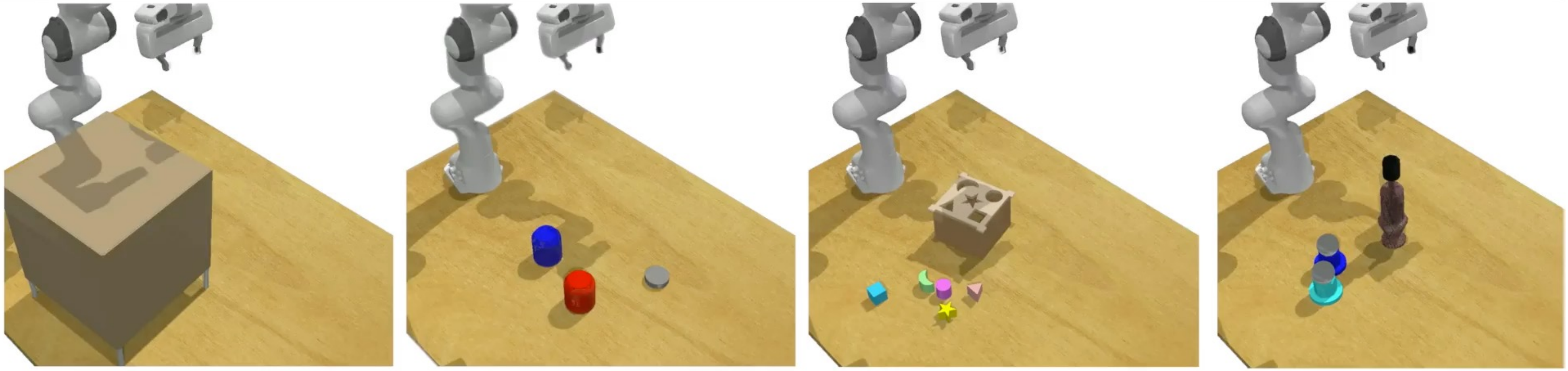
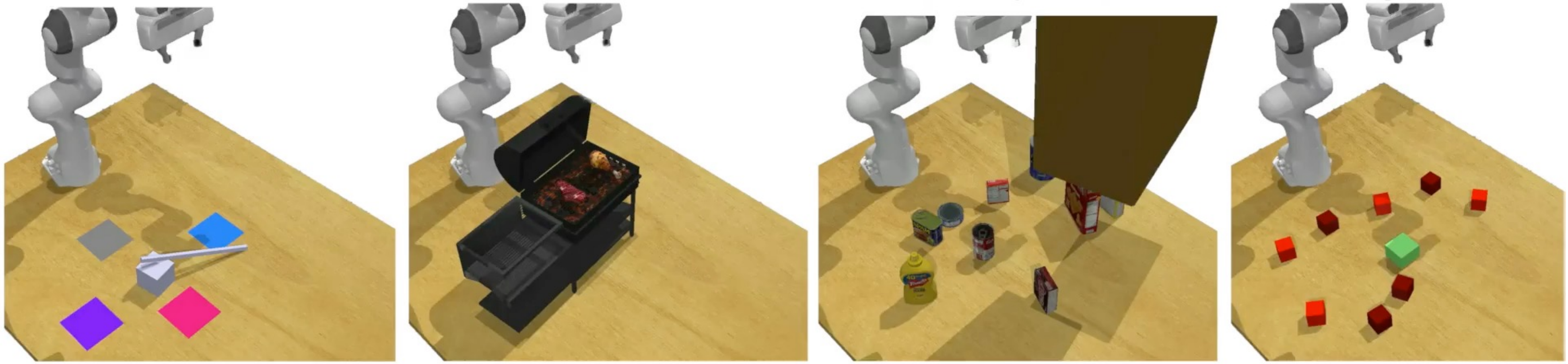
Acting with Perception and Language

Mohit Shridhar



"open the middle drawer" "close the blue jar" "put the cylinder in the shape sorter" "screw in the blue light bulb"



"use the stick to drag the cube onto the gray target" "take the steak off the grill" "put the mustard in the cupboard" "stack 2 maroon blocks"

So far ...

Object Detection

Object Pose Estimation

What are the **objects** here?

There is something weird

about **objects** 🤔

David Marr

“Vision is a computational process that transforms the retinal image into an objective representation of 3D shape.”

Other Perspectives?

Prof. Lana Lazebnik
UIUC

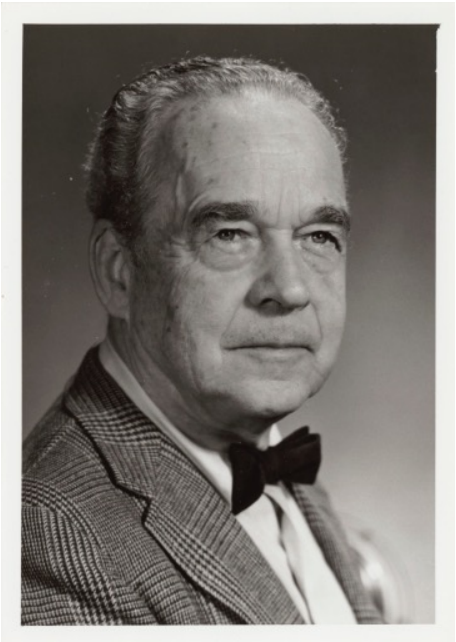
https://slazebni.cs.illinois.edu/spring20

Three Perspectives on Vision

David Marr

"Vision is a computational process that transforms the retinal image into an objective representation of 3D shape."

James Gibson



"There is no computation. There is no retinal image. There are no representations. There is no 3D shape. There is only direct pickup of ecologically relevant variants and invariants. Vision is in the world, not the observer."

Credit: Lana Lazebnik

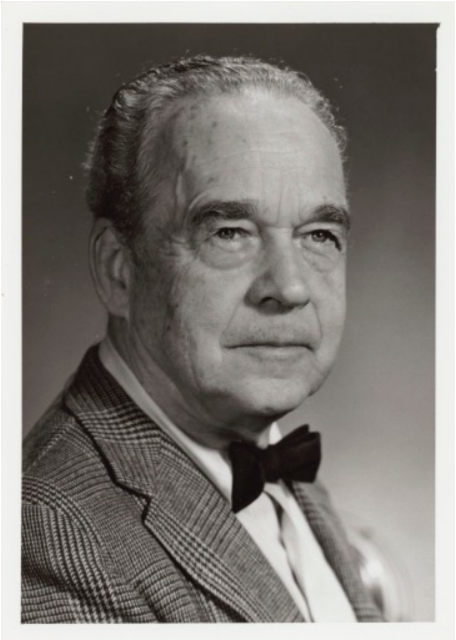
Affordances

Three Perspectives on Vision

David Marr

“Vision is a computational process that transforms the retinal image into an objective representation of 3D shape.”

James Gibson



“There is no computation. There is no retinal image. There are no representations. There is no 3D shape. There is only direct pickup of ecologically relevant variants and invariants. Vision is in the world, not the observer.”

Jan Koenderink

“There is no objective world, only the observer’s *umwelt*. Thus, vision cannot be in the world but is a creative act of the observer.”

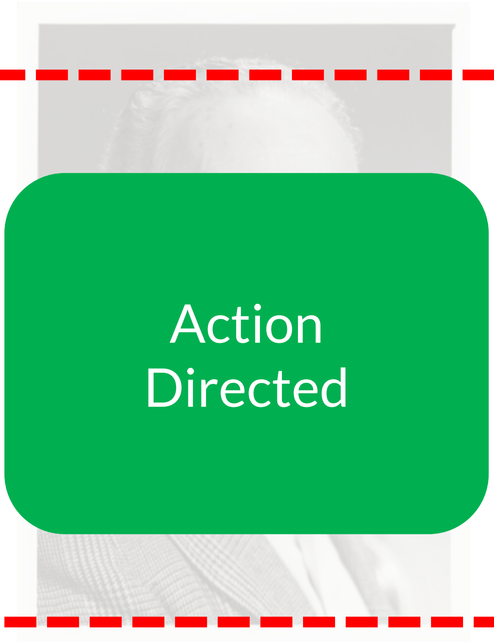
Credit: Lana Lazebnik


Umwelt
An ant has difference
experience and goals

Three Perspectives on Vision

David Marr

James Gibson



This Lecture

Jan Koenderink

"Vision is a computational process that transforms the retinal image into an objective representation of 3D shape."

"There is no computation. There is no retinal image. There are no representations. There is no 3D shape. There is only direct pickup of ecologically relevant variants and invariants. Vision is in the world, not the observer."

"There is no objective world, only the observer's *umwelt*. Thus, vision cannot be in the world but is a creative act of the observer."

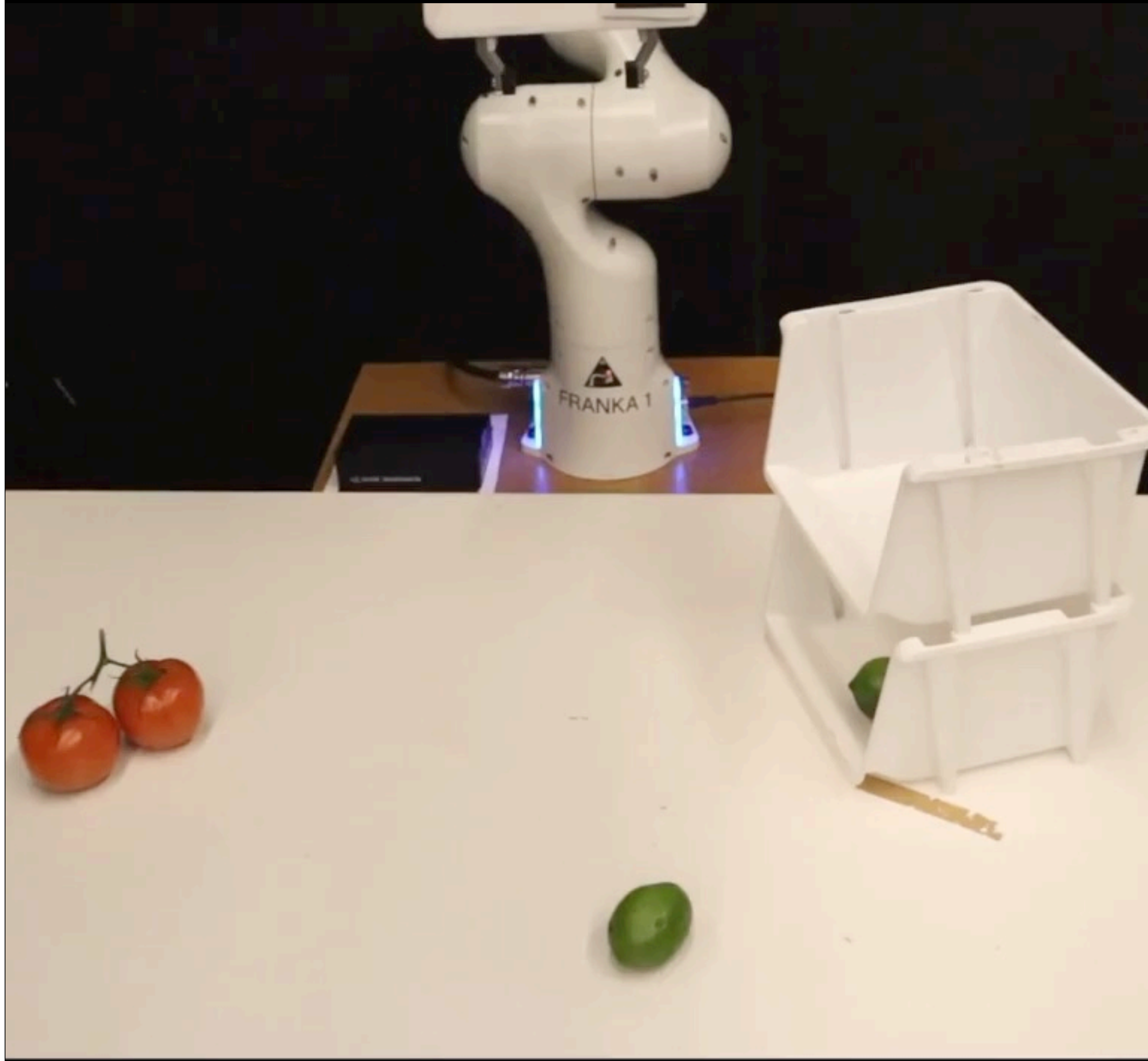
Credit: Lana Lazebnik

Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation

Mohit Shridhar[1], Lucas Manuelli[2], Dieter Fox[1, 2]
[1]University of Washington, [2]NVIDIA

