Creating Generalist Robot Models by Physical Intelligence

BuzzRobot
95.4K subscribers
Subscribe
99
Share
Ask
Save

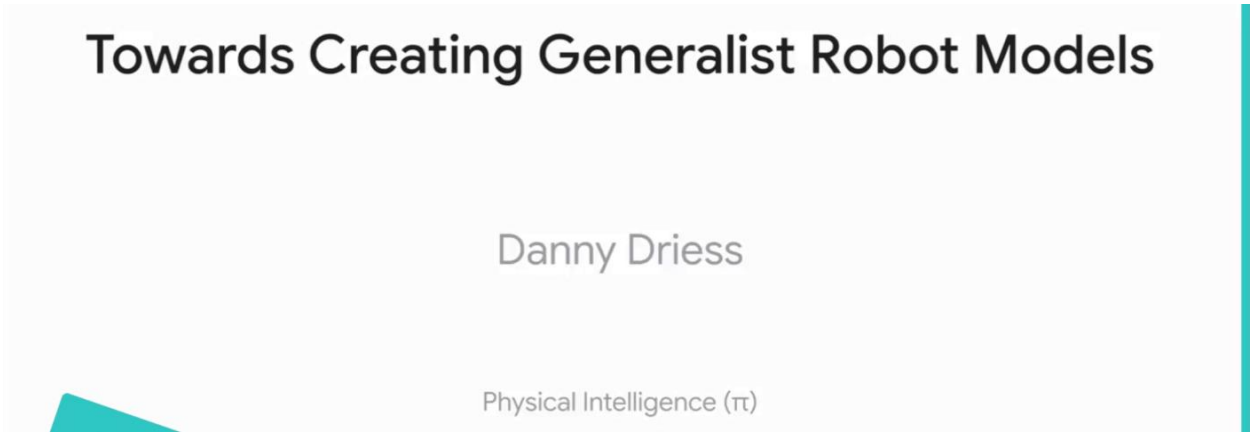
1,990 views Feb 20, 2025 #robotics #educationalvideos #robot

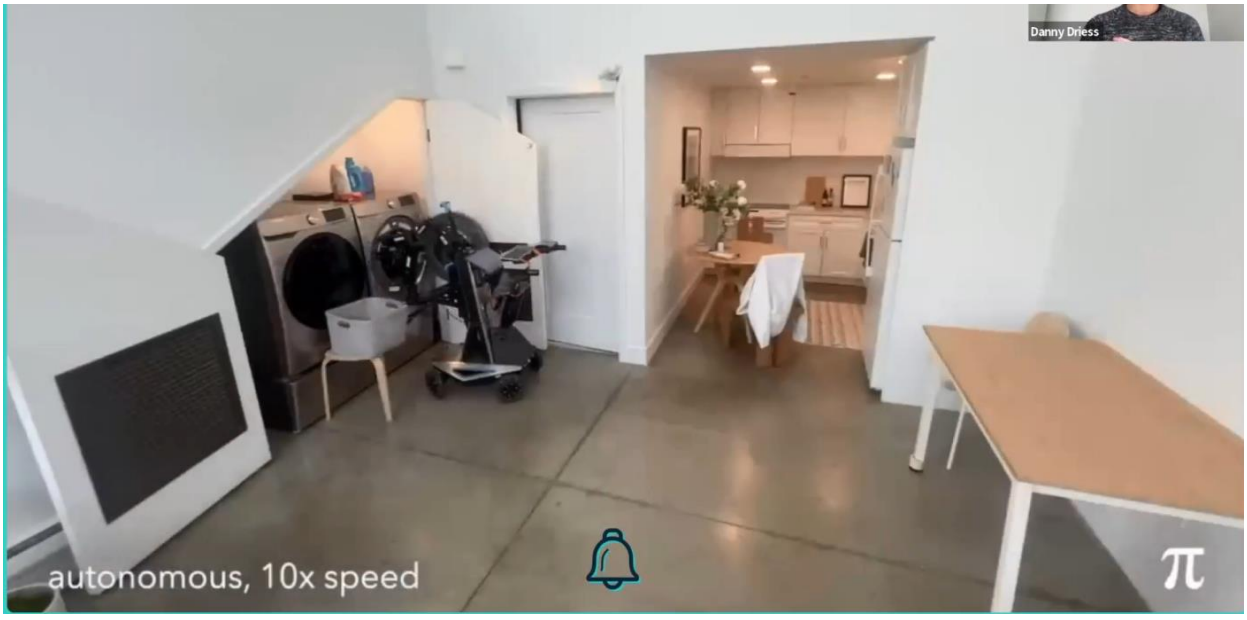
In this talk, Danny Driess, a research scientist at Physical Intelligence (Pi), explores the path to creating generalist #robot models—robot policies capable of solving any task in any environment.

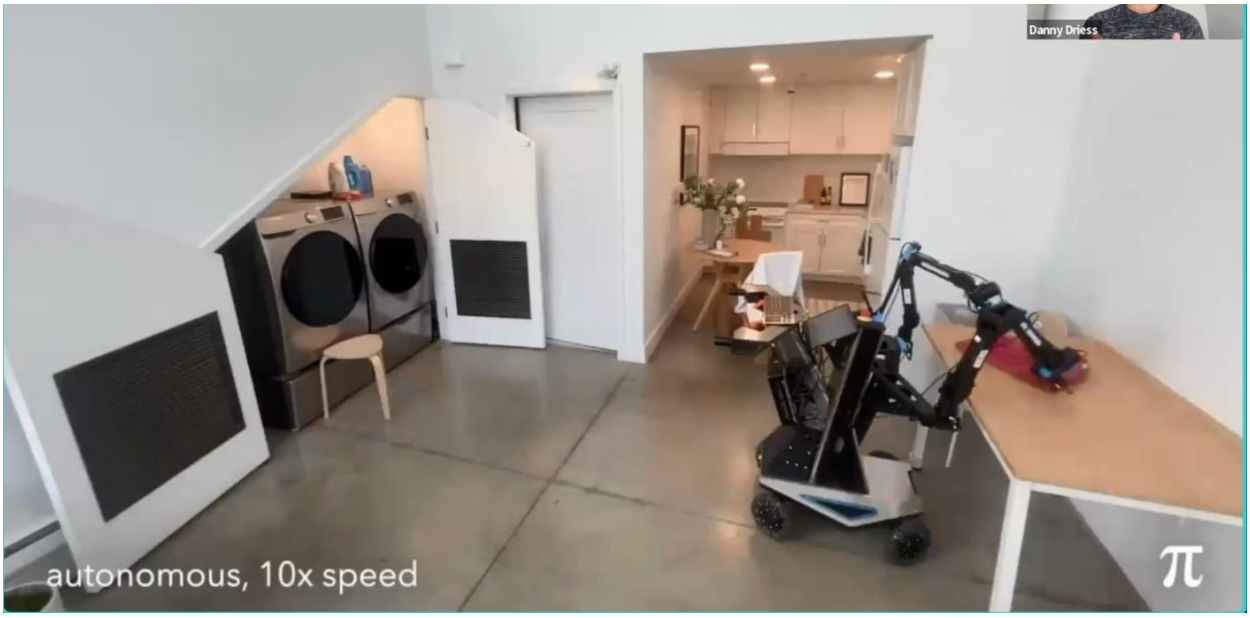
Danny introduces PaLM-E, one of the first large embodied #ai models, which demonstrated how general vision-language knowledge can transfer to robotics for high-level reasoning. He then explains how these insights contributed to the development of RT-2, Pi0, and Pi0-FAST. Notably, with Pi0, the team showcased one of the first generalist policies capable of fully autonomous, long-horizon dexterous tasks, such as unloading a dryer and folding laundry.

Bio: Danny Driess is a research scientist at Physical Intelligence (Pi). Prior to that, he was a senior research scientist at @googledeepmind, working in the intersection between robotics and #gemini.

Timestamps:



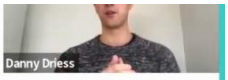




LLMs & VLMs

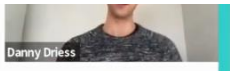
- Solve a large variety of tasks
- Work basically “all the time”
- Remarkable generalization capabilities

Robots



- Limited set of tasks
- Limited generalization capabilities
- Transfer between robot tasks unclear
- Brittle

Physical Intelligence (π)

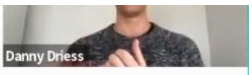


Build AI systems that operate in the real, **physical world**

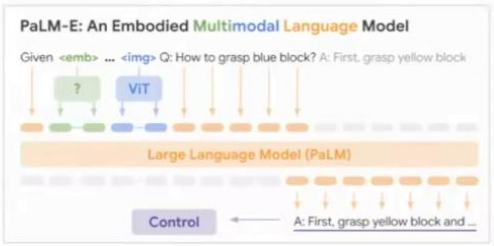
Generalist robot model that works in **every environment**

In this talk

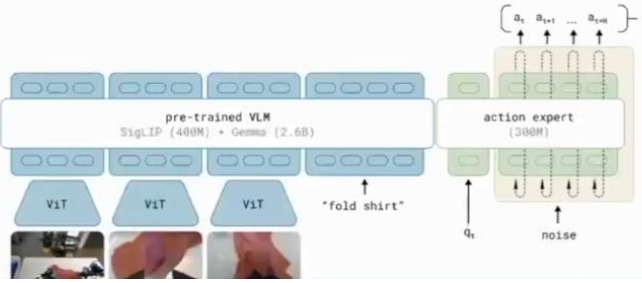
LLMs & VLMs



Embodied VLMs



Generalist Robot Models



PaLM-E: An Embodied Multimodal Language Model

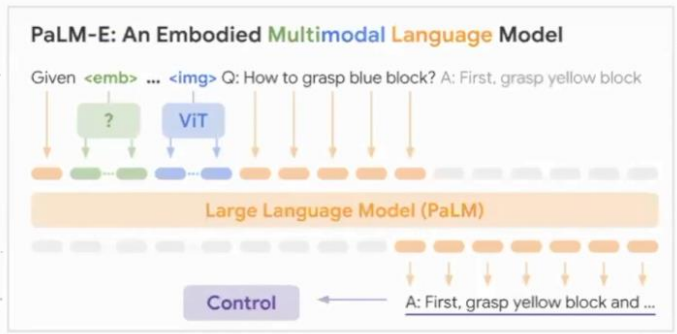
Danny Driess, Fei Xia, Mehdi Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor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

Google DeepMind

ICML 2023

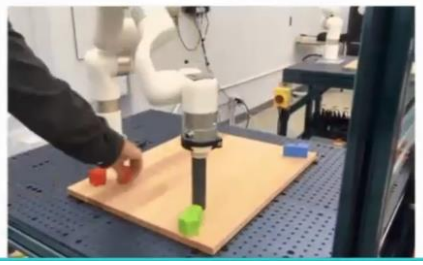
"One model"
- Embodied robotics tasks
- Vision-language
- Language
- *... across multiple robot embodiments*
- *... across multiple modalities (vision, states, neural scenes)*

positive transfer

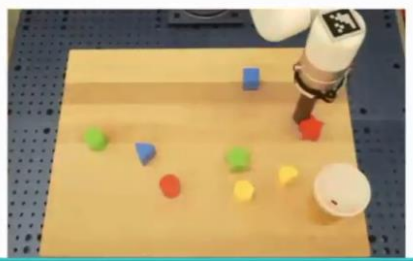


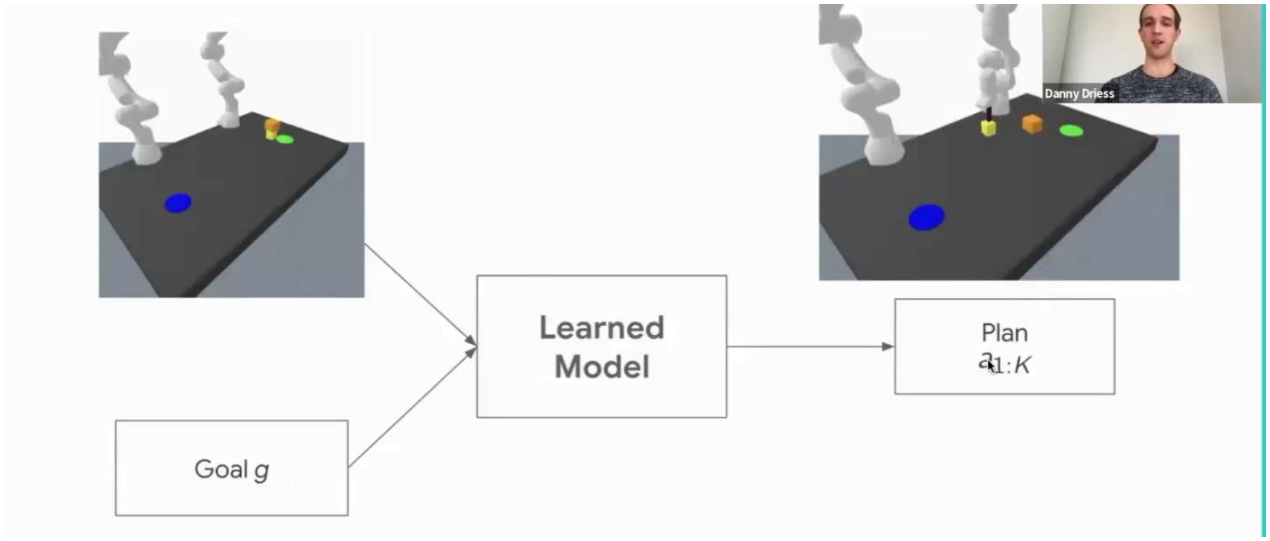
closed-loop end-to-end planning
("Given <img>... Bring me the rice chips from the drawer ")

