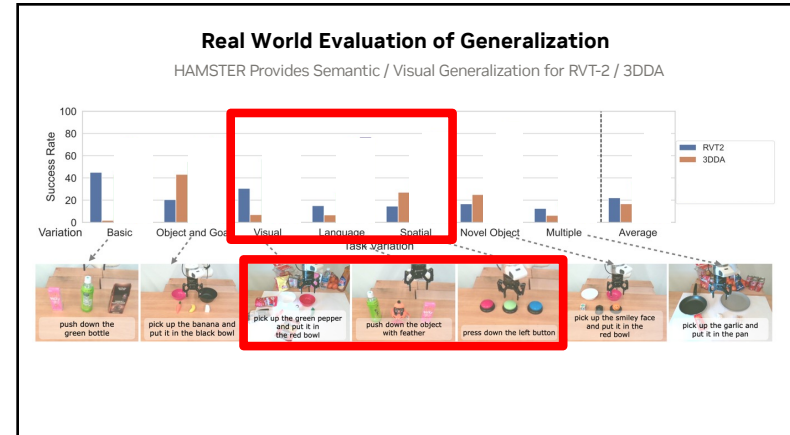
31



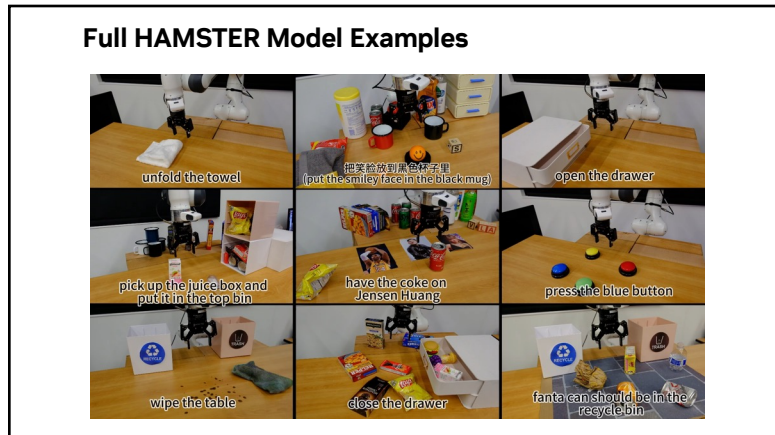
32



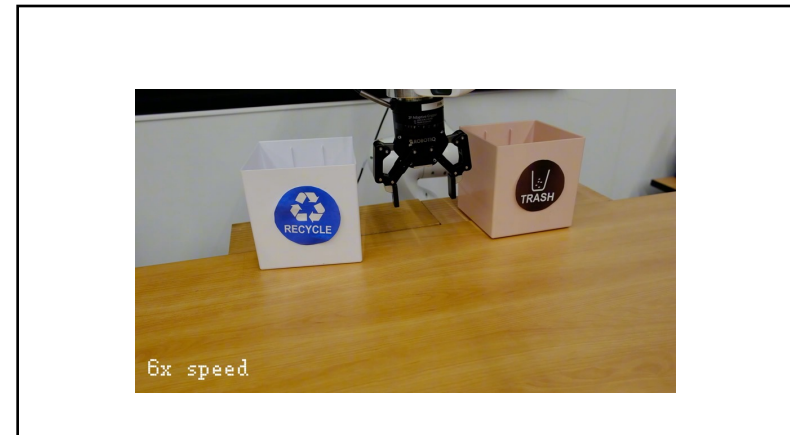
33



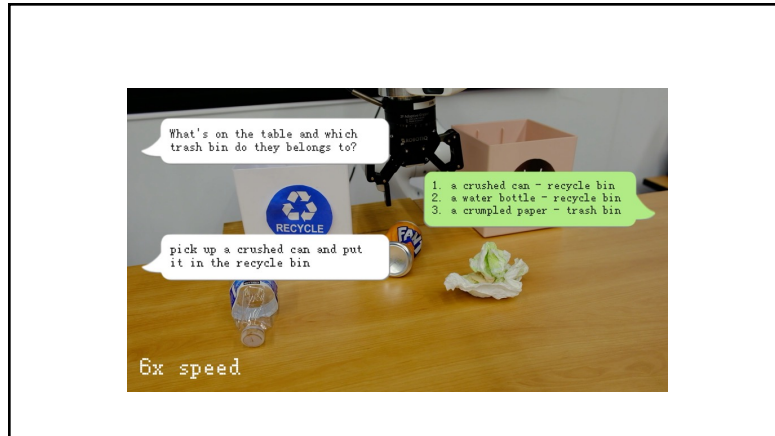
34



35



36



37

Where is RobotGPT?

- Computer vision and NLP communities have shown that **learning at scale enables powerful models** for images, videos, and language, but **RobotGPT still has a way to go**
- Simulation helps overcome data starvation**
 - generate the kind of data** necessary to train foundation manipulation capabilities (diversity and scaling via **automatic asset, scene, and task generation**; **privileged information** enables generation of demonstrations via TAMP, control, RL)
 - benchmark** models and techniques, enabling community to measure progress
 - broadening community participation** due to manageable cost
- Hierarchical action models** combine **open-world semantic reasoning of VLMs** with **reactive motion generation**
 - High level performs embodiment-agnostic semantic and spatial reasoning to generate guidance for low level policy
 - Low level policy generates reactive, 3D motion controls from relatively small demonstration data

28 NVIDIA

38