Perceiver-Actor

Multi-task 6-DoF manipulation agent

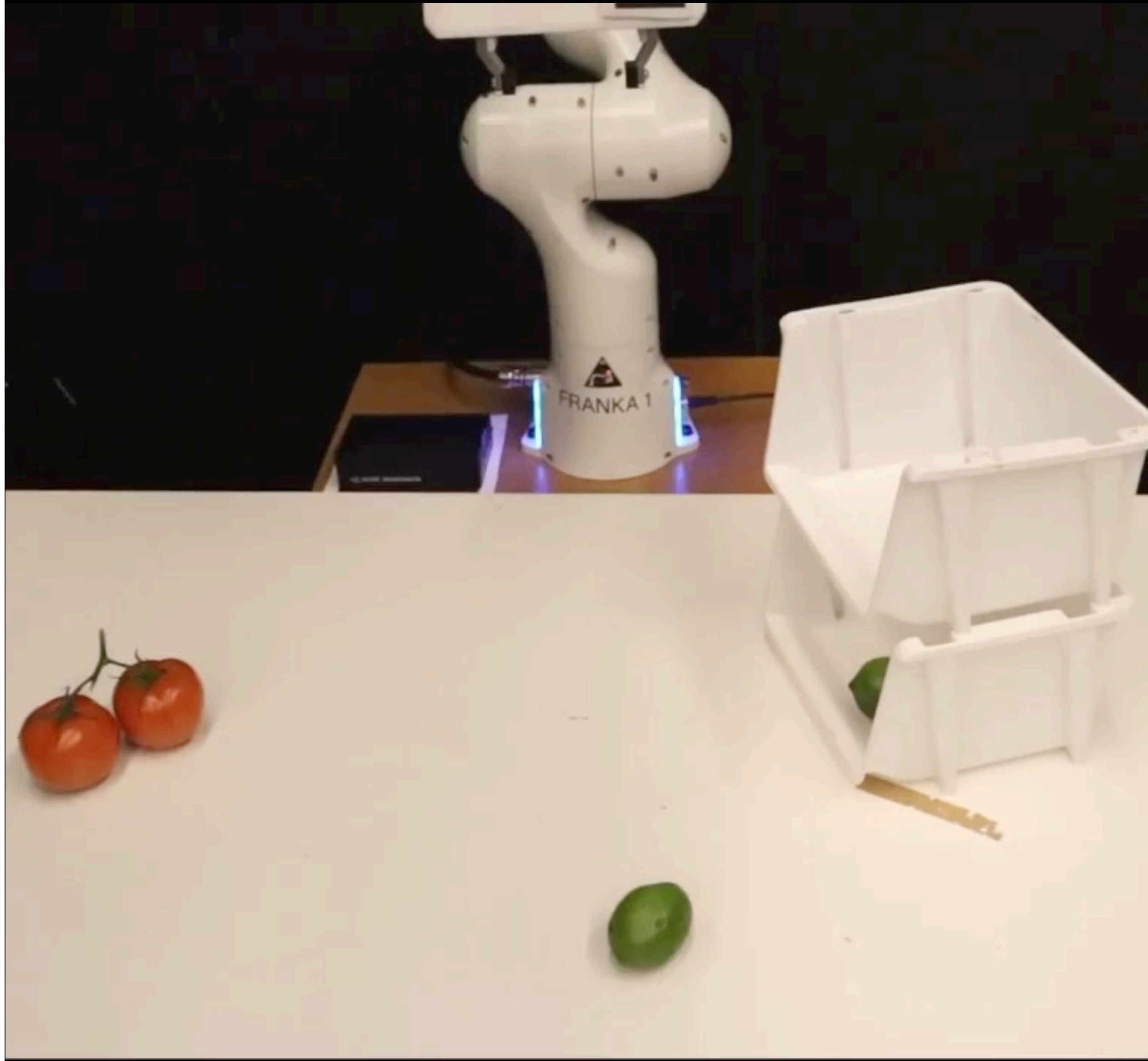


“put the tomatoes in the top bin”

Perceiver-Actor

Multi-task 6-DoF manipulation agent

End-to-end few-shot imitation learning



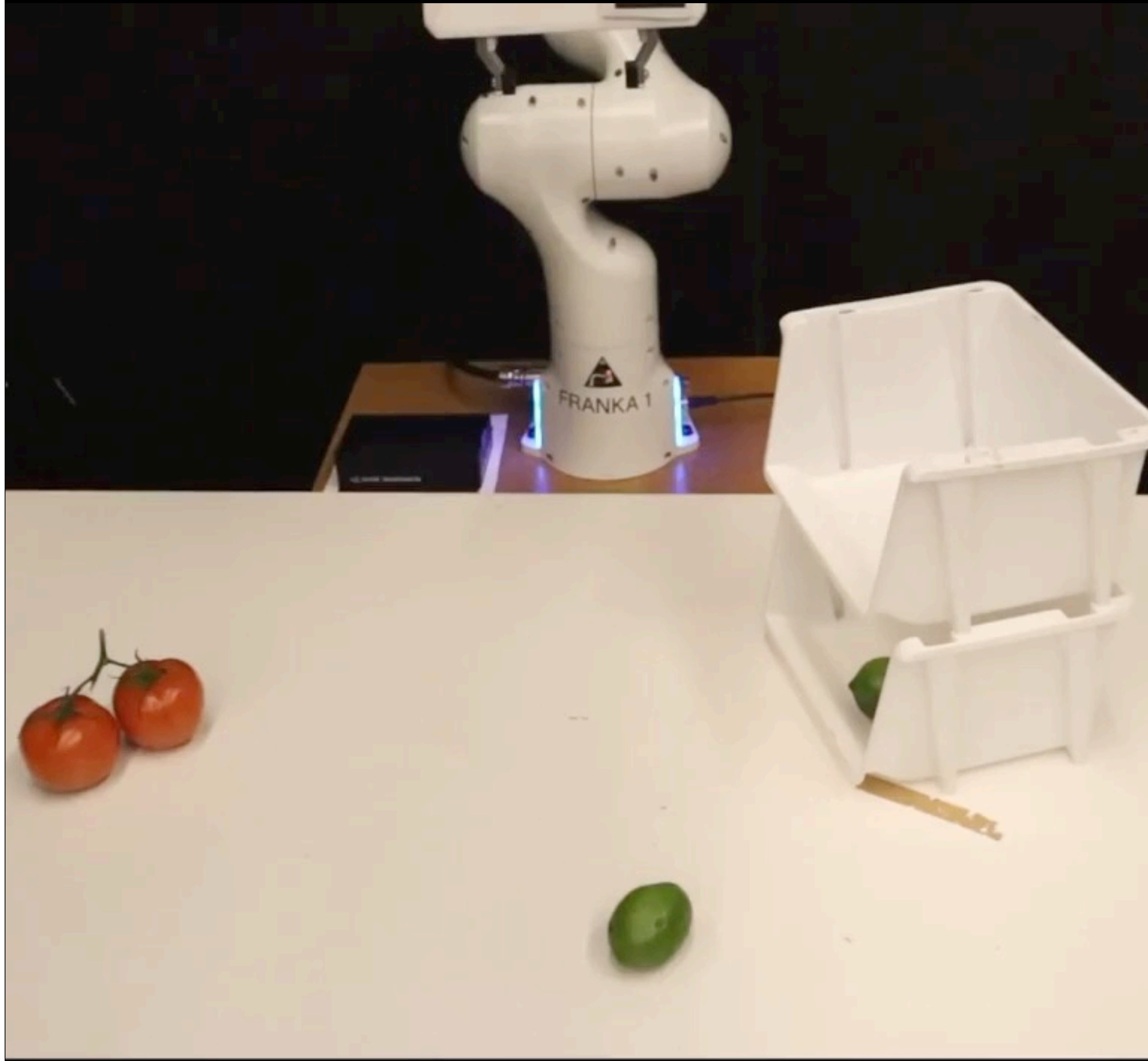
"put the tomatoes in the top bin"

Perceiver-Actor

Multi-task 6-DoF manipulation agent

End-to-end few-shot imitation learning

Input: RGB-D Voxels & Language Goal



"put the tomatoes in the top bin"

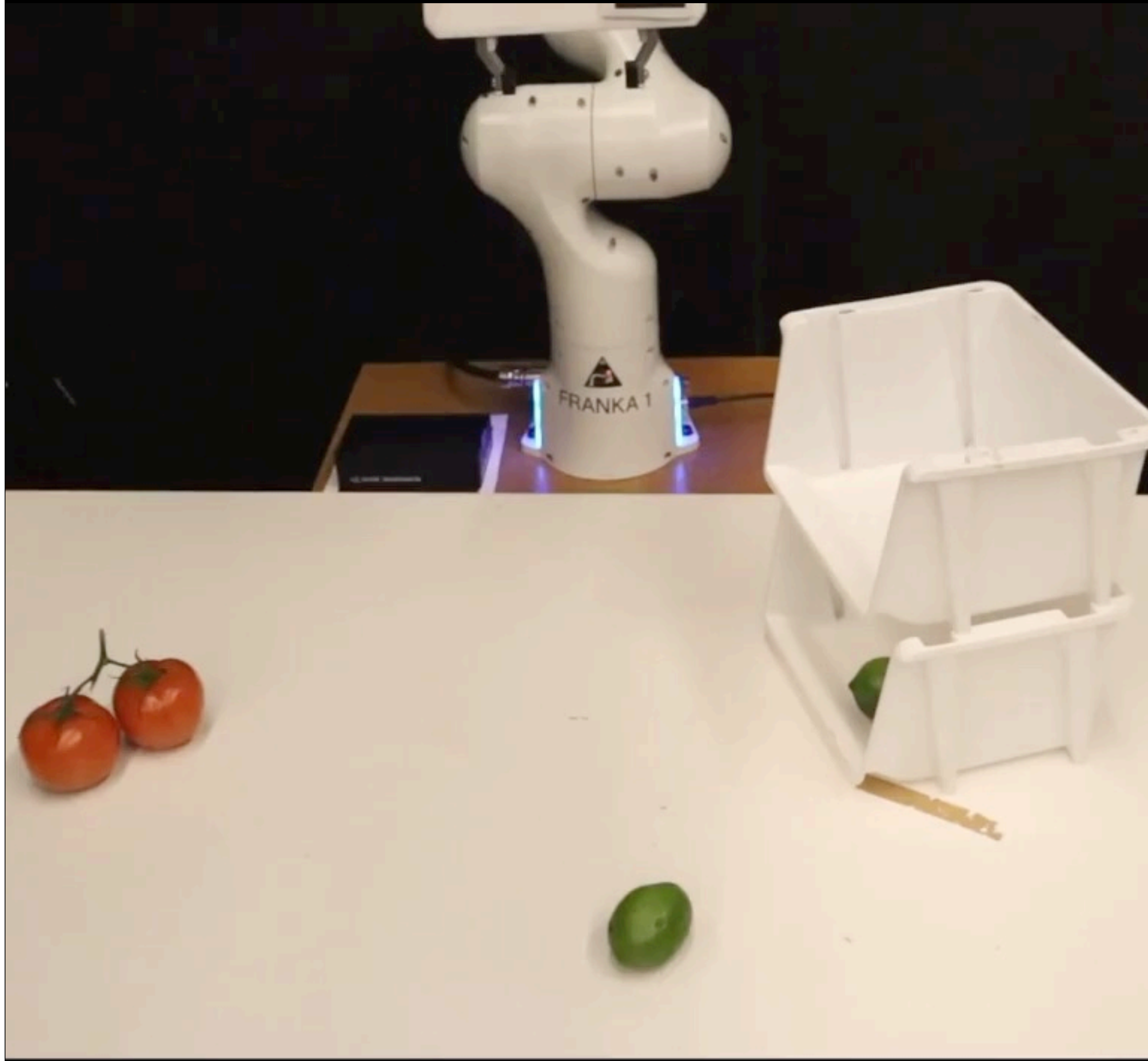
Perceiver-Actor

Multi-task 6-DoF manipulation agent

End-to-end few-shot imitation learning

Input: RGB-D Voxels & Language Goal

Output: Discretized 6-DoF action + open/close



“put the tomatoes in the top bin”