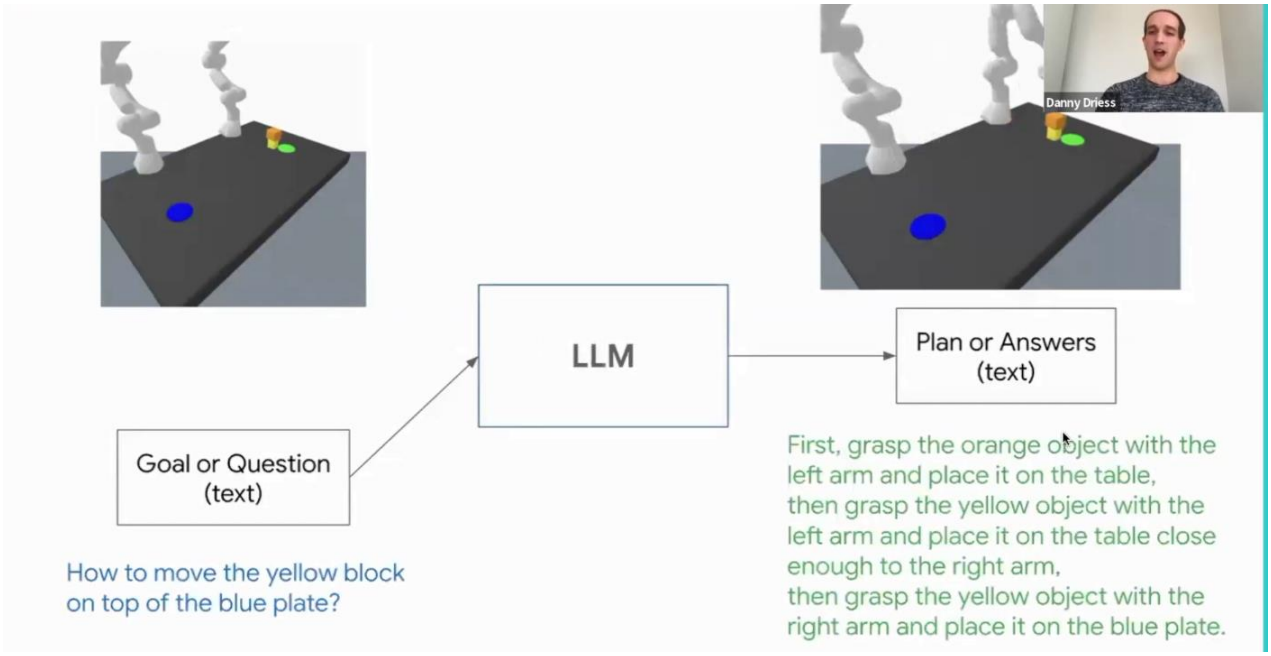
long-horizon tasks
("Given <img>.... Sort the blocks by colors into corners")

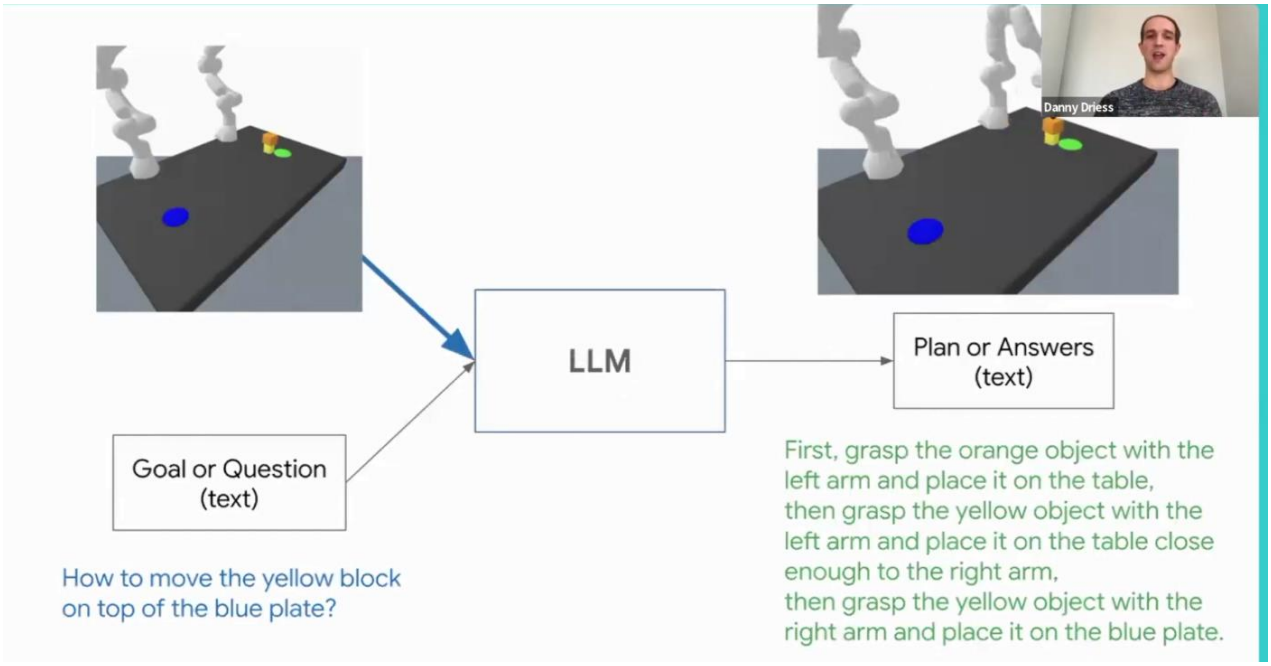


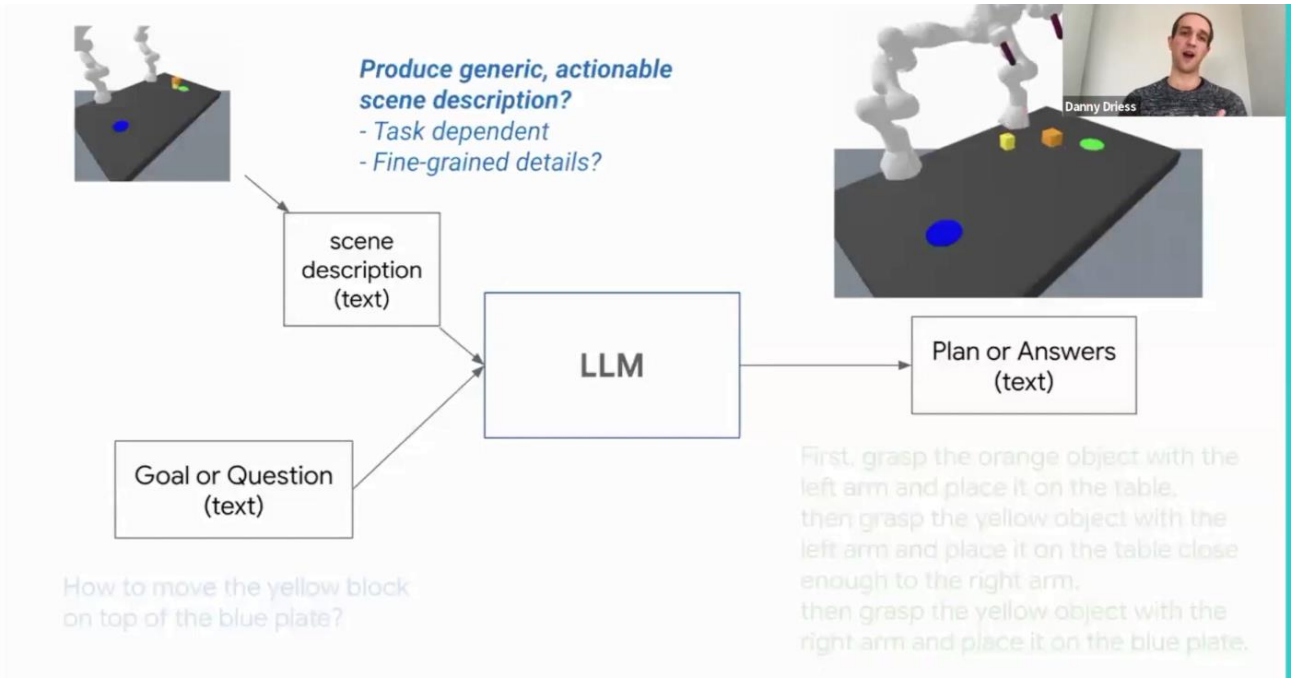
zero-shot generalization
(unseen object pairings, or objects)