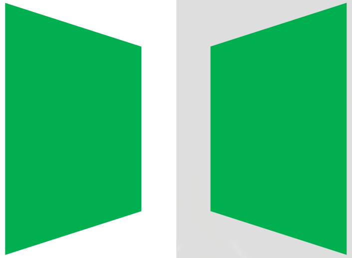
Perceiver-Actor

Multi-task 6-DoF manipulation agent

End-to-end few-shot imitation learning

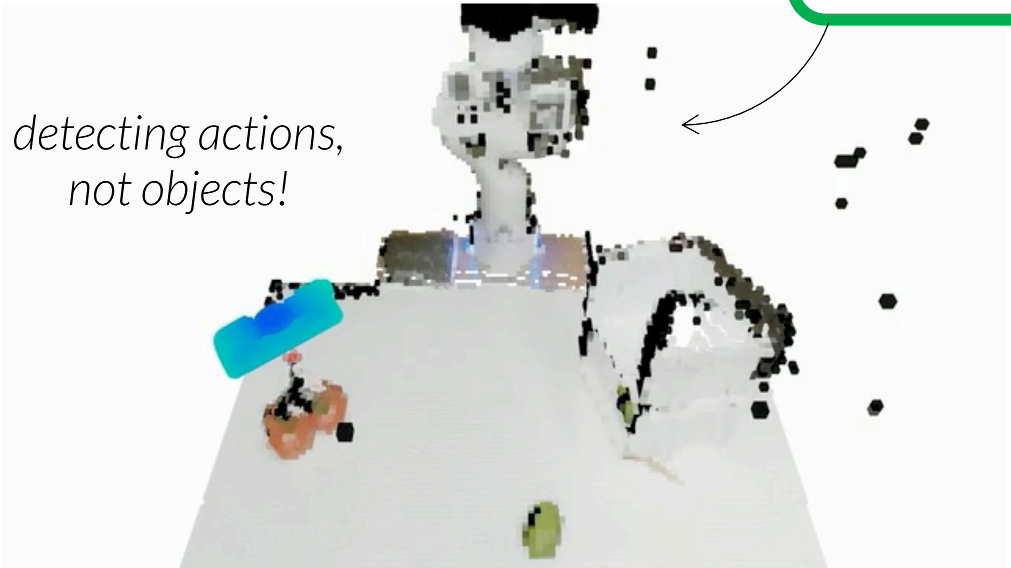
Input: RGB-D Voxels & Language Goal

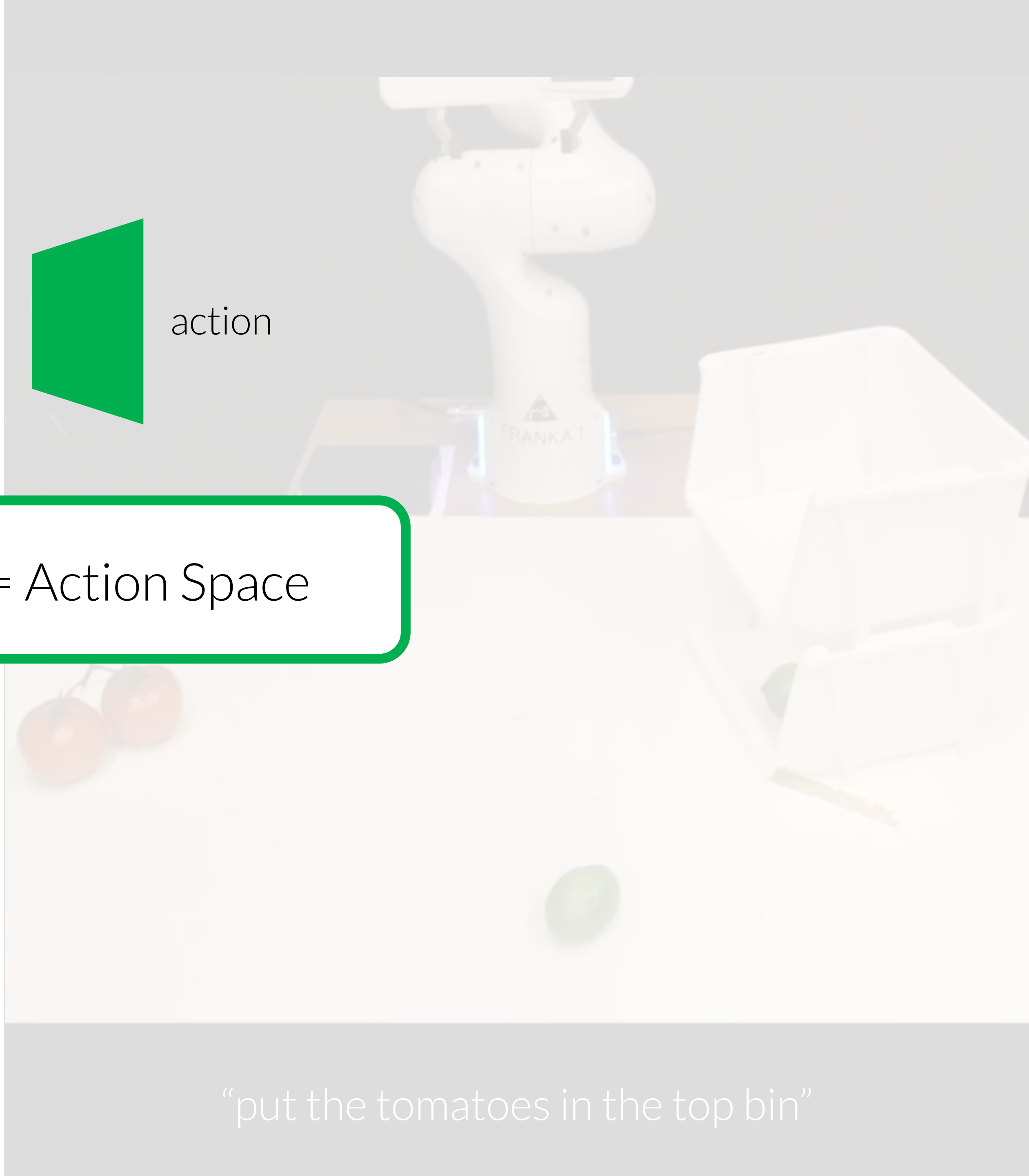
Output: Discretized 6-DoF action + c



Obs Space = Action Space

detecting actions, not objects!





Perceiver-Actor

Multi-task 6-DoF manipulation agent

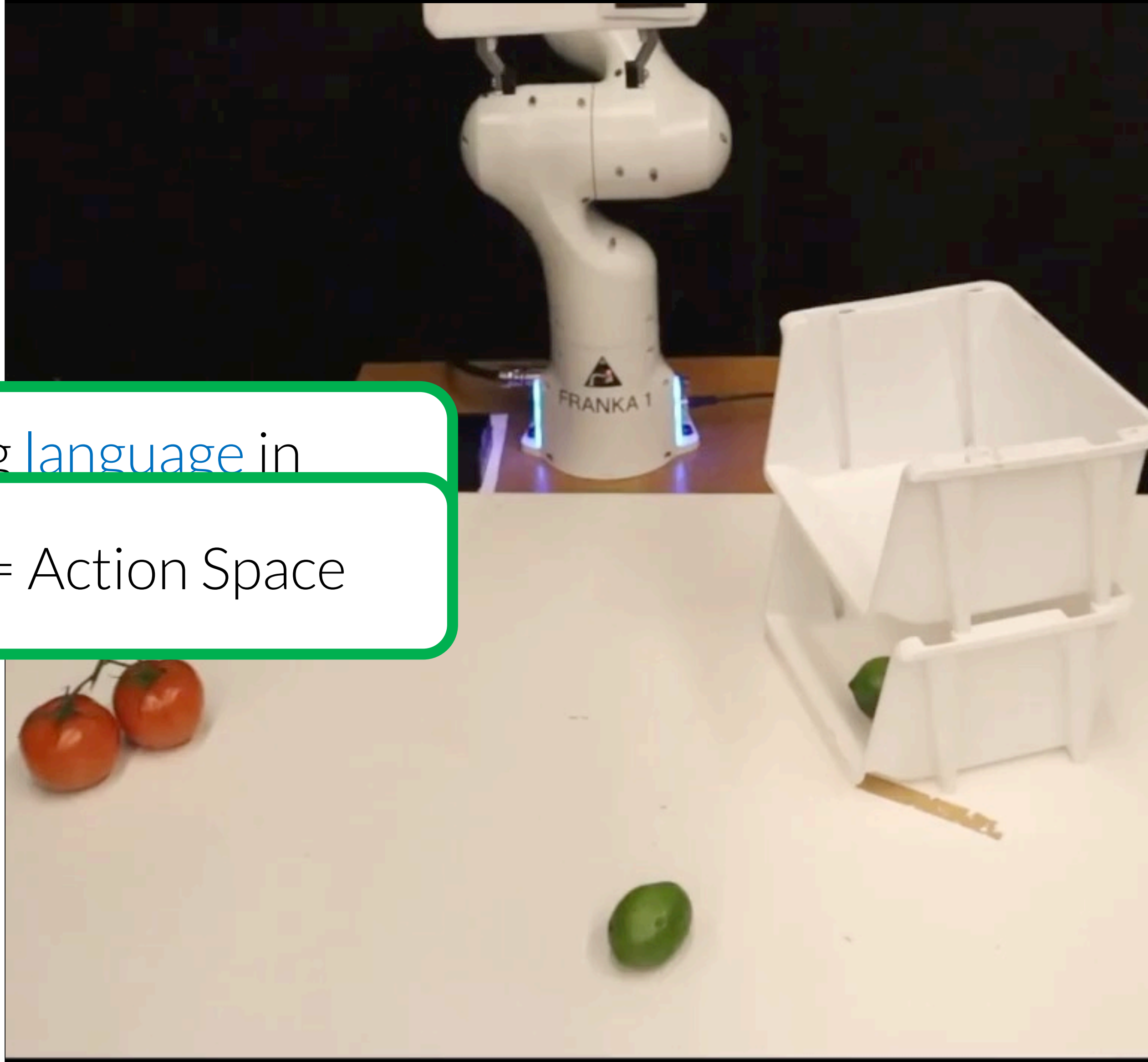
End-to-end few-shot imitation learning

Input: RGB-D Voxels & Language Goa

Output: Discretized 6-DoF action + c

Grounding language in

Obs Space = Action Space

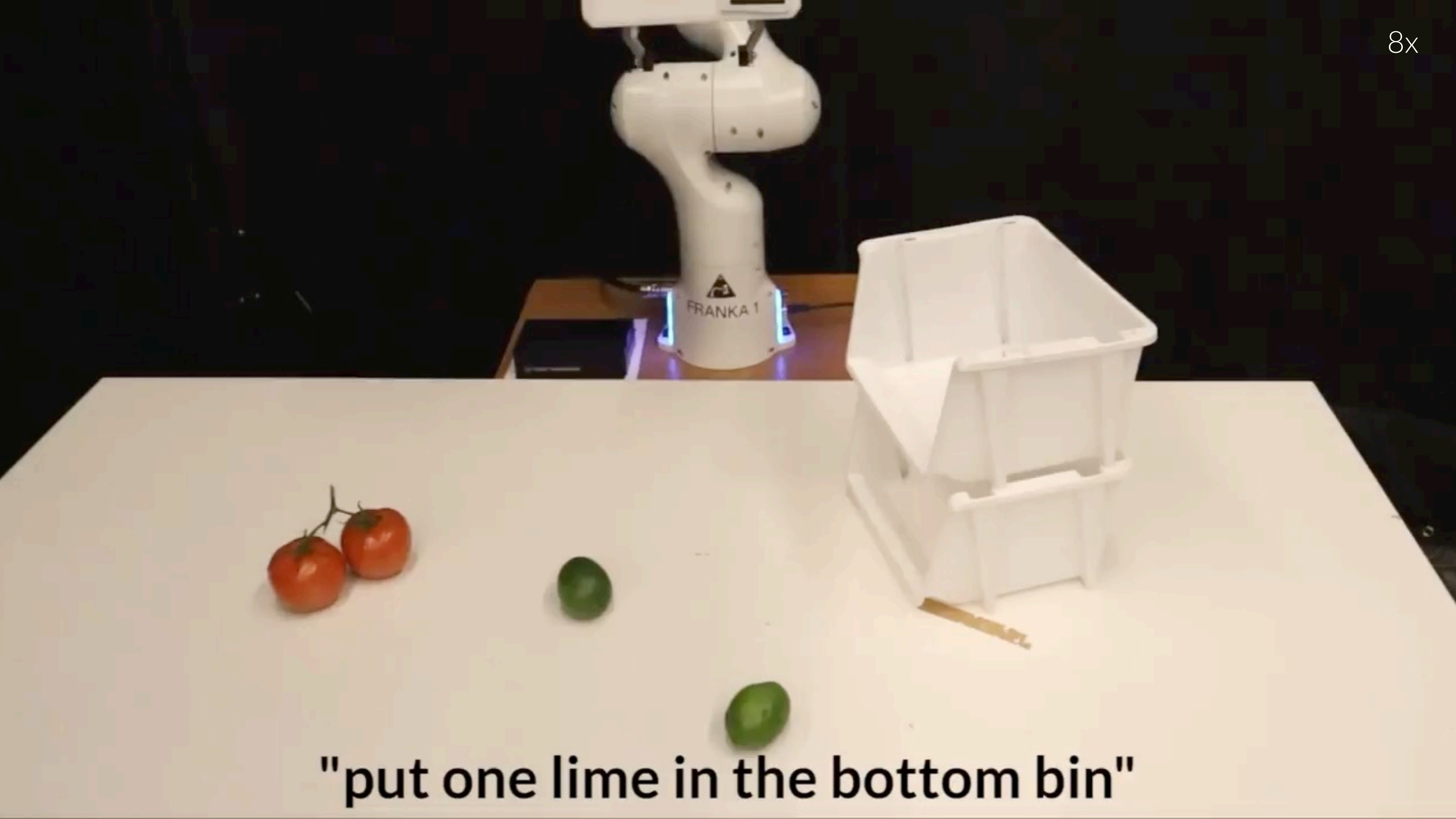


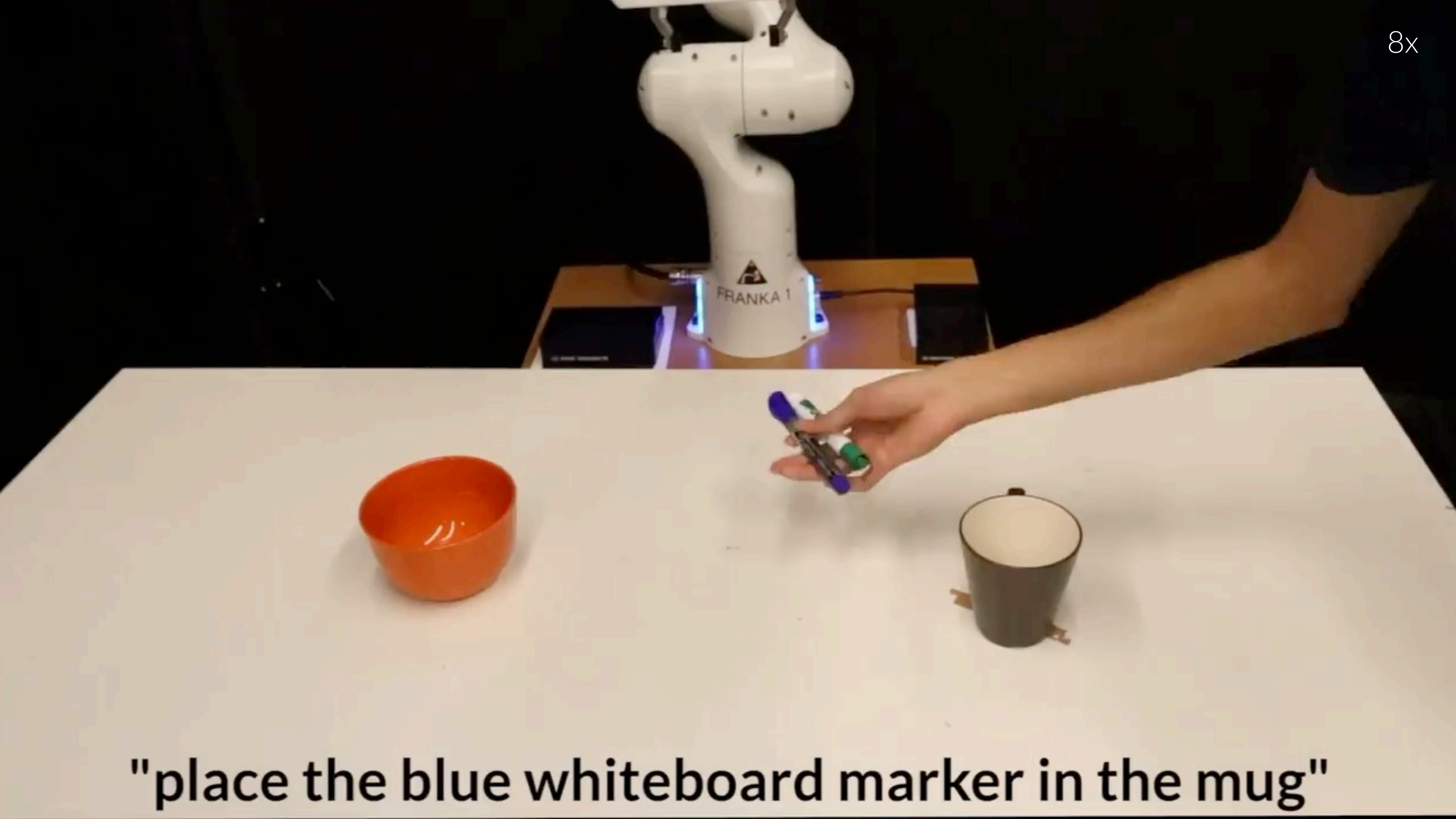
“put the tomatoes in the top bin”

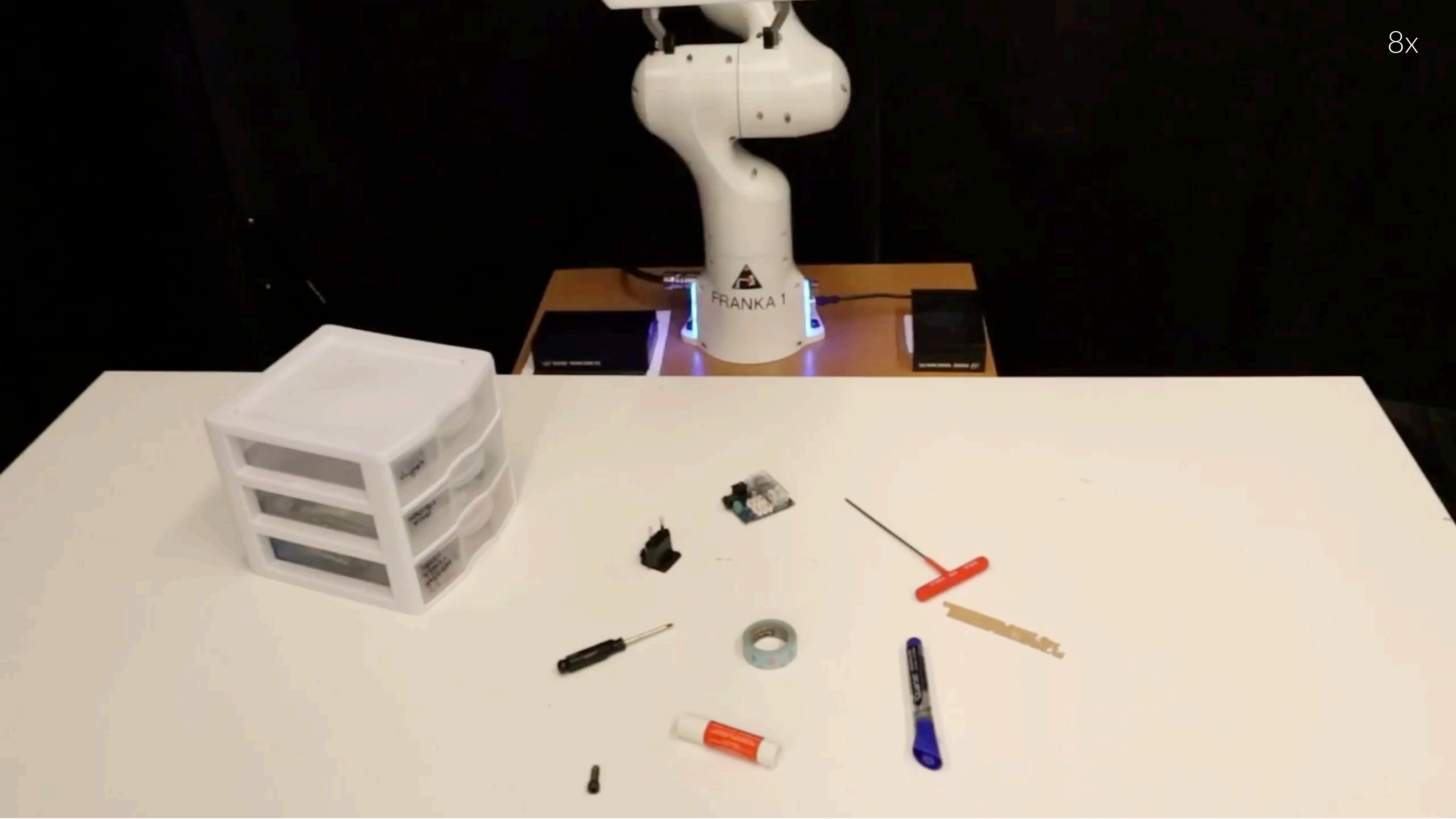
These results are from a one multi-task Transformer

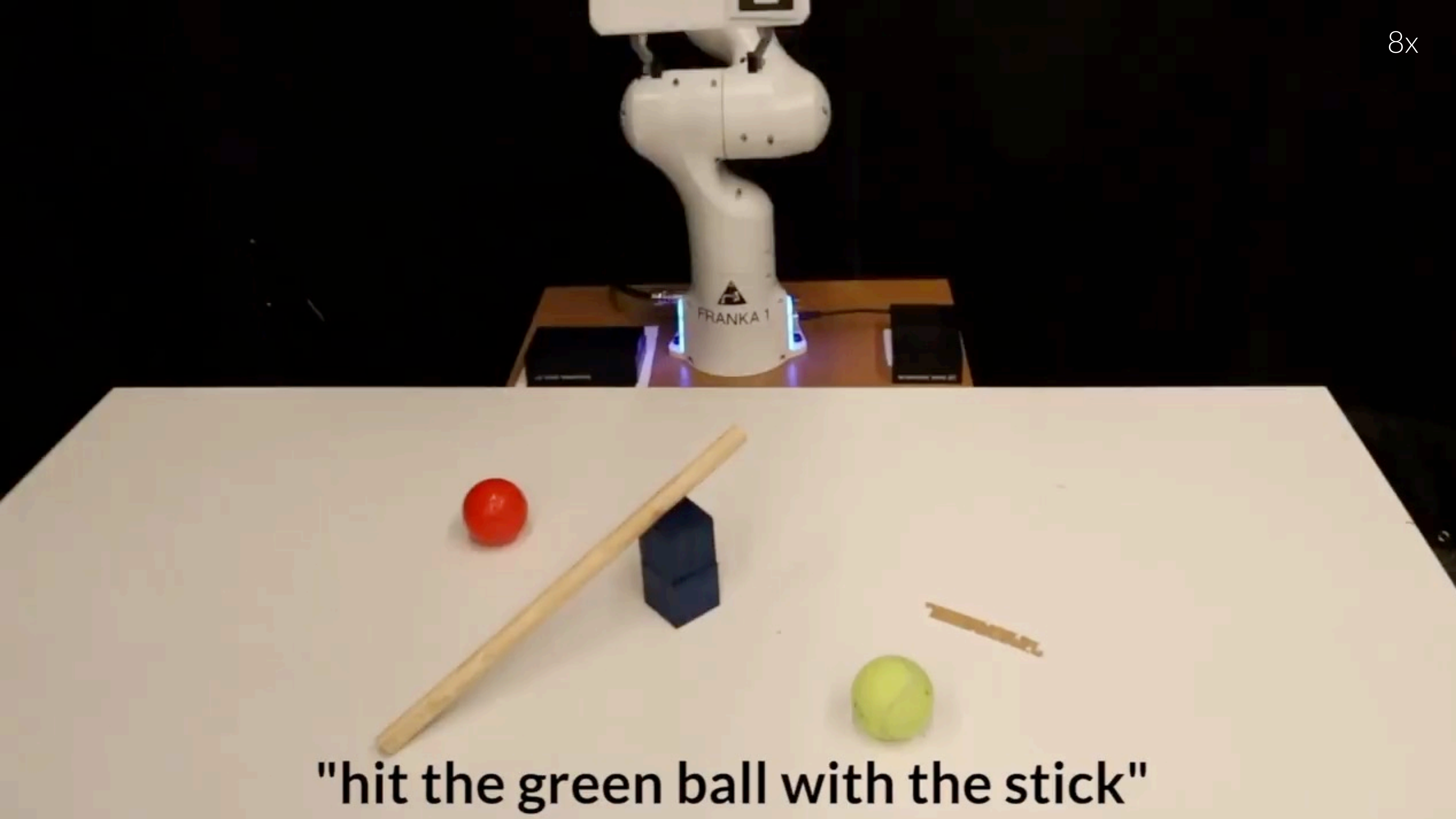
trained *from scratch* with just **53 demos**

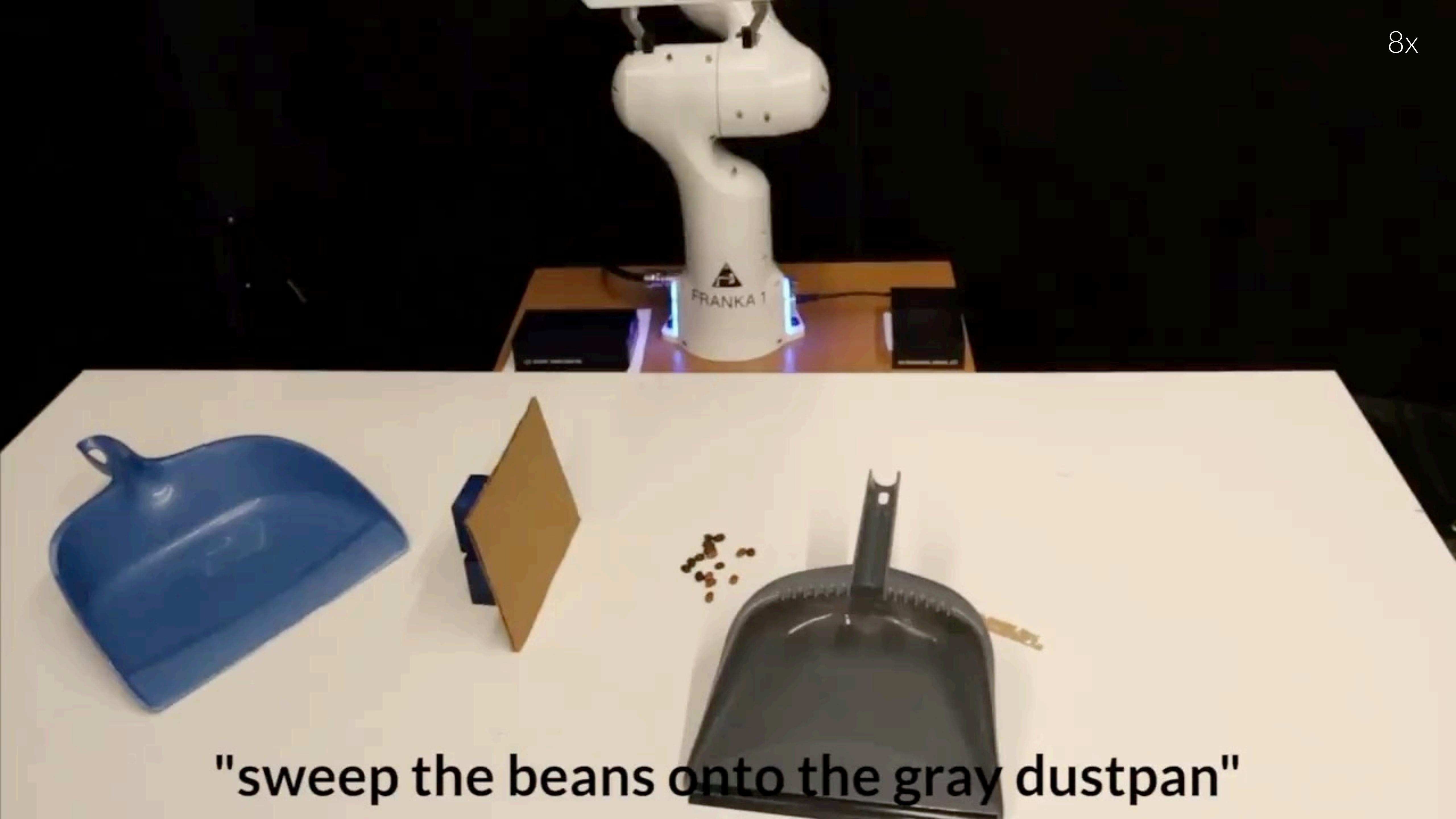

8x
FRANKA 1
"press the hand san"


8x
FRANKA 1
"put one lime in the bottom bin"