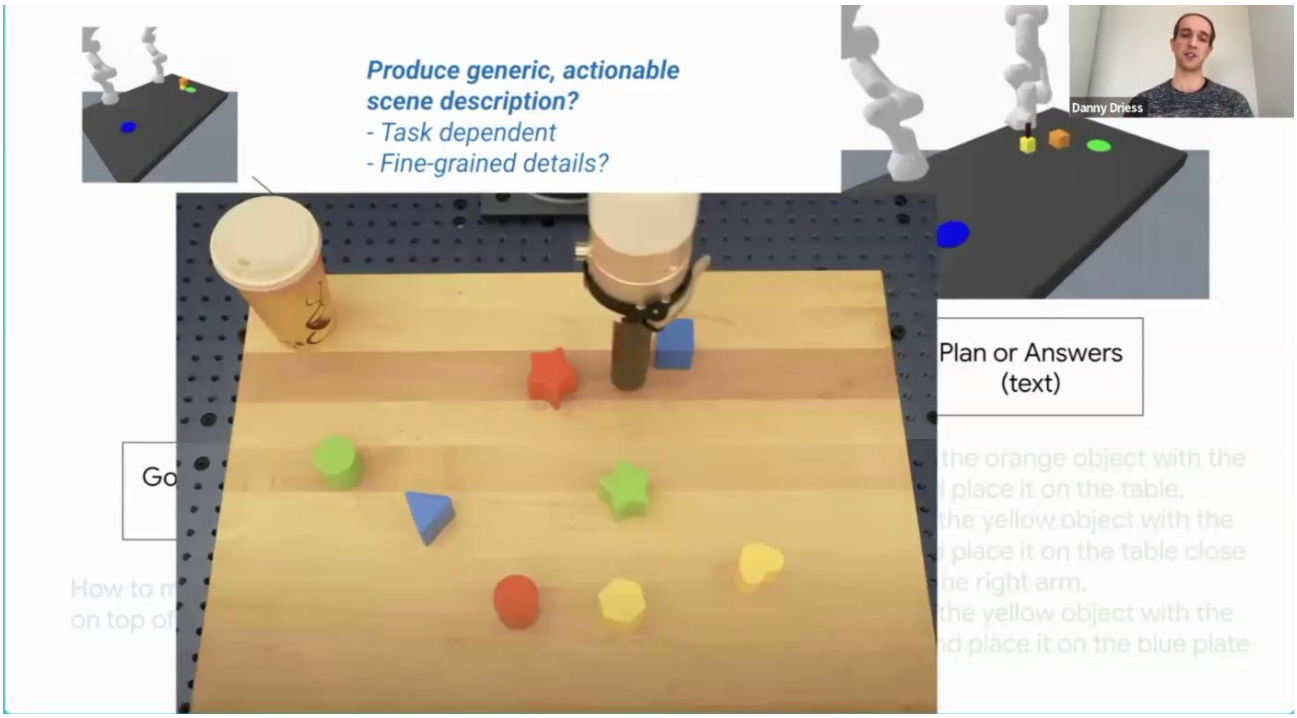


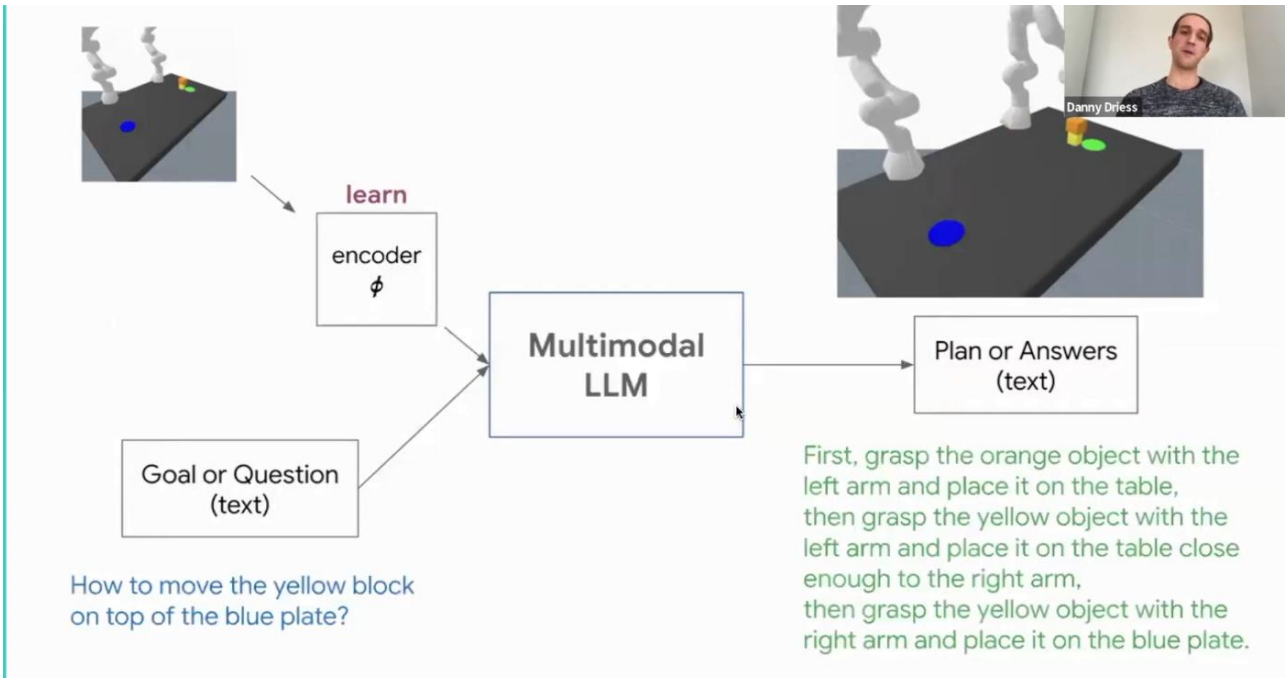

Danny Driess
Learned Model
Goal g
Plan $a_{1:K}$


Danny Driess
LLM
Goal or Question (text)
Plan or Answers (text)
How to move the yellow block on top of the blue plate?
First, grasp the orange object with the left arm and place it on the table, then grasp the yellow object with the left arm and place it on the table close enough to the right arm, then grasp the yellow object with the right arm and place it on the blue plate.

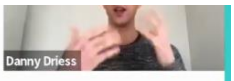

Danny Driess
LLM
Goal or Question (text)
Plan or Answers (text)
How to move the yellow block on top of the blue plate?
First, grasp the orange object with the left arm and place it on the table, then grasp the yellow object with the left arm and place it on the table close enough to the right arm, then grasp the yellow object with the right arm and place it on the blue plate.


Produce generic, actionable scene description?
- Task dependent
- Fine-grained details?
scene description (text)
LLM
Plan or Answers (text)
Goal or Question (text)
Danny Driess
How to move the yellow block on top of the blue plate?
First, grasp the orange object with the left arm and place it on the table.
then grasp the yellow object with the left arm and place it on the table close enough to the right arm.
then grasp the yellow object with the right arm and place it on the blue plate.


Produce generic, actionable scene description?
- Task dependent
- Fine-grained details?
Plan or Answers (text)
Danny Driess



Research Questions

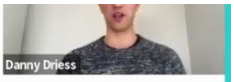


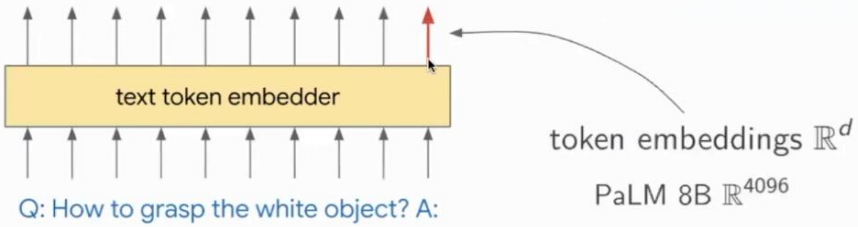
1. Can multimodal features (pixels or states) be input to an LLM while leveraging the generalization of LLMs for embodied reasoning and planning?

2. Large-scale model + large scale data – do we get ***transfer***?

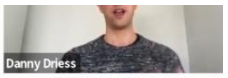
   - Scaling the LLM (8B, 62B, 540B)
   - Scaling the ViT (4B, 22B)
   - Scaling to all robotics data
   - Co-training on general VQA tasks

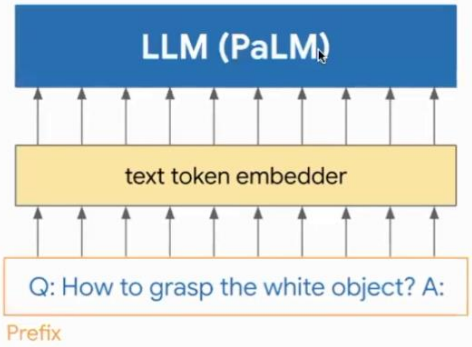
PaLM-E – Injecting multimodal information into PaLM



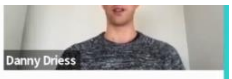


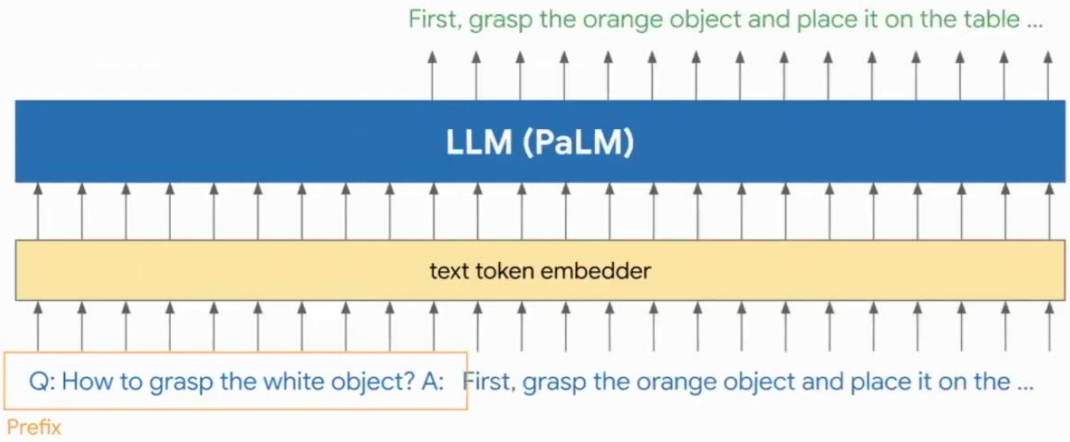
PaLM-E – Injecting multimodal information into PaLM



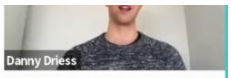


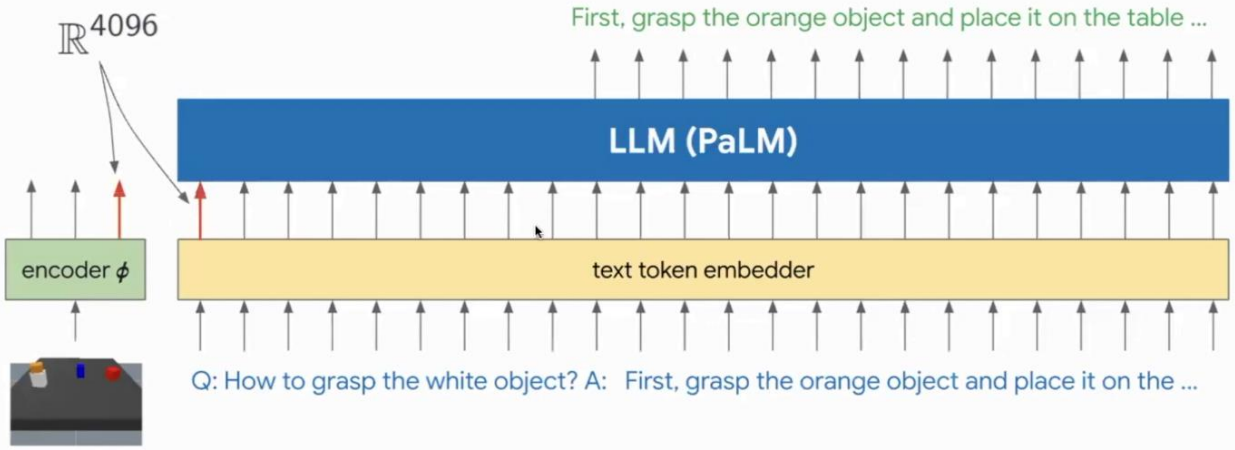
PaLM-E – Injecting multimodal information into PaLM





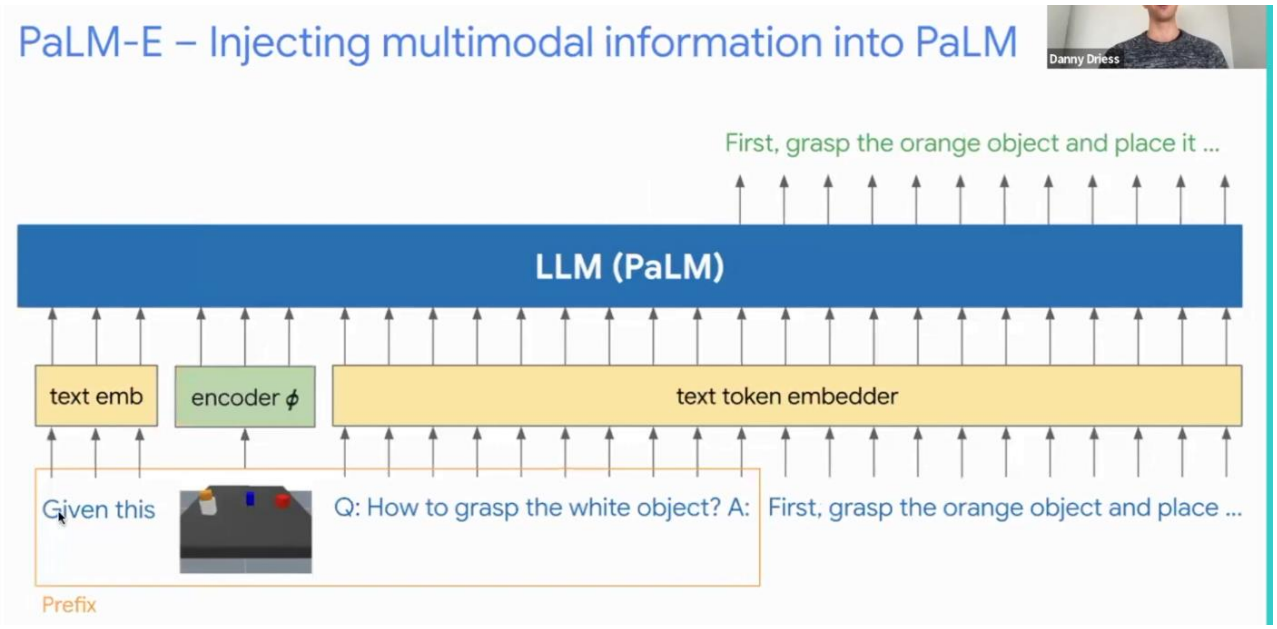
PaLM-E – Injecting multimodal information into PaLM



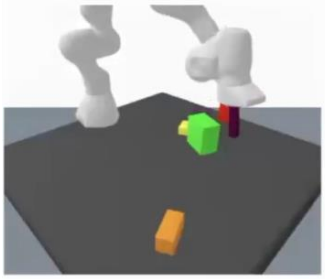


$$\phi : \mathbb{R}^{h \times w \times 3} \rightarrow \mathbb{R}^{m \times 4096}$$

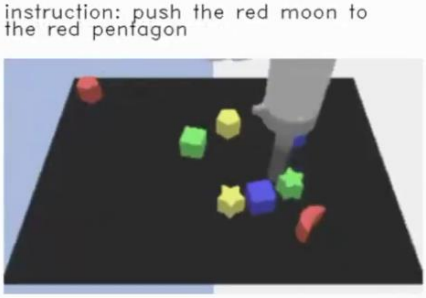
PaLM-E – Injecting multimodal information into PaLM



PaLM-E: Robot Environments



Task and Motion Planning



Language Table (Lynch et al., 2023)

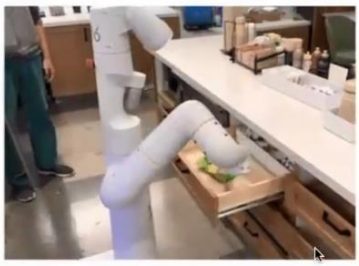
Sim and Real

SayCan (Ahn et al., 2022)

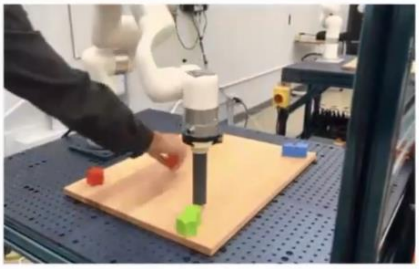
PaLM-E: Training Data – Full mixture

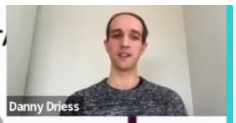
Dataset in full mixture	Sampling frequency	%
Webli (Chen et al., 2022)	100	52.4
VQ^{2}A (Changpinyo et al., 2022)	25	13.1
VQG (Changpinyo et al., 2022)	10	5.2
CC3M (Sharma et al., 2018)	25	13.1
Object Aware (Piergiovanni et al., 2022)	10	5.2
OKVQA (Marino et al., 2019)	1	0.5
VQAv2 (Goyal et al., 2017)	1	0.5
COCO (Chen et al., 2015)	1	0.5
Wikipedia text	1	0.5
(robot) Mobile Manipulator, real	6	3.1
(robot) Language Table (Lynch et al., 2022), sim and real	8	4.2
(robot) TAMP, sim	3	1.6

closed-loop end-to-end planning
("bring me the rice chips from the drawer ")



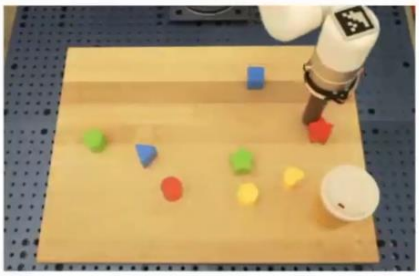
long-horizon tasks
("sort the blocks by colors into corners")

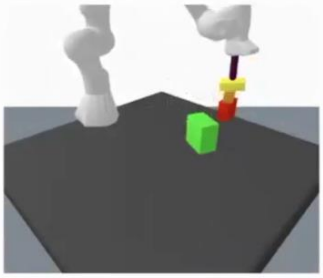




cross-embodiment generalization

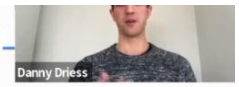
zero-shot





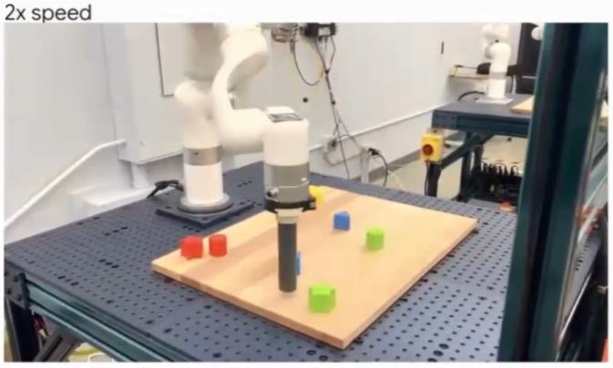
SayCan Tasks

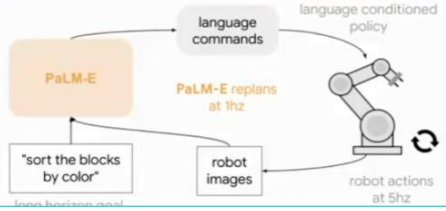
Language Table (48-



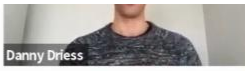
4x speed

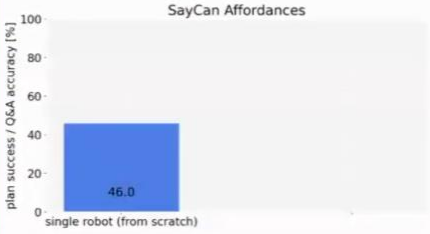
2x speed

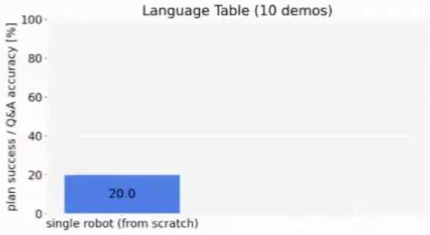


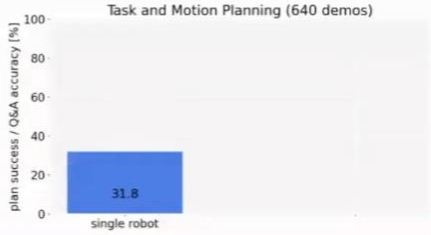


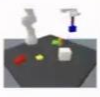
Transfer



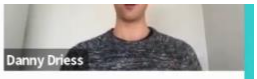


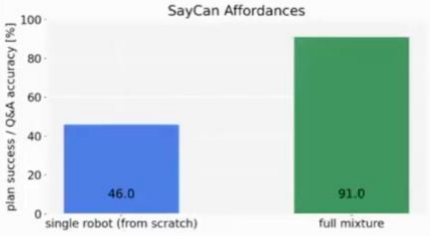


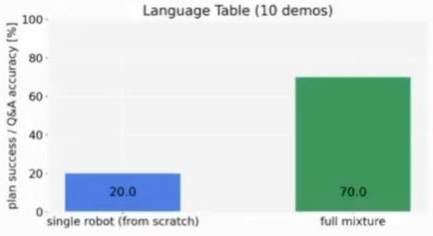




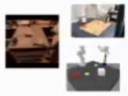
Transfer







all robot data



\+
webli, VQA v2
COCO, OKVQA etc.

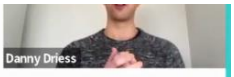
all robot data

\+
webli, VQA v2
COCO, OKVQA etc.

all robot data

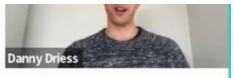
\+
webli, VQA v2
COCO, OKVQA etc.

PaLM-E is also a competent general VLM model



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

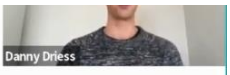
Qualitative Examples

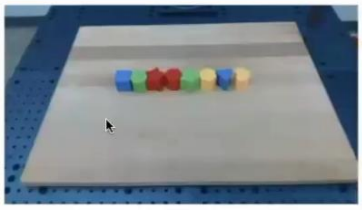


Given **<img>**. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.

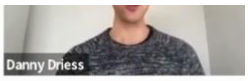
Given **<img>**. Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step. 1. do not enter. 2. except bicycles. 3. do not entry except bicycles. 4. yes.

Qualitative Examples

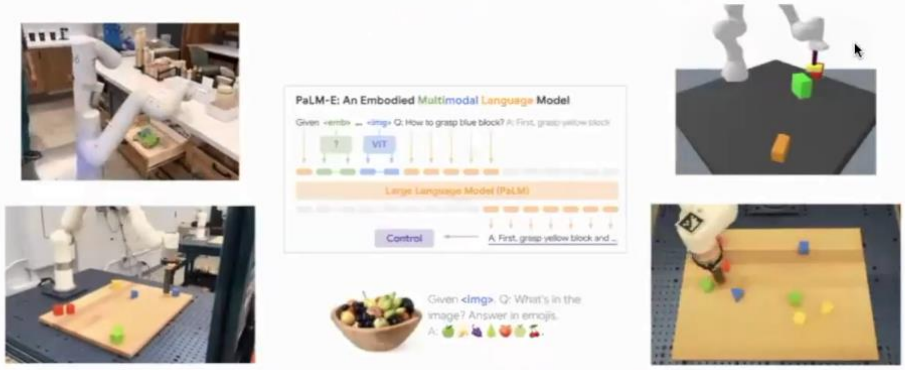


Given  . Q: What was the task? A: put all the blocks in a horizontal line.

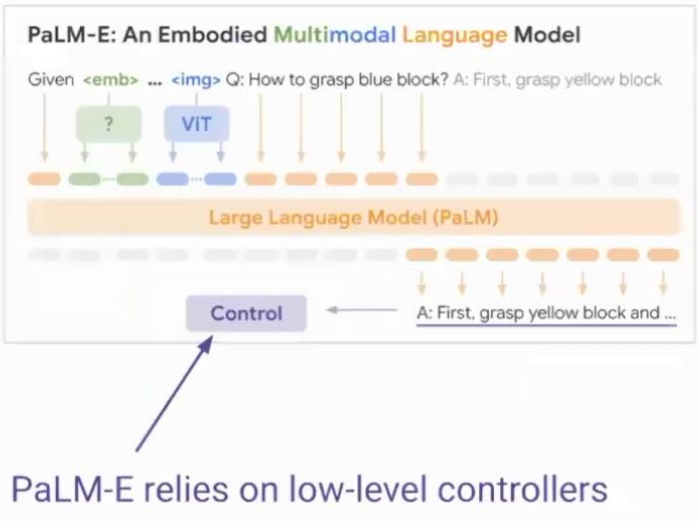
PaLM-E Conclusion



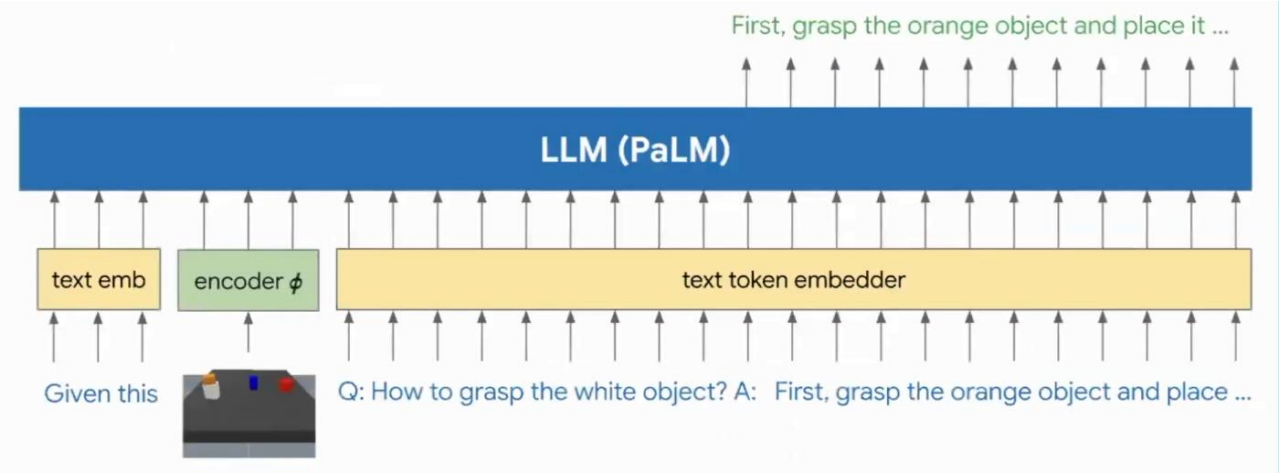
- Embodied reasoning as vision-language modeling
- Transfer across **robot** and general **vision-language** datasets
- PaLM-E also competitive in general VQA, captioning, language tasks