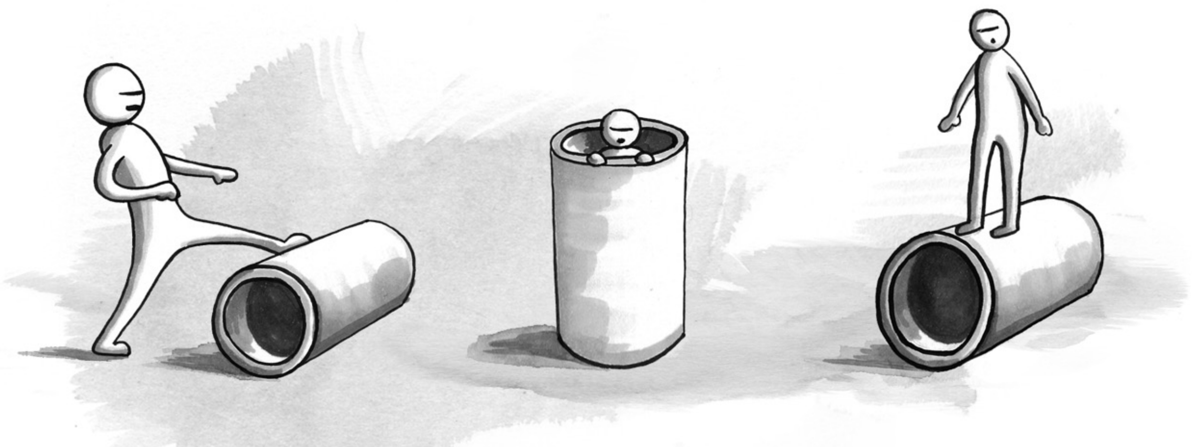

8x
FRANKA 1
"place the blue whiteboard marker in the mug"


8x
FRANKA 1

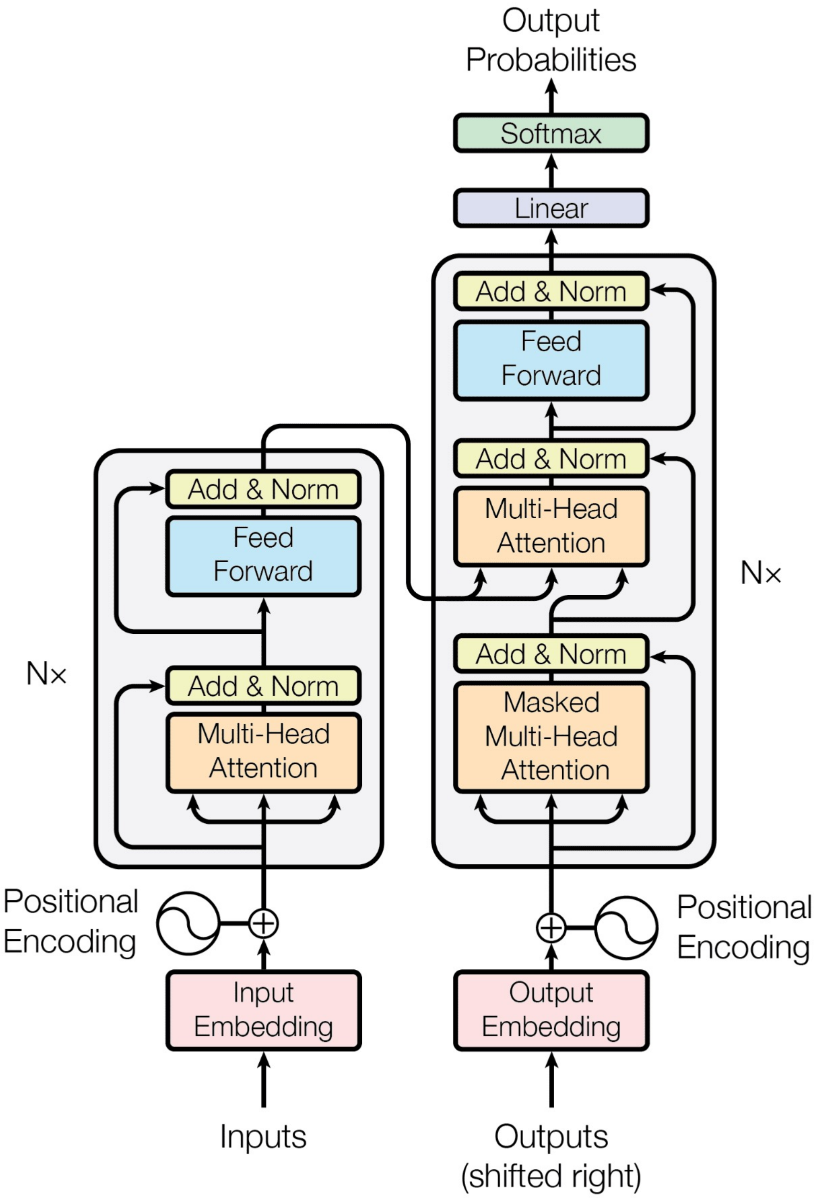

8x
FRANKA 1
"hit the green ball with the stick"


8x
FRANKA 1
"sweep the beans onto the gray dustpan"

How does it work?

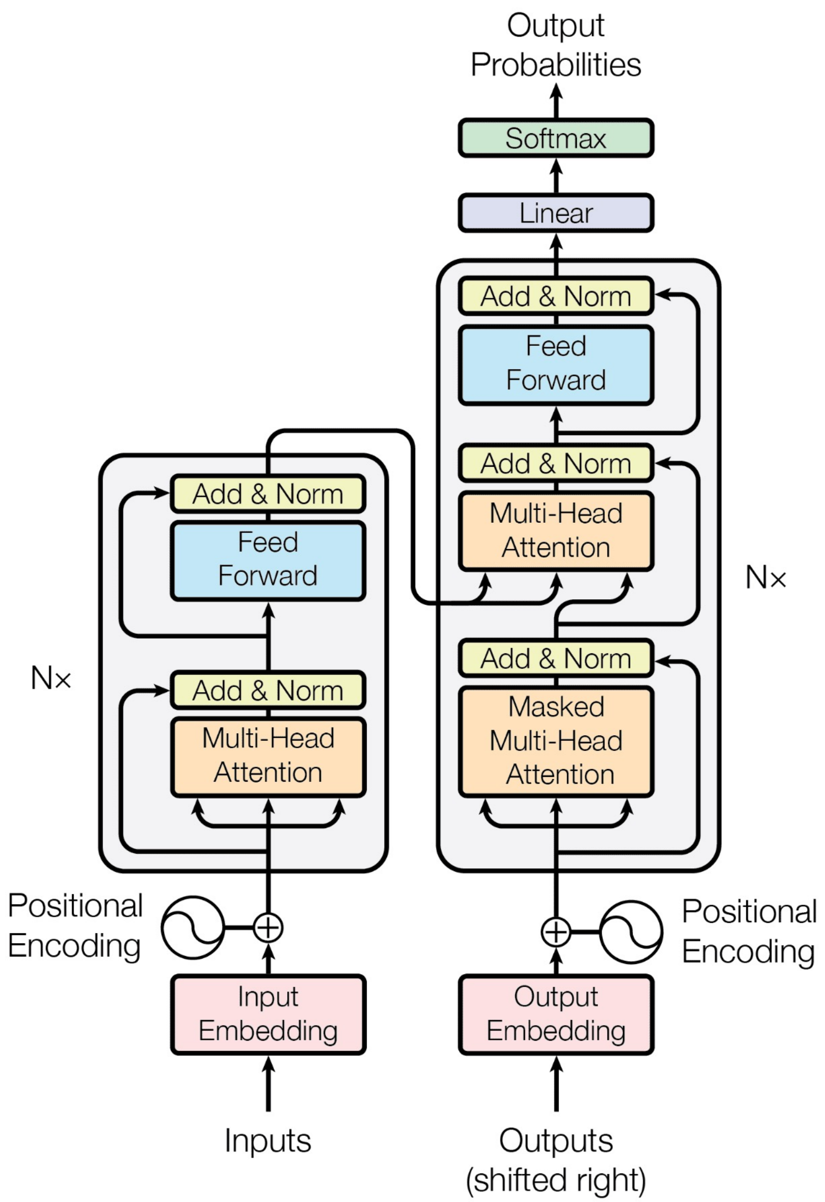


Credit: https://uxdesign.cc/



Transformer

Vaswani et al, 2017



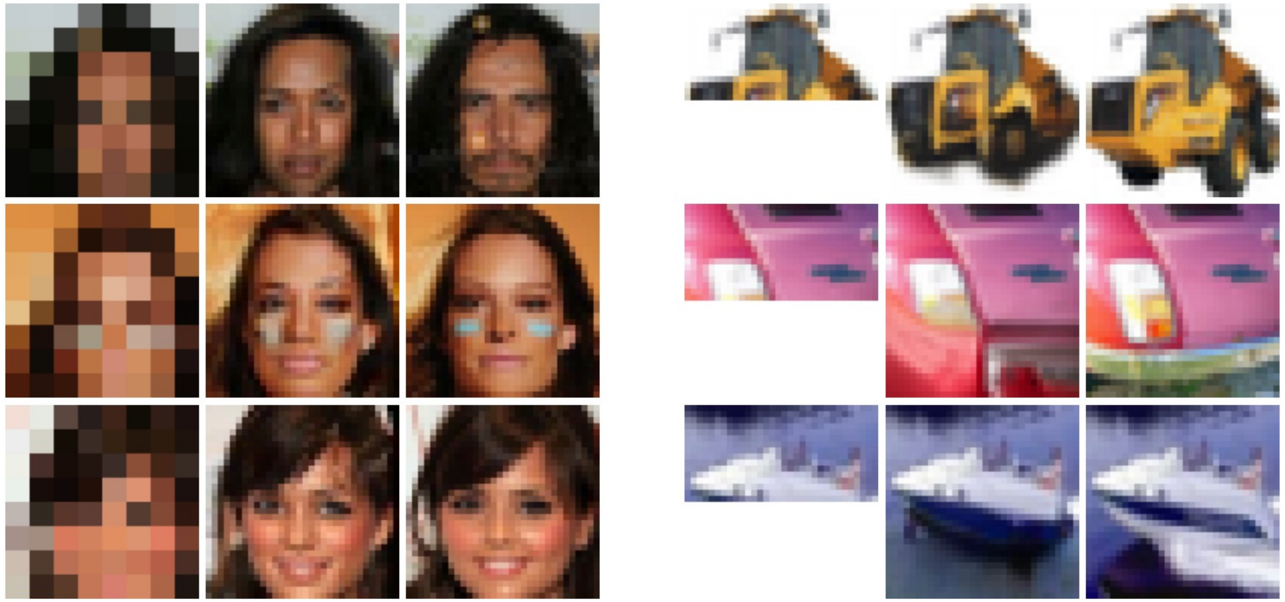
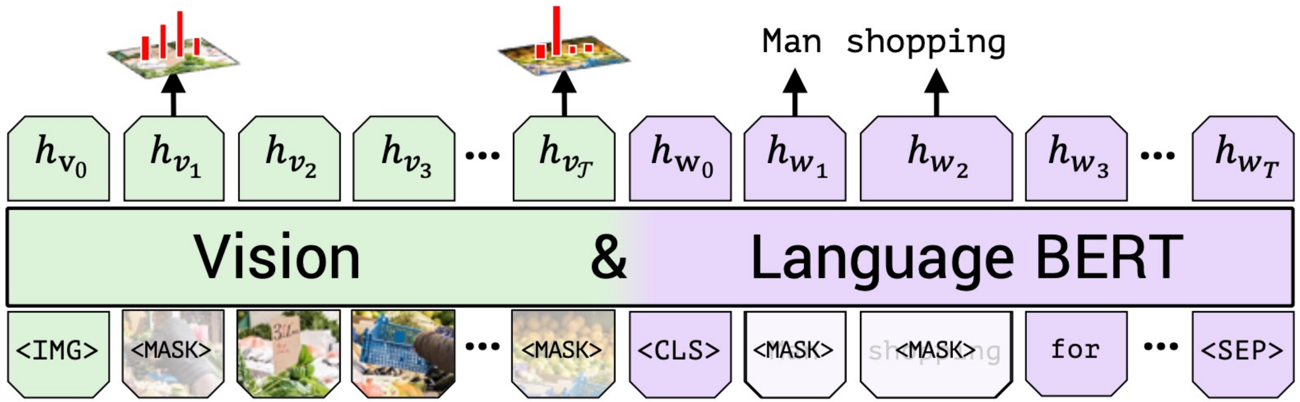


Image Transformer
Parmar et al, 2018

Transformer
Vaswani et al, 2017

Problem:

vs. word



Vision-Lang BERTs
Lu et al, 2019
Tan et al, 2019

Image Transformer
Parmar et al, 2018

Transformer
Vaswani et al, 2017

Problem:

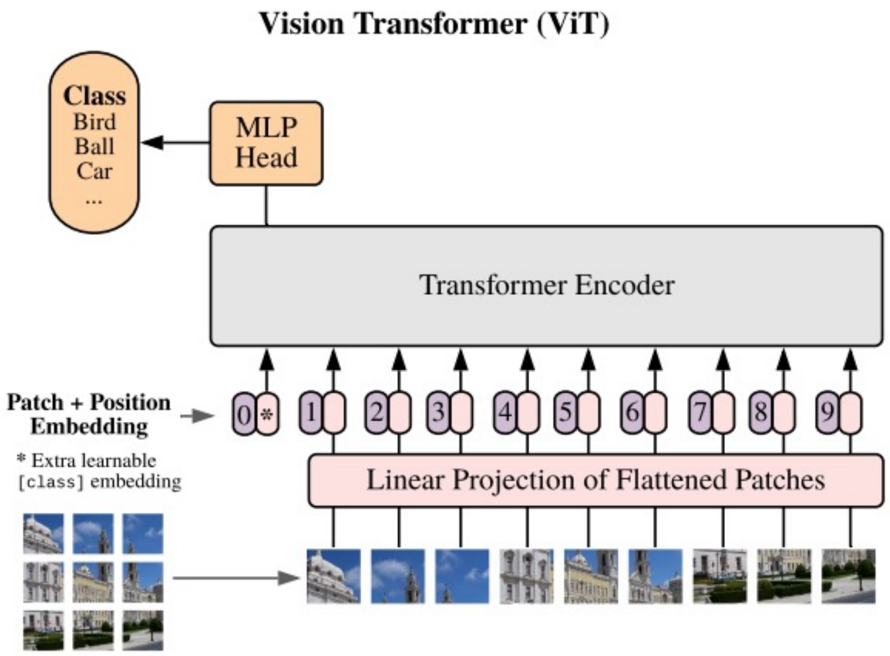
Object Detections

ViT
Dosovitskiy et al, 2020

Vision-Lang BERTs
Lu et al, 2019
Tan et al, 2019

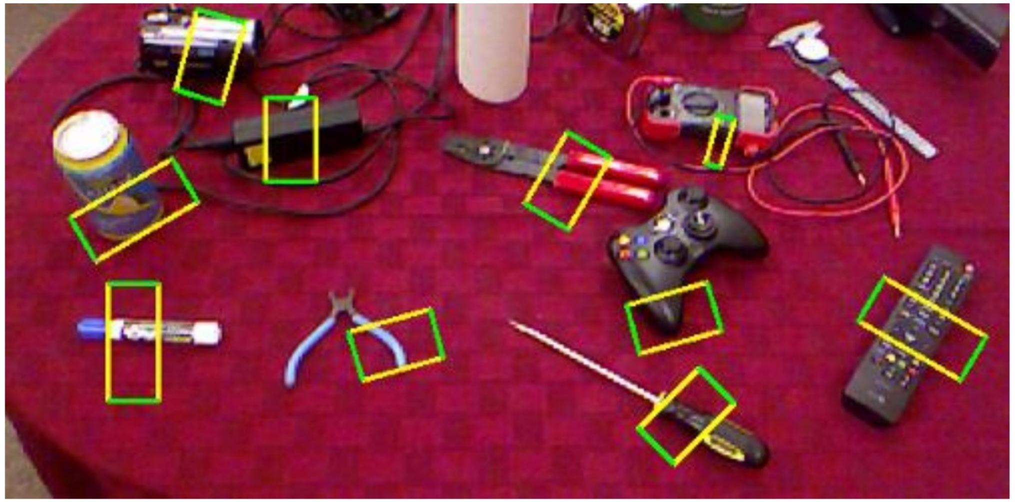
Image Transformer
Parmar et al, 2018

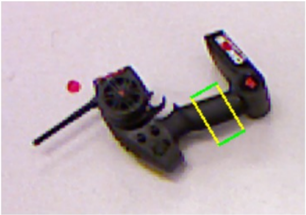
Transformer
Vaswani et al, 2017

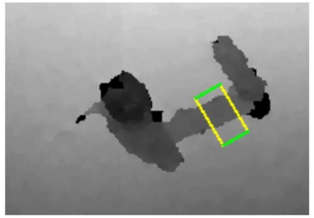


Solution:

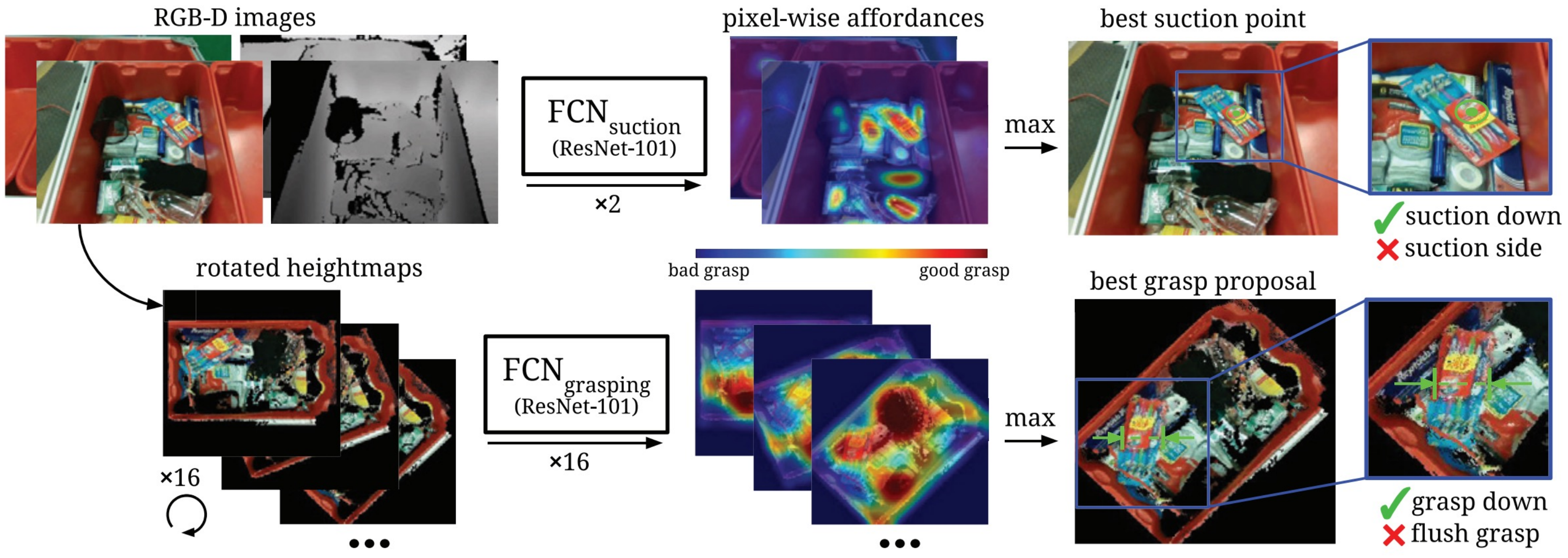
2D patches!

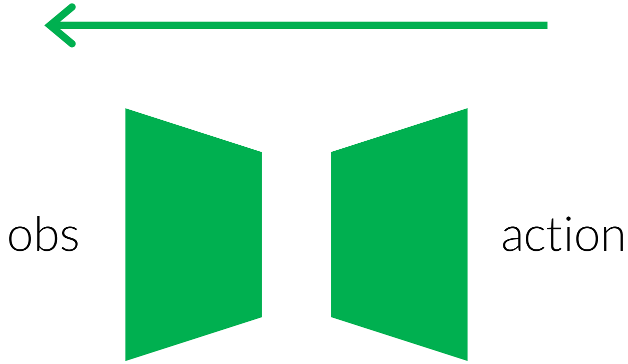






Deep Grasping
Lenz et al, 2014





Visual Affordances
Zeng et al, 2017
Zeng et al, 2019

Deep Grasping
Lenz et al, 2014

Visual Affordances
Zeng et al, 2017
Zeng et al, 2019

Deep Grasping
Lenz et al, 2014

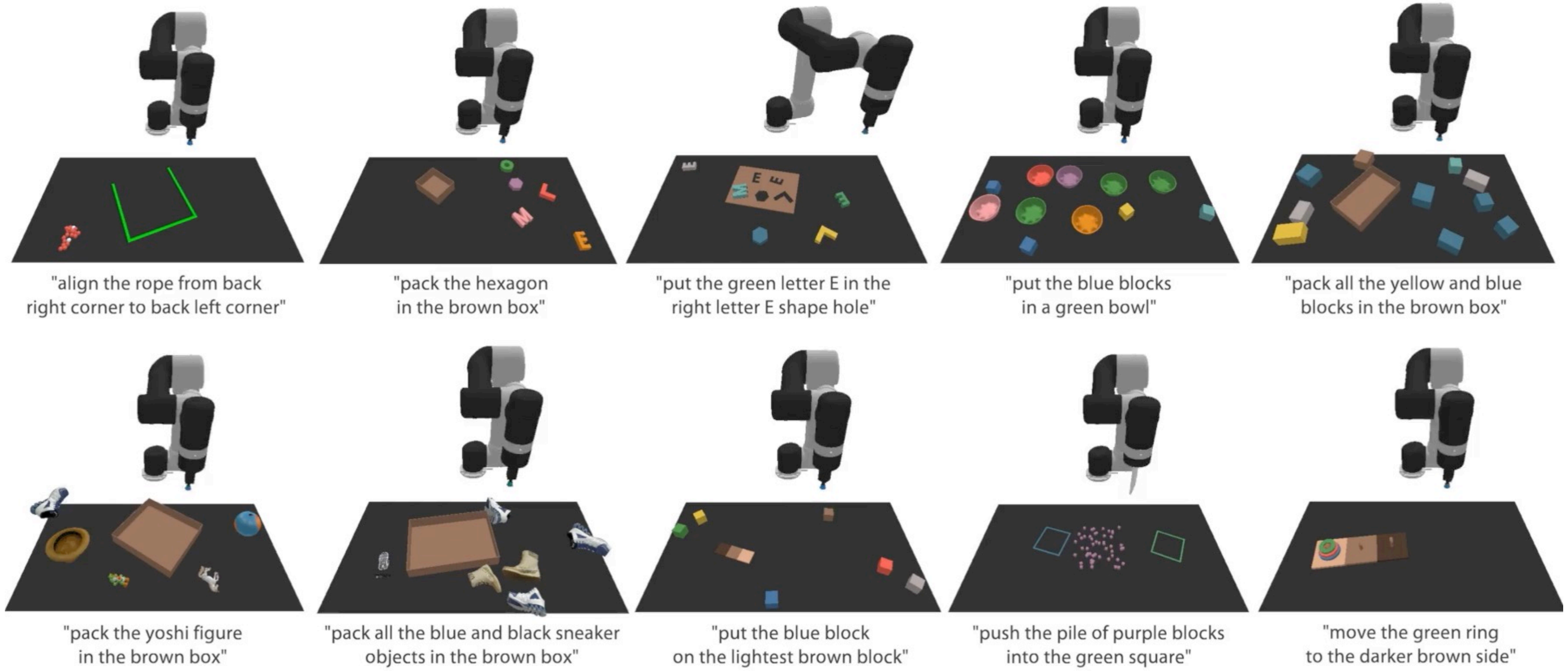
(a) (b) (e) (f)

Problem:

Missing Natural Interface
for Goal-Conditioning

Visual Affordances
Zeng et al, 2017
Zeng et al, 2019

Deep Grasping
Lenz et al, 2014



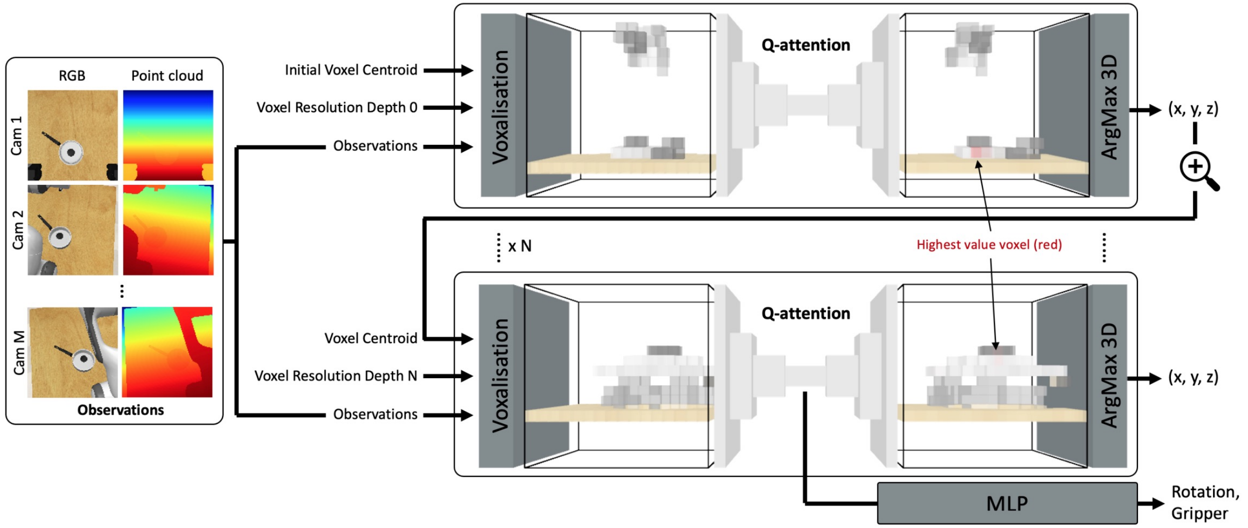
CLIPort
Shridhar et al, 2021

Problem:

2D Obs & Action Space

Visual Affordances
Zeng et al, 2017
Zeng et al, 2019

Deep Grasping
Lenz et al, 2014



Problem:

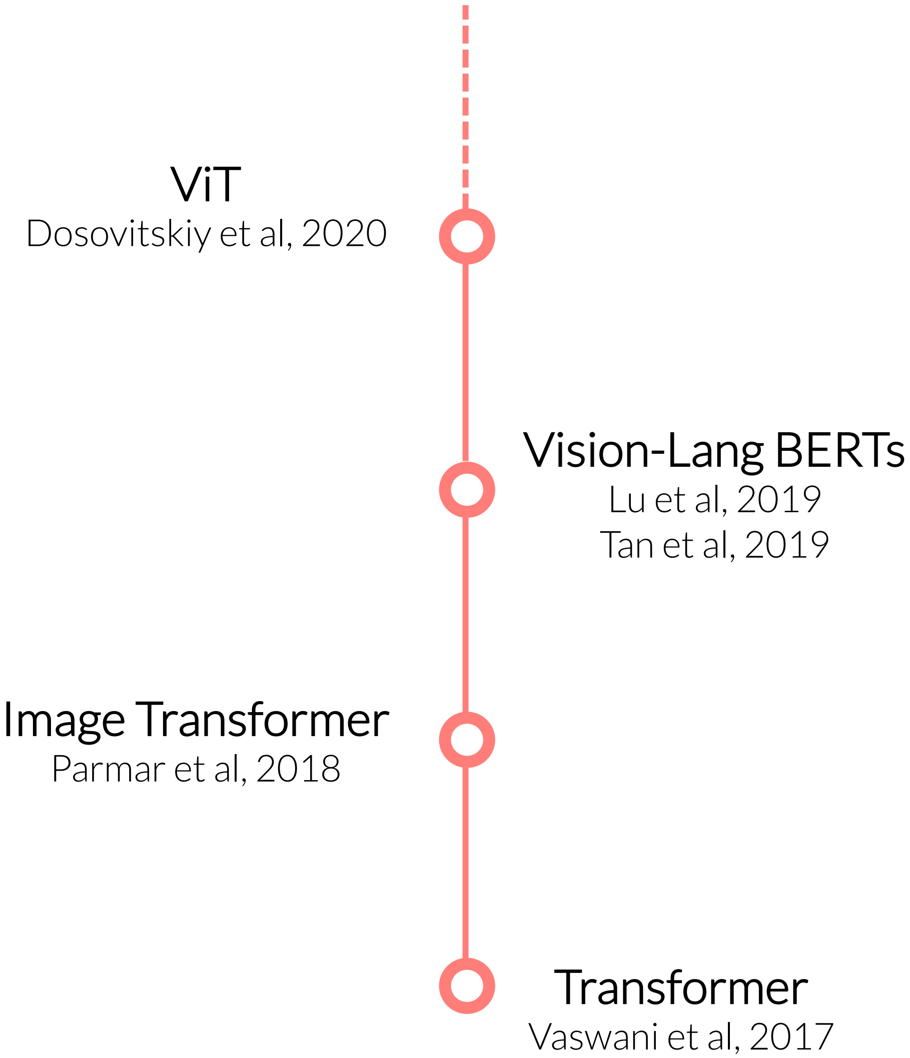
High Dimensional Input

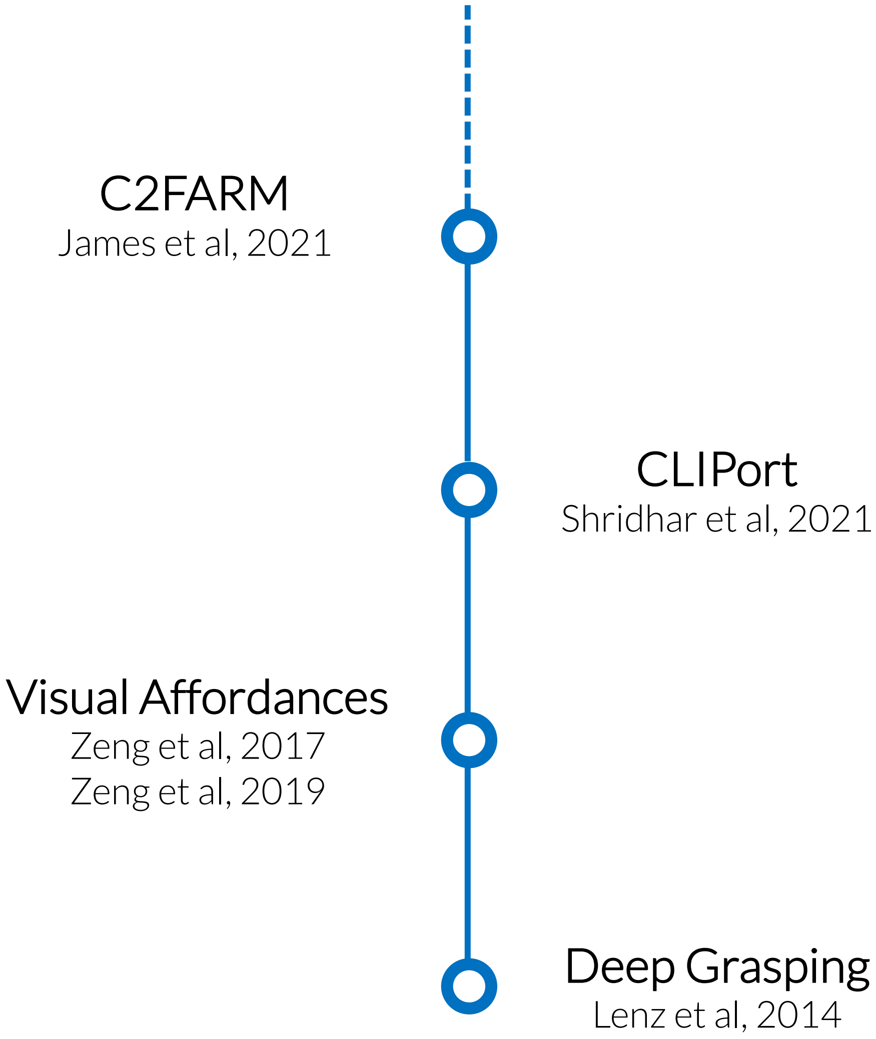
C2FARM
James et al, 2021

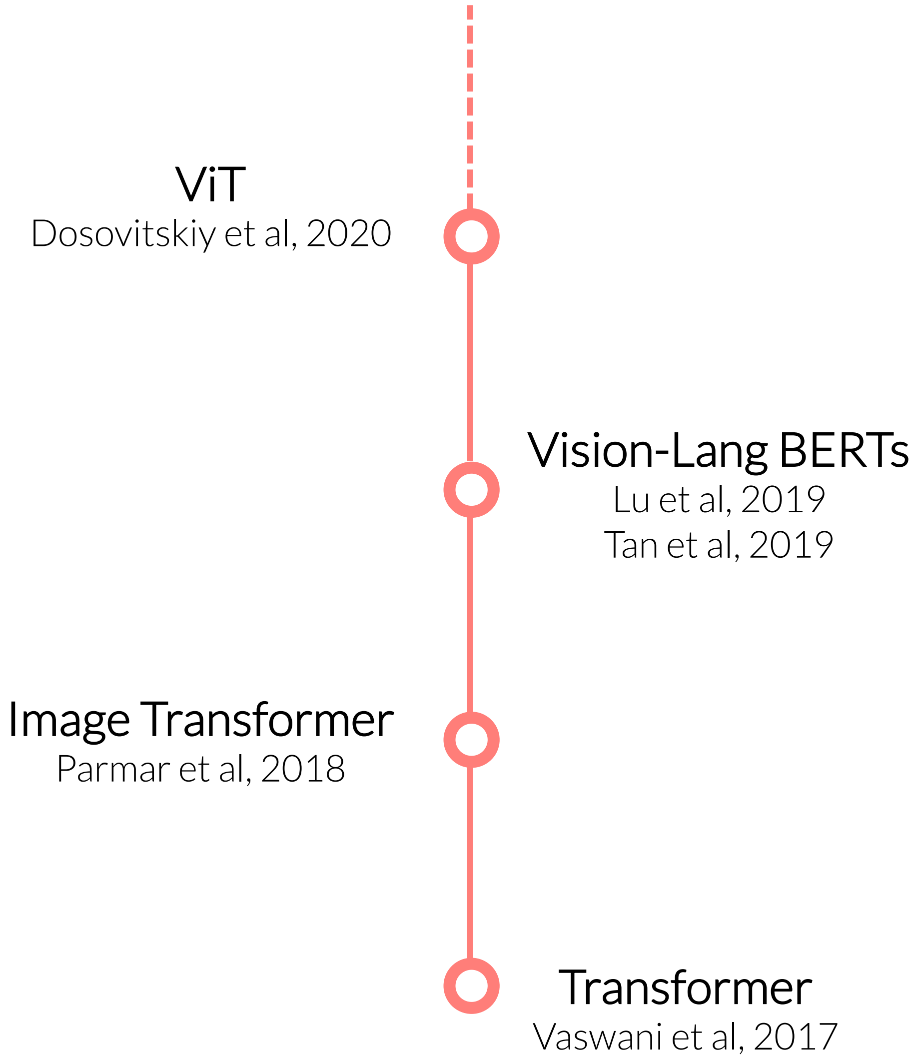
CLIPort
Shridhar et al, 2021

Visual Affordances
Zeng et al, 2017
Zeng et al, 2019

Deep Grasping
Lenz et al, 2014


ViT
Dosovitskiy et al, 2020
Vision-Lang BERTs
Lu et al, 2019
Tan et al, 2019
Image Transformer
Parmar et al, 2018
Transformer
Vaswani et al, 2017


C2FARM
James et al, 2021
CLIPort
Shridhar et al, 2021
Visual Affordances
Zeng et al, 2017
Zeng et al, 2019
Deep Grasping
Lenz et al, 2014



What are the right **tokens** for manipulation?

3D Voxel Patches

Problems with
2D Static Monocular RGB

hand-eye coordination

depth cues

camera perturbations

distractors

spatial data augmentation?

What are the right **tokens** for manipulation?

3D Voxel Patches

How to deal with
high dimensional input?

Latent-Space Transformer

C2FARM
James et al, 2021

CLIPort
Shridhar et al, 2021

Visual Affordances
Zeng et al, 2017
Zeng et al, 2019

Deep Grasping
Lenz et al, 2014

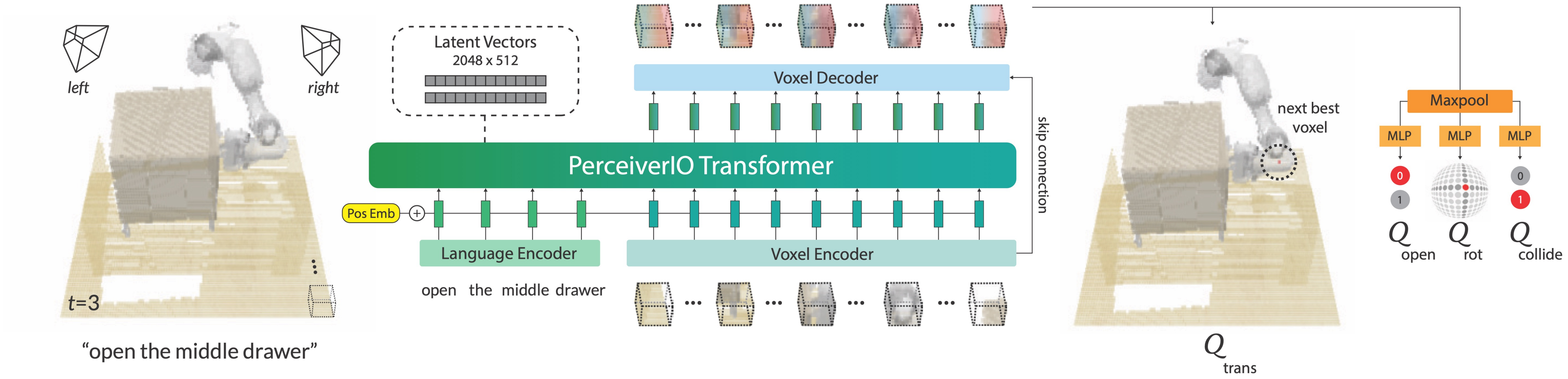
How to deal with
high dimensional input?

Latent-Space Transformer

What are the right
tokens for manipulation?

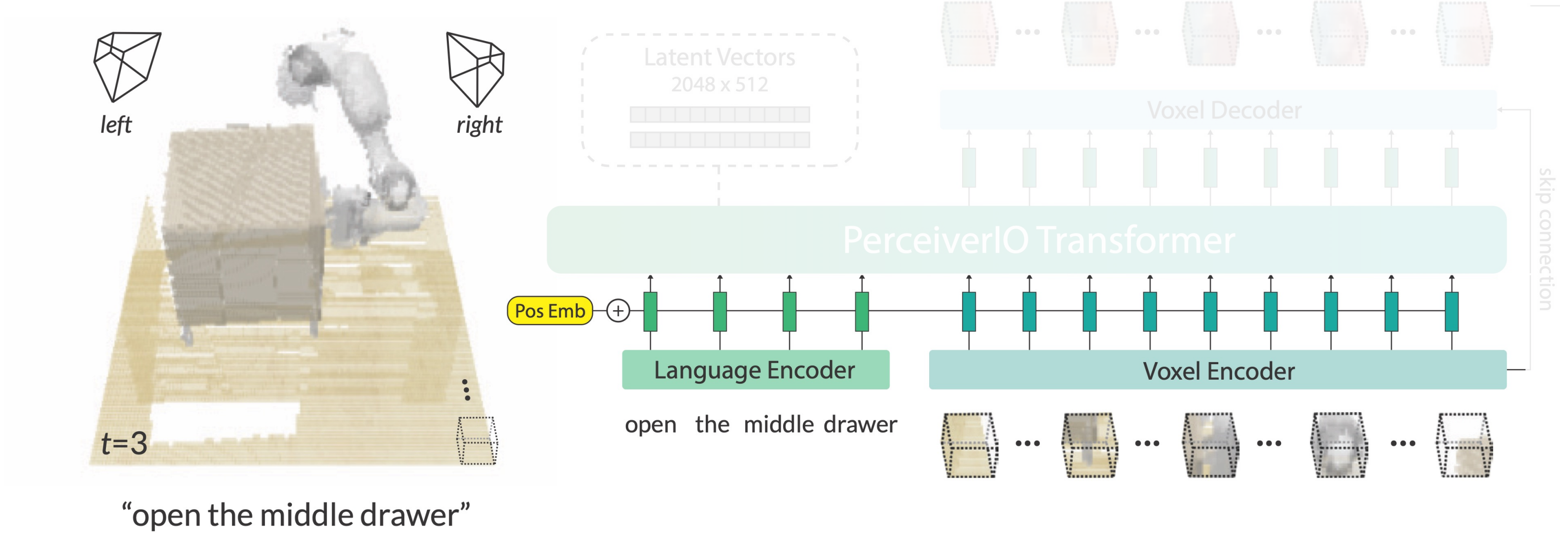
3D Voxel Patches

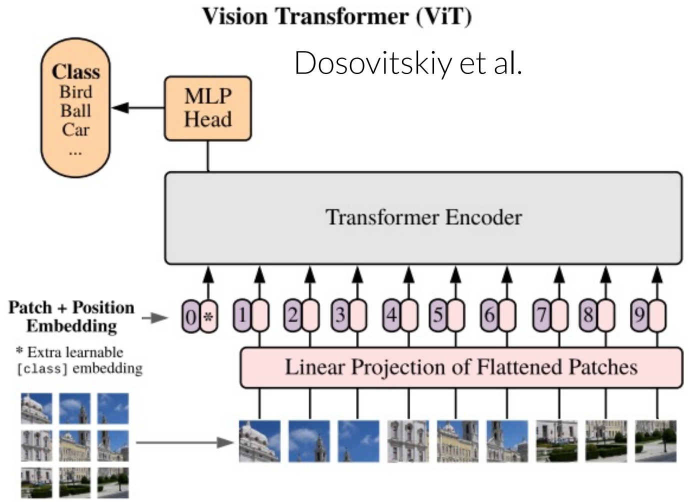
Perceiver-Actor



"open the middle drawer"