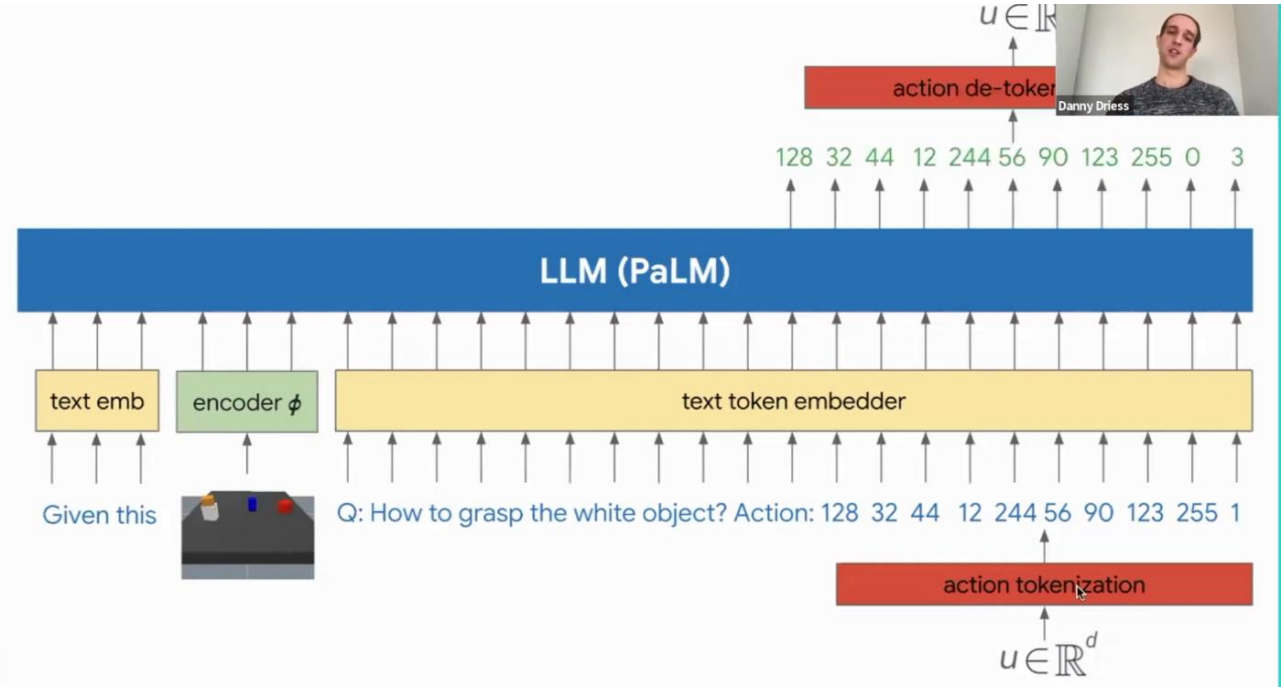


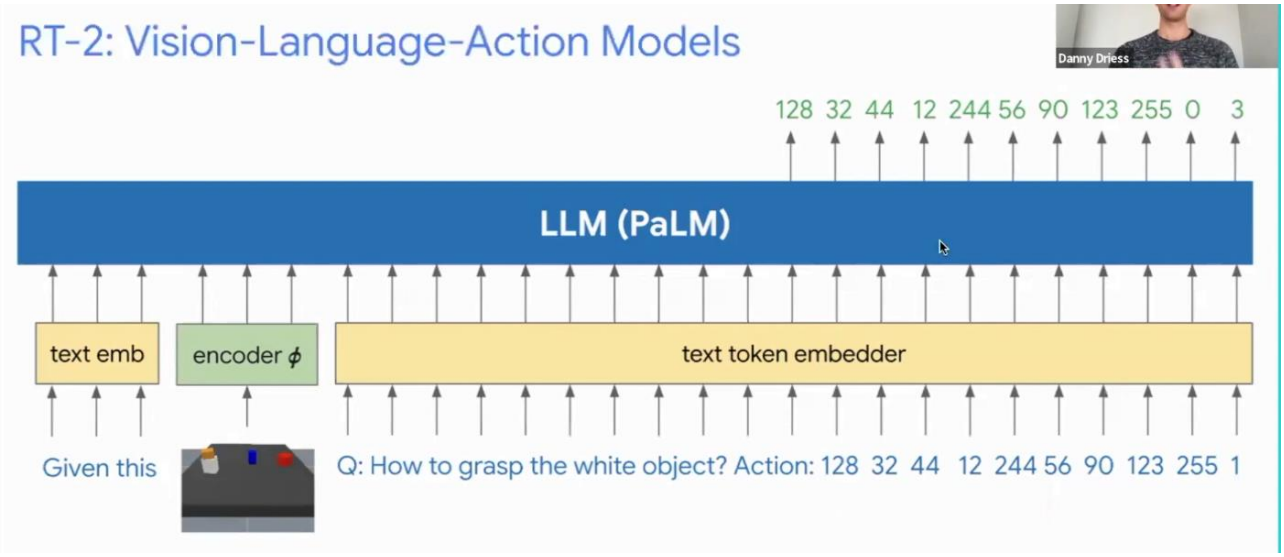
However



PaLM-E relies on low-level controllers

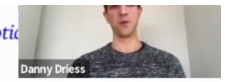

First, grasp the orange object and place it ...
LLM (PaLM)
text emb
encoder ϕ
text token embedder
Given this
Q: How to grasp the white object? A: First, grasp the orange object and place ...


$u \in \mathbb{R}$
action de-toke
Danny Driess
128 32 44 12 244 56 90 123 255 0 3
LLM (PaLM)
text emb
encoder ϕ
text token embedder
Given this
Q: How to grasp the white object? Action: 128 32 44 12 244 56 90 123 255 1
action tokenization
$u \in \mathbb{R}^d$


RT-2: Vision-Language-Action Models
Danny Driess
128 32 44 12 244 56 90 123 255 0 3
LLM (PaLM)
text emb
encoder ϕ
text token embedder
Given this
Q: How to grasp the white object? Action: 128 32 44 12 244 56 90 123 255 1

Google DeepMind

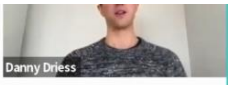
https://robotic

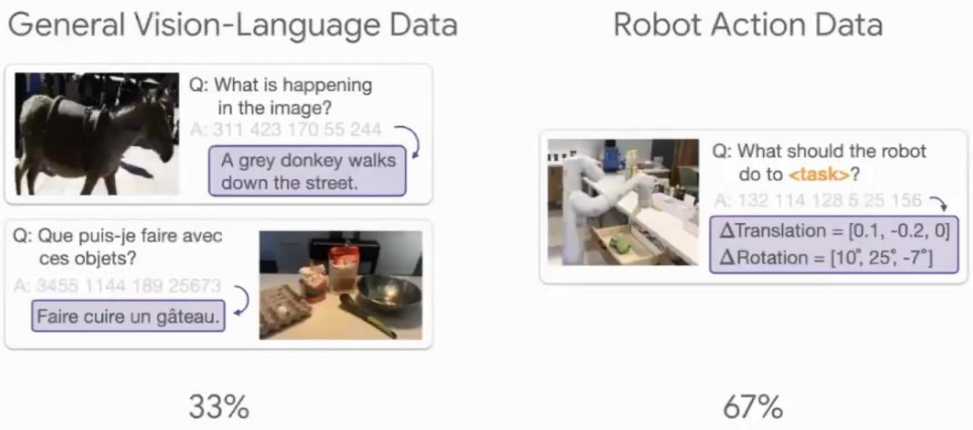


RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

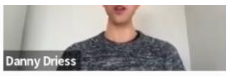
Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex 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, and Brianna Zitkovich

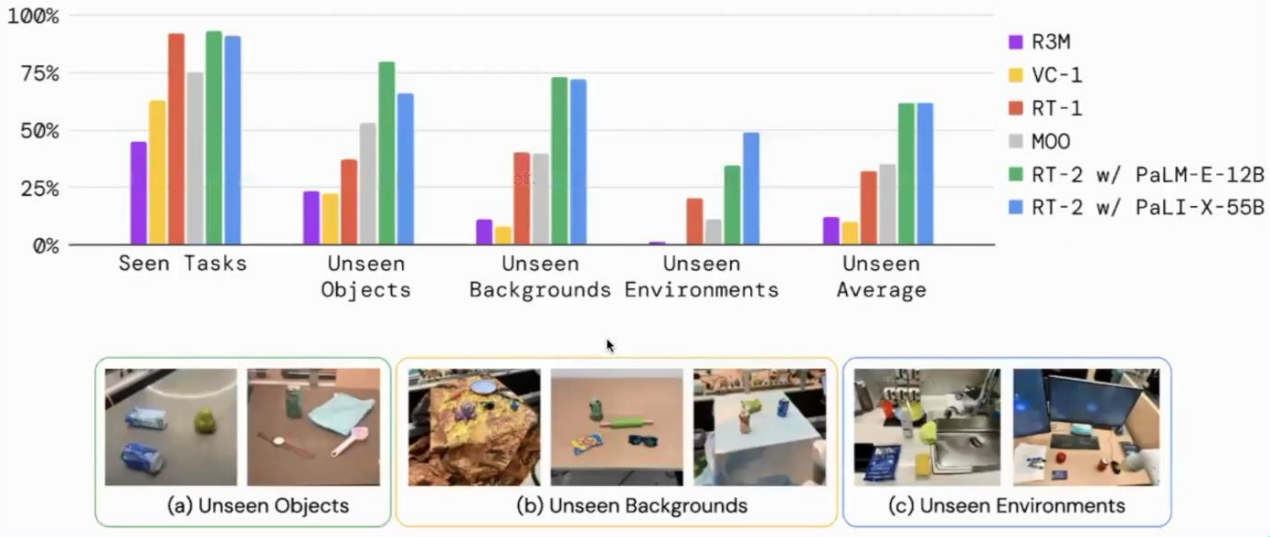
RT-2: Training Data



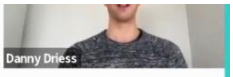


Generalization





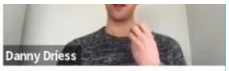
Generalization

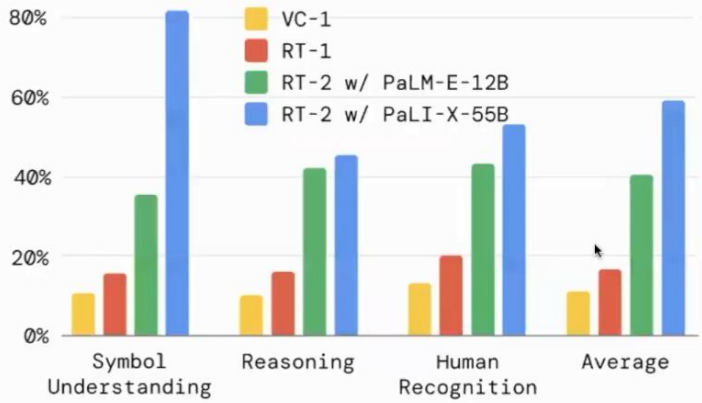


Pick the extinct animal

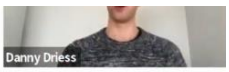
Pick the lion

Generalization





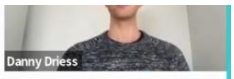
RT-2 Conclusion



- Transfer of internet-scale semantic knowledge into robotics !!!!!
- Mostly semantic pick-and-place

dexterity?

Dexterity & Action Chunking



- Predict "action chunk"

Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware

Tony Z. Zhao[1] Vikash Kumar[3] Sergey Levine[2] Chelsea Finn[1]
[1] Stanford University [2] UC Berkeley [3] Meta

ALOHA Unleashed: A Simple Recipe for Robot Dexterity

Tony Z. Zhao, Jonathan Tompson, Danny Driess, Pete Florence,
Kamyar Ghasemipour, Chelsea Finn, Ayzaan Wahid*

Google DeepMind

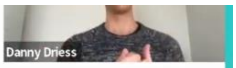
Action chunk is all about predicting a whole trajectory of what the robot should be doing next

π_0: A Vision-Language-Action Flow Model for General Robot Control

Physical Intelligence

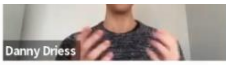
Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, Ury Zhilinsky

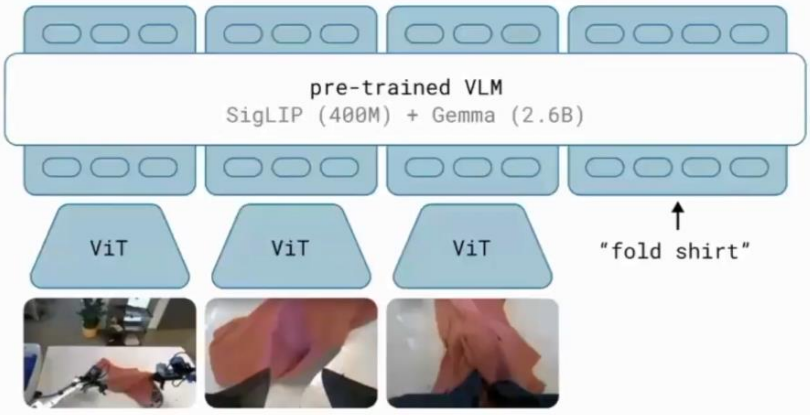
Goals



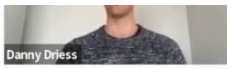
- Generalist robot model for multiple embodiments
- Long-horizon, **dexterous** tasks
- Fast enough inference

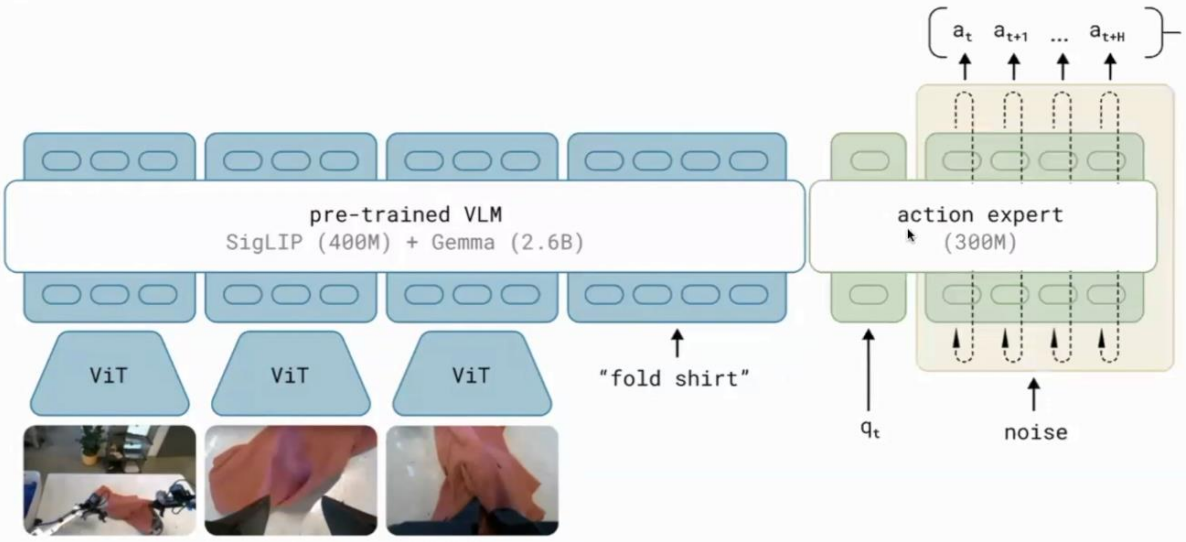
π_0 Model



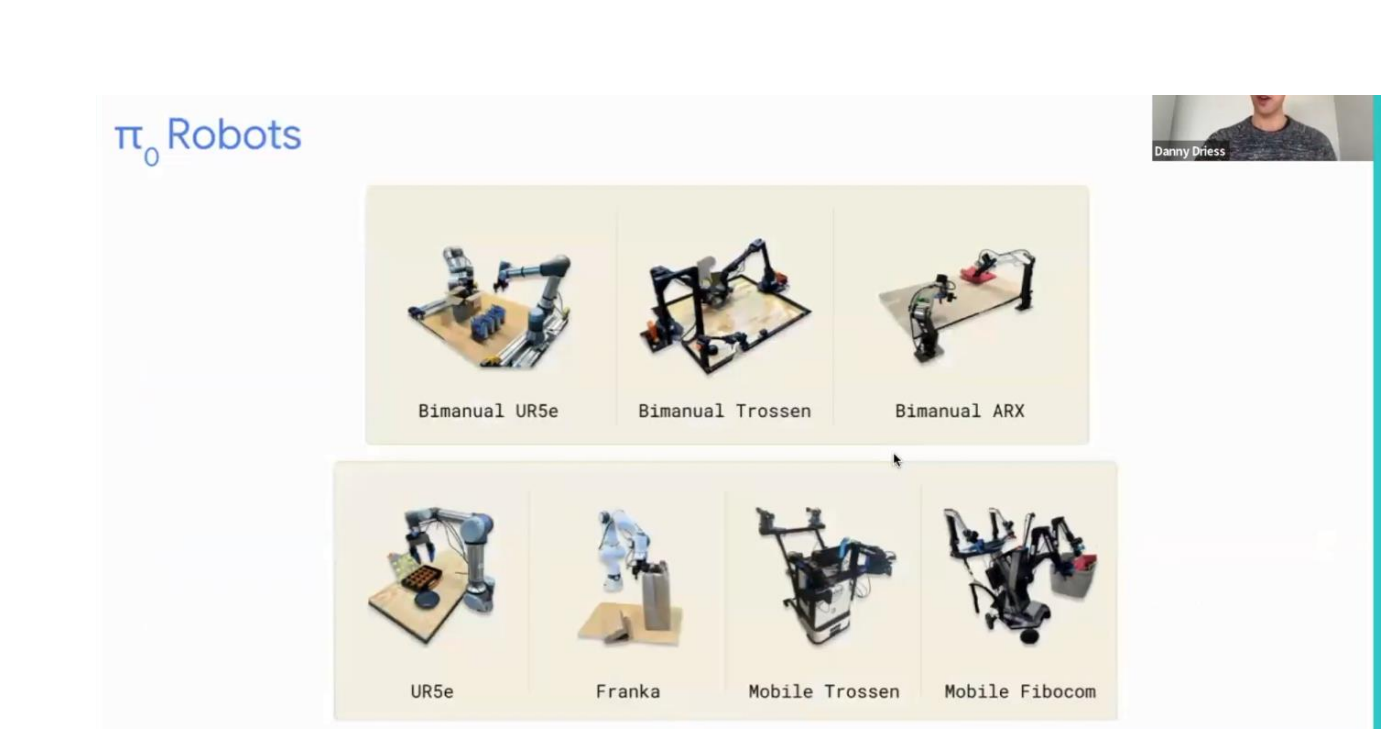


π_0 Model

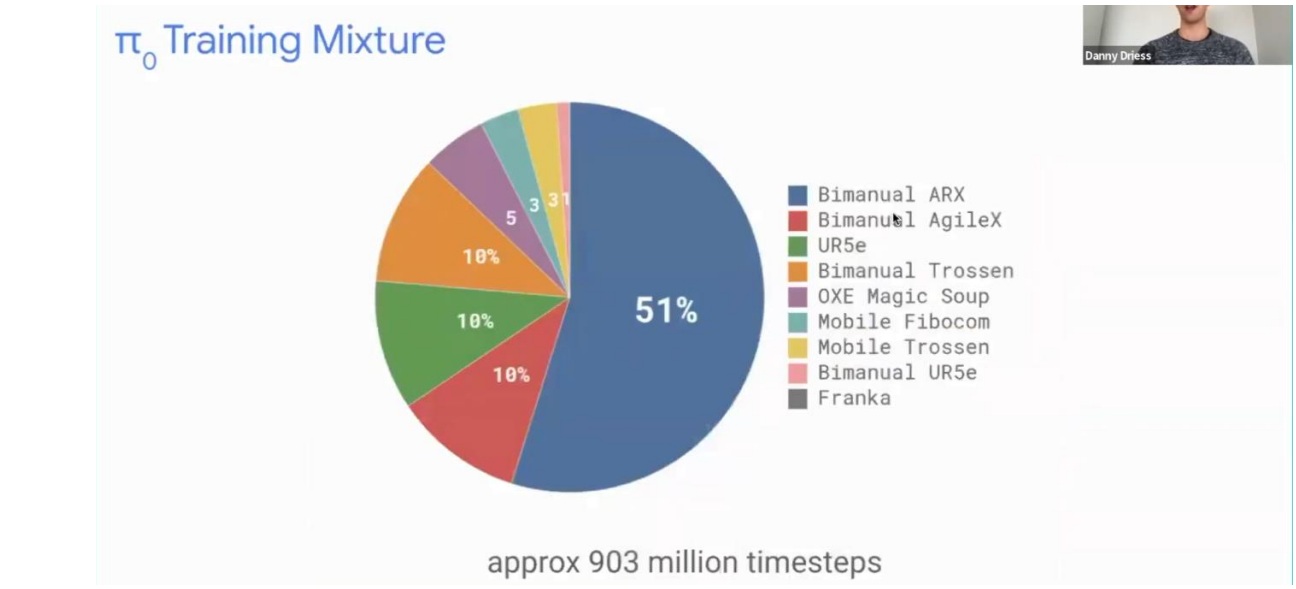




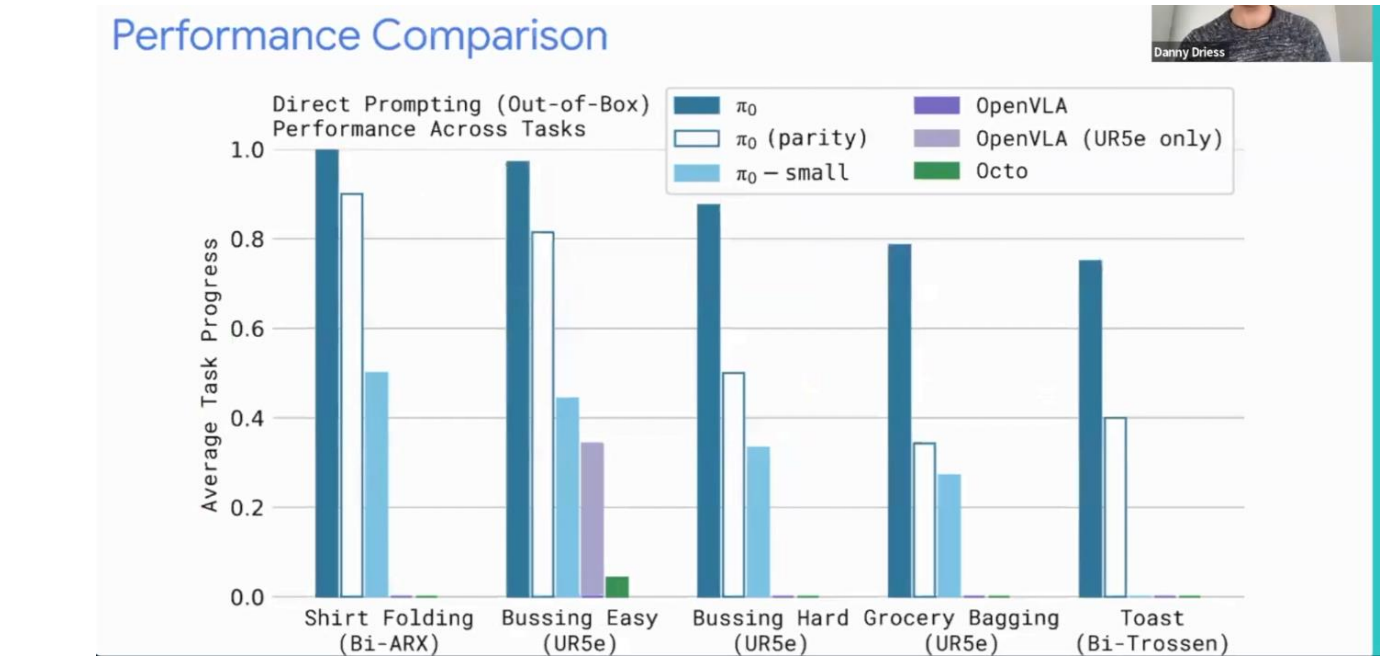
π_0 Robots



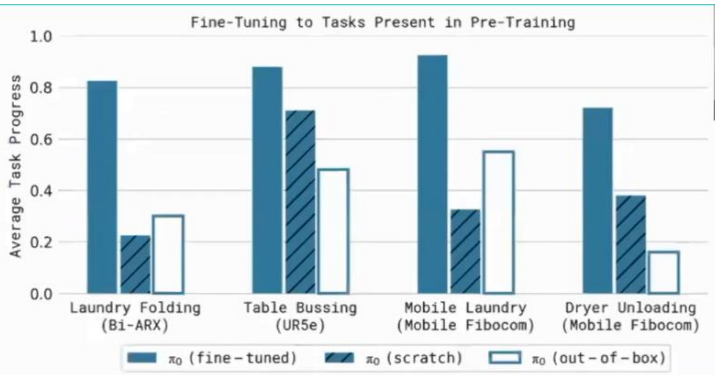
π_0 Training Mixture

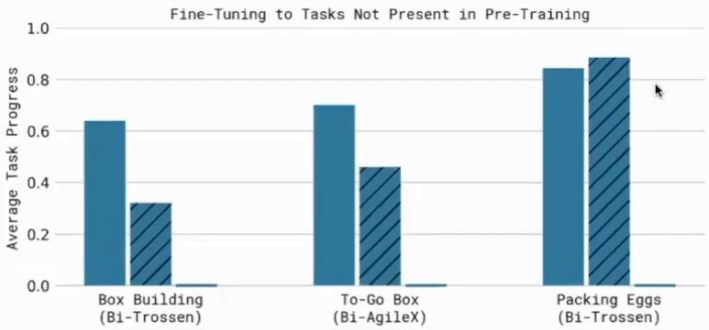


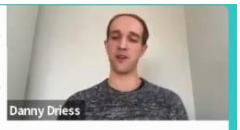
Performance Comparison

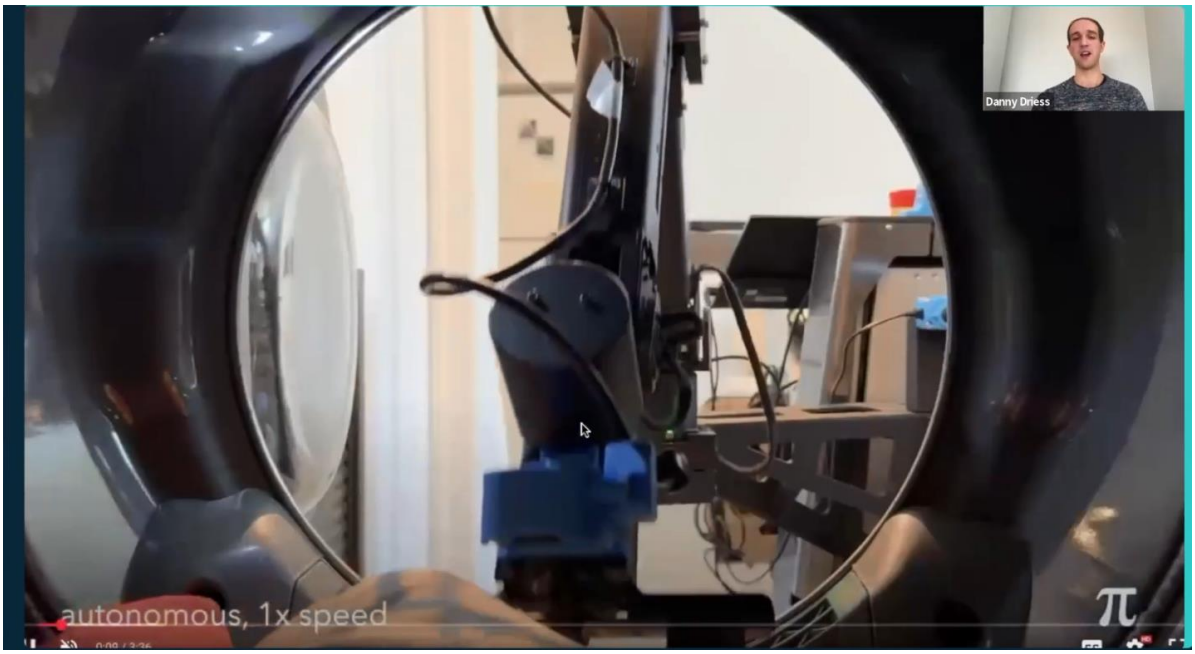


Finetuning

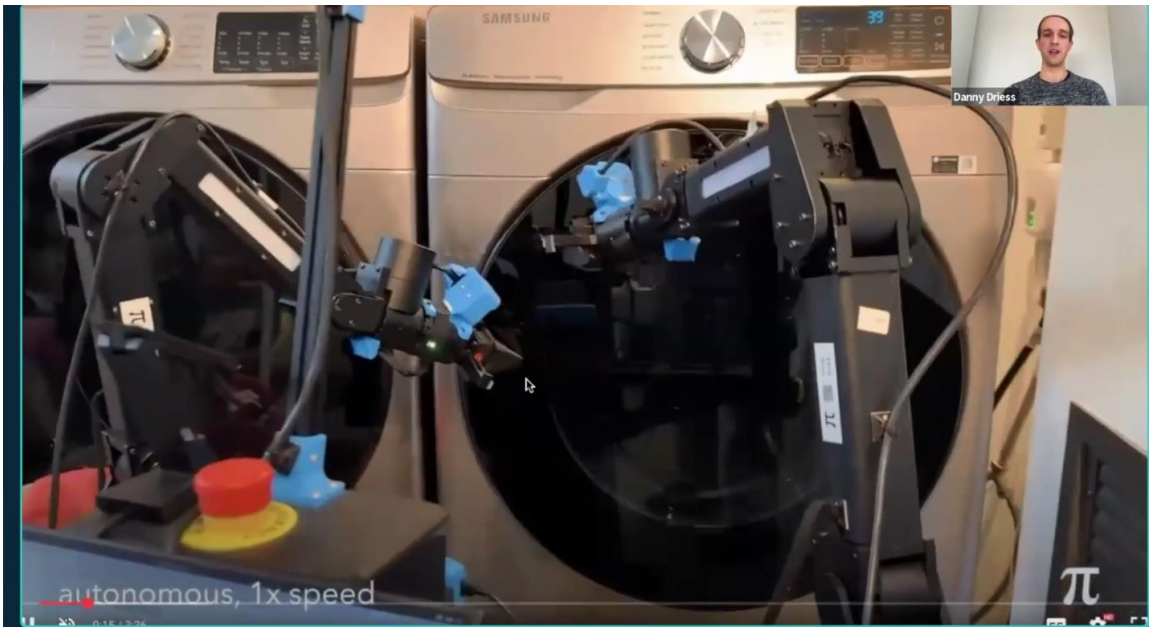

Fine-Tuning to Tasks Present in Pre-Training
Average Task Progress
1.0
0.8
0.6
0.4
0.2
0.0
Laundry Folding (Bi-ARX)
Table Bussing (UR5e)
Mobile Laundry (Mobile Fibocom)
Dryer Unloading (Mobile Fibocom)
π_0 (fine-tuned)
π_0 (scratch)
π_0 (out-of-box)


Fine-Tuning to Tasks Not Present in Pre-Training
Average Task Progress
1.0
0.8
0.6
0.4
0.2
0.0
Box Building (Bi-Trossen)
To-Go Box (Bi-AgileX)
Packing Eggs (Bi-Trossen)


Danny Driess

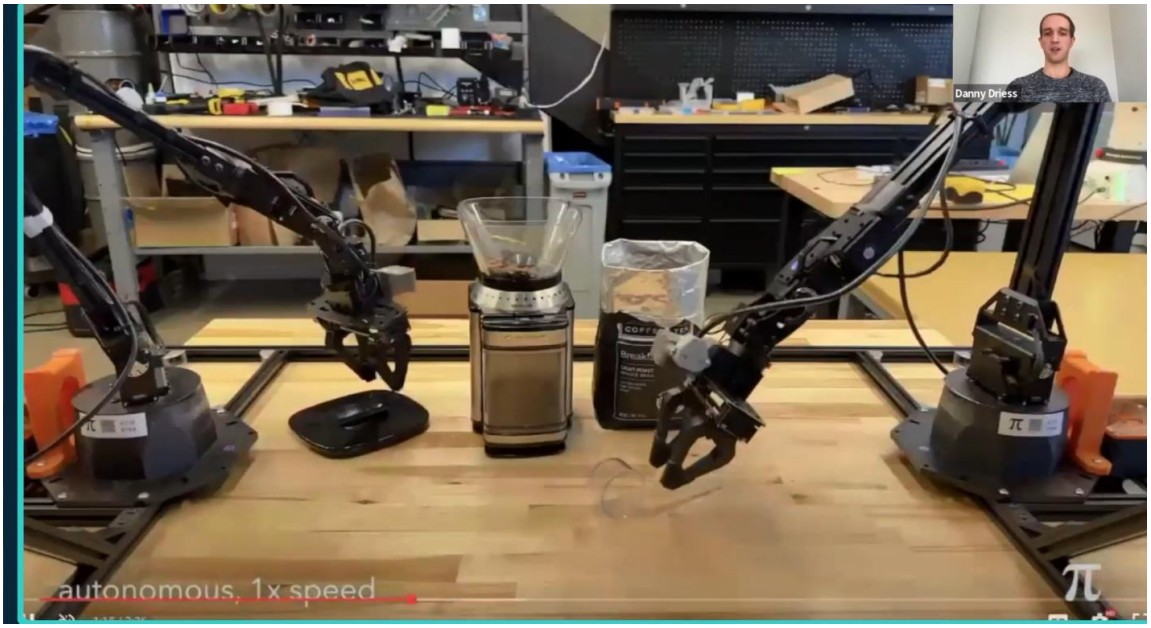

Danny Driess
autonomous, 1x speed

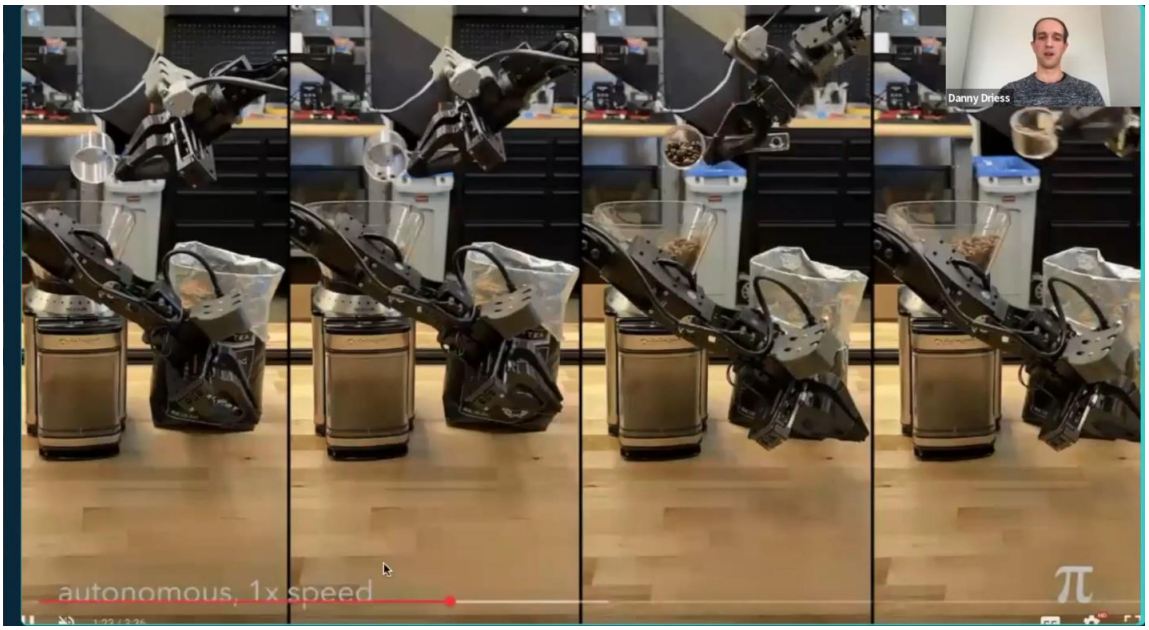

Danny Driess
autonomous, 1x speed

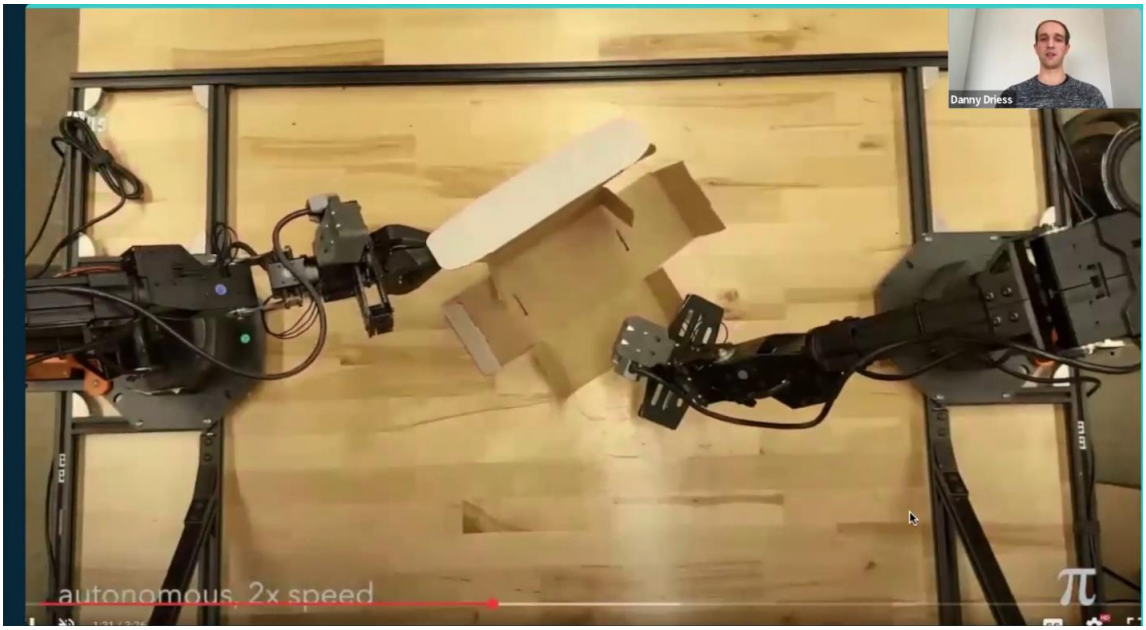

SAMSUNG
Danny Driess
autonomous, 1x speed


Danny Driess
autonomous, 1x speed


Danny Driess
autonomous, 1x speed


Danny Driess
autonomous, 1x speed

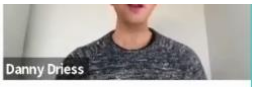

Danny Driess
autonomous, 1x speed

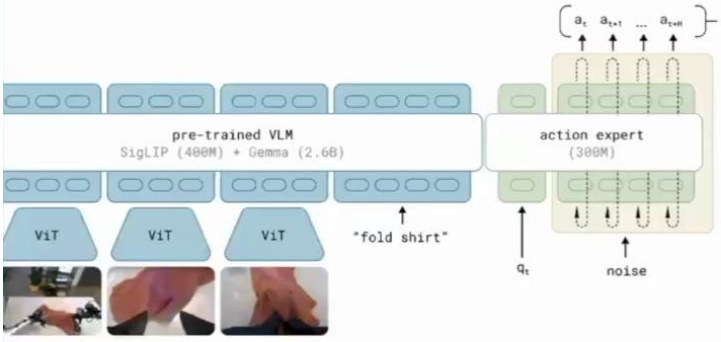

Danny Driess
autonomous, 2x speed


Danny Driess
autonomous, 1x speed


Danny Driess
autonomous, 4x speed

Why not RT-2 style VLA?

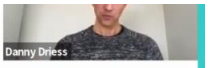




π_0 model

RT-2 style model

Why not RT-2 style VLA?



- Action-chunking required to make dexterous policies work
- Action-chunking with RT-2

700 tokens for 50 Hz biarm robot

- Slow inference
- Models struggle to learn

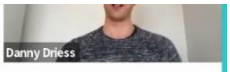
FAST: Efficient Action Tokenization for Vision-Language-Action Models

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, Sergey Levine

FAST allows us to train VLAs on high frequency action data

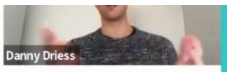
FAST Tokenizer

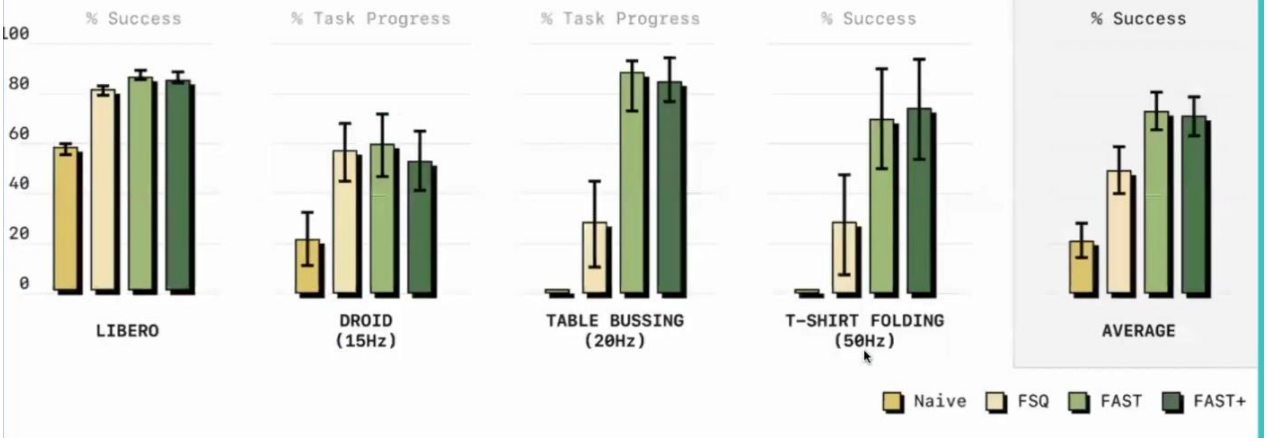
FAST Tokenizer Compression



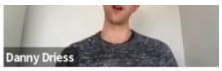
Dataset	Action Dimension	Control Frequency	Avg. Token Naive	Avg. Token FAST	Compression
BridgeV2	7	5 Hz	35	20	1.75
DROID	7	15 Hz	105	29	3.6
Bussing	7	20 Hz	140	28	5.0
Shirt Fold	14	50 Hz	700	53	13.2

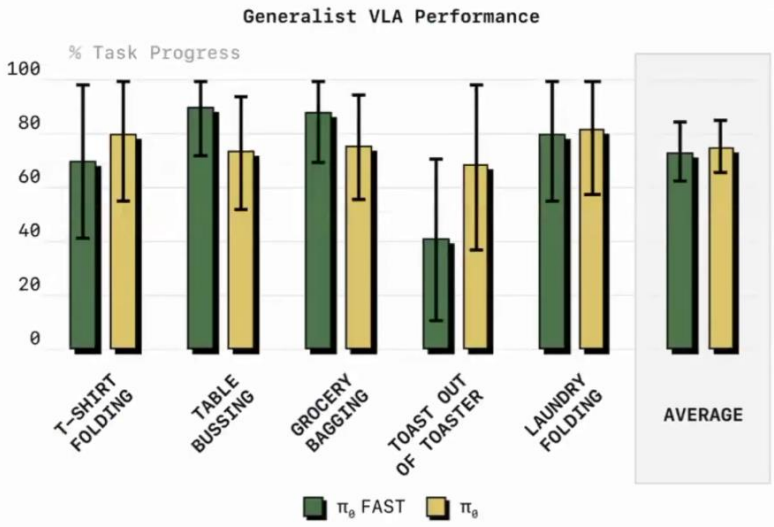
Comparison To Other Tokenization Schemes



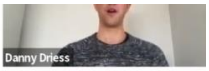


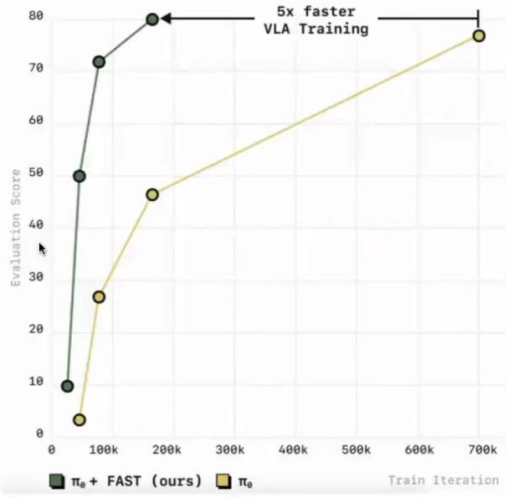
Comparison to Flow-Matching



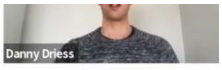


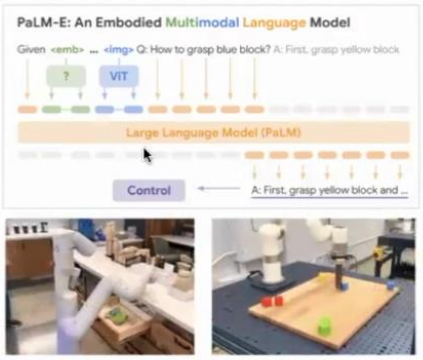
π_0-FAST trains faster

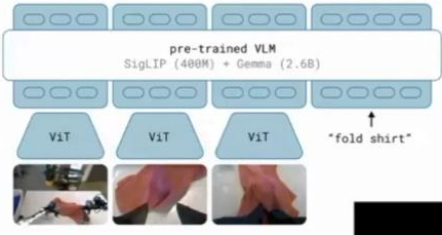


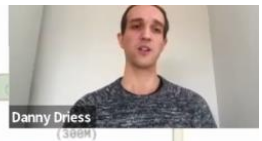


π_0-FAST on DROID (https://droid-dataset.github.io/)



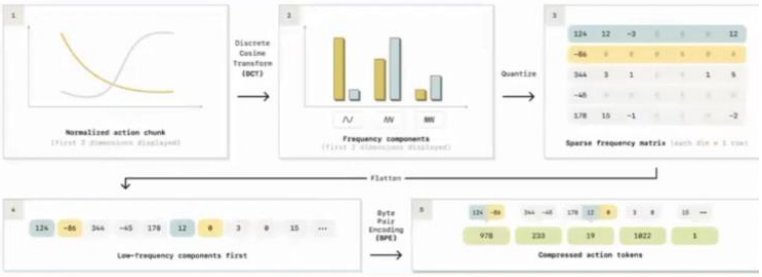

PaLM-E: An Embodied Multimodal Language Model
Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block
?
ViT
Large Language Model (PaLM)
Control
A: First, grasp yellow block and ...


pre-trained VLM
SigLIP (400M) + Gemma (2.6B)
(300M)
ViT
ViT
ViT
"fold shirt"
q_t
noise


Danny Driess


π_0
2:28 / 3:36


$u \in \mathbb{R}^d$
action de-tokenization
128 32 44 12 244 56 90 123 255 0 3
LLM (PaLM)
text emb
encoder
text token embedder
Given this
Q: How to grasp the white object? Action: 128 32 44 12 244 56 90 123 255 1
action tokenization
$u \in \mathbb{R}^d$


1
Normalized action chunk
Discrete Cosine Transform (DCT)
2
Frequency components
Quantize
3
Sparse frequency matrix
Flatten
4
Low-frequency components first
Byte Pair Encoding (BPE)
5
Compressed action tokens