Perceiver-Actor

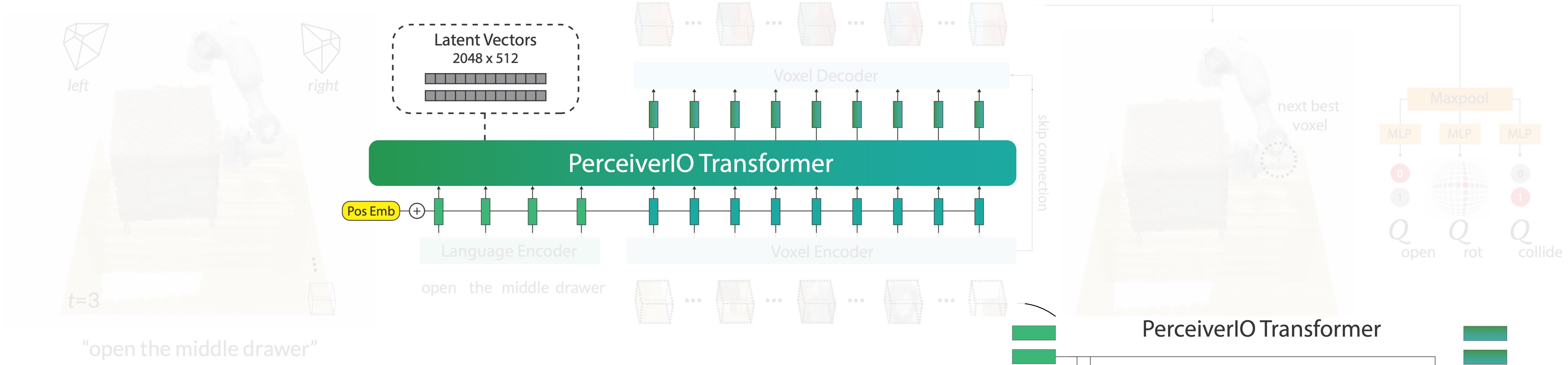




Dosovitskiy et al.

ViT : 2D image patches
PerAct : 3D voxel patches

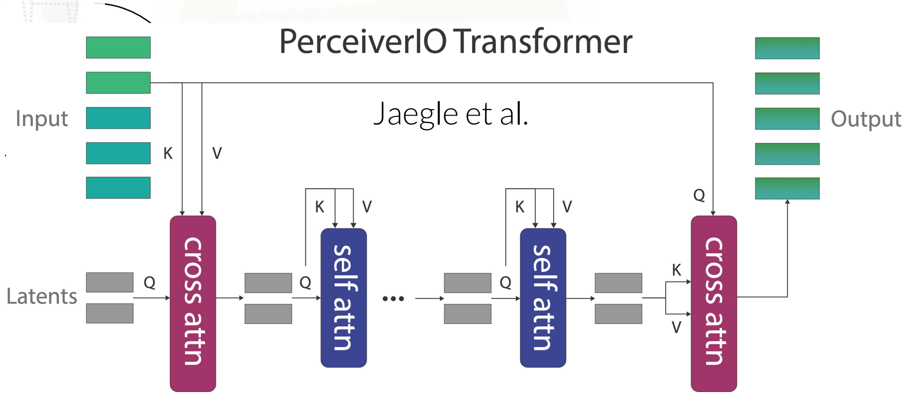
Perceiver-Actor



5^3 patches wi

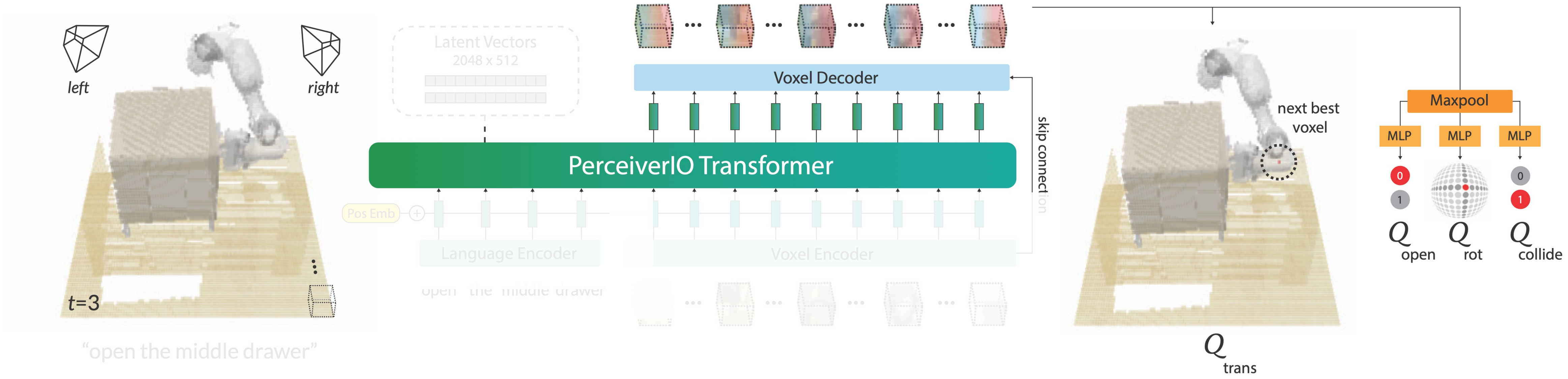
2048 x 512

randomly initialized
and trained end-to-end

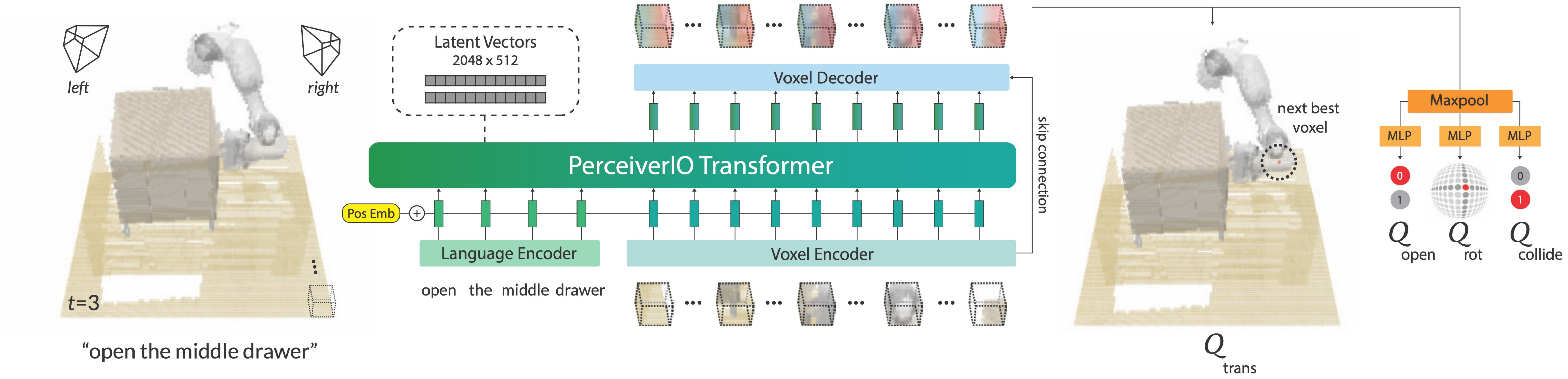


Perceiver-Actor

100 x 100 x 100 x 64 features

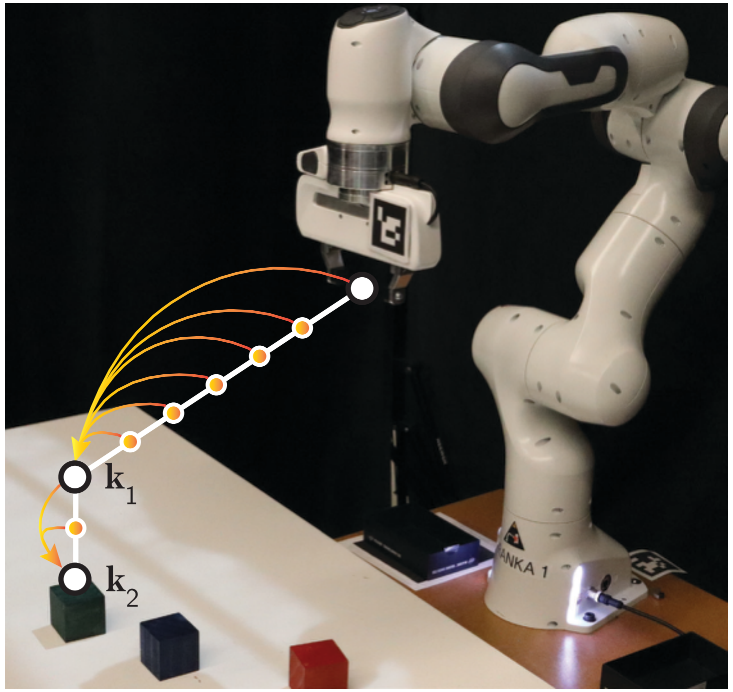


Perceiver-Actor



"open the middle drawer"

Dataset Setup



Predict the "next best keyframe action" classification task

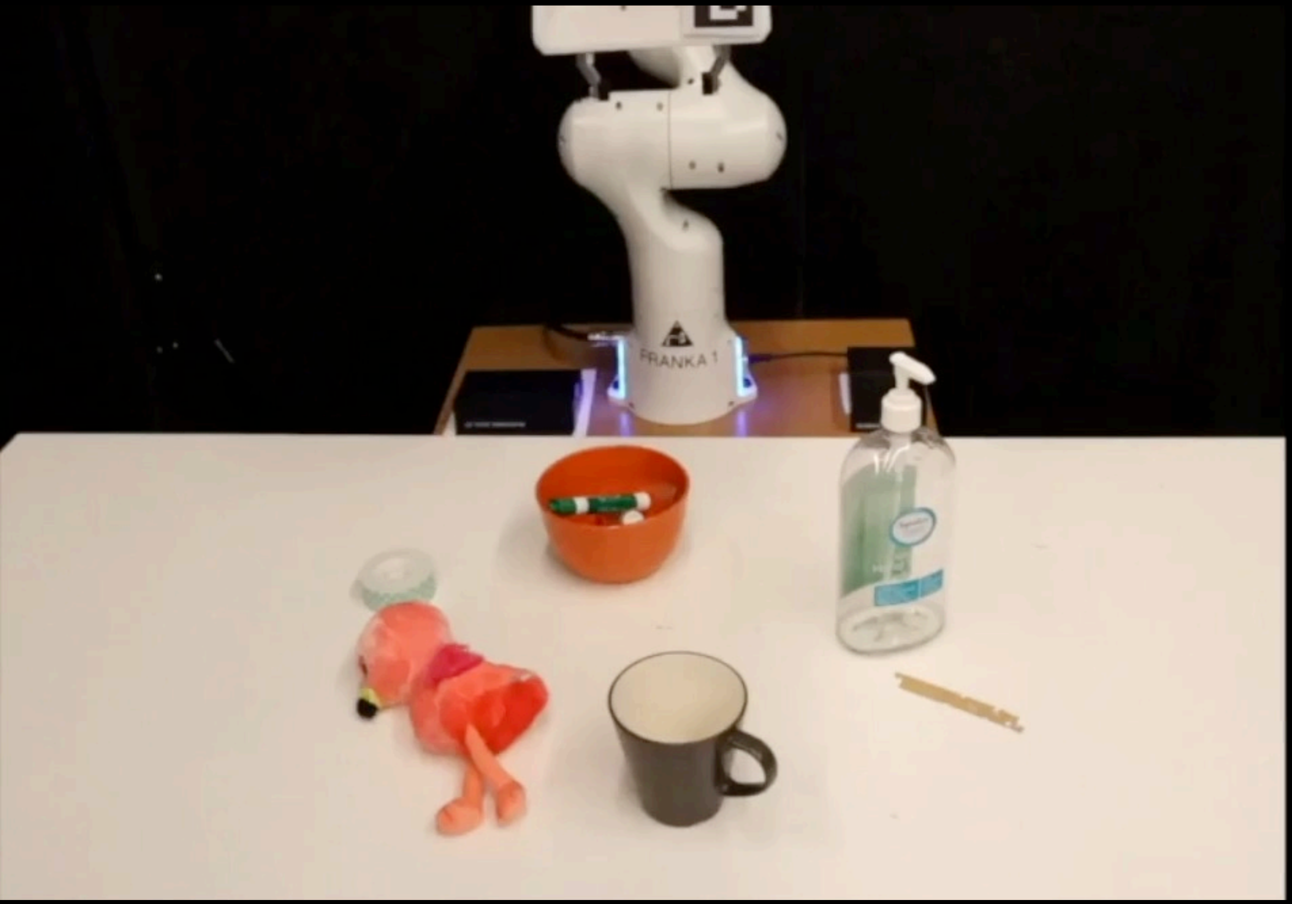
James et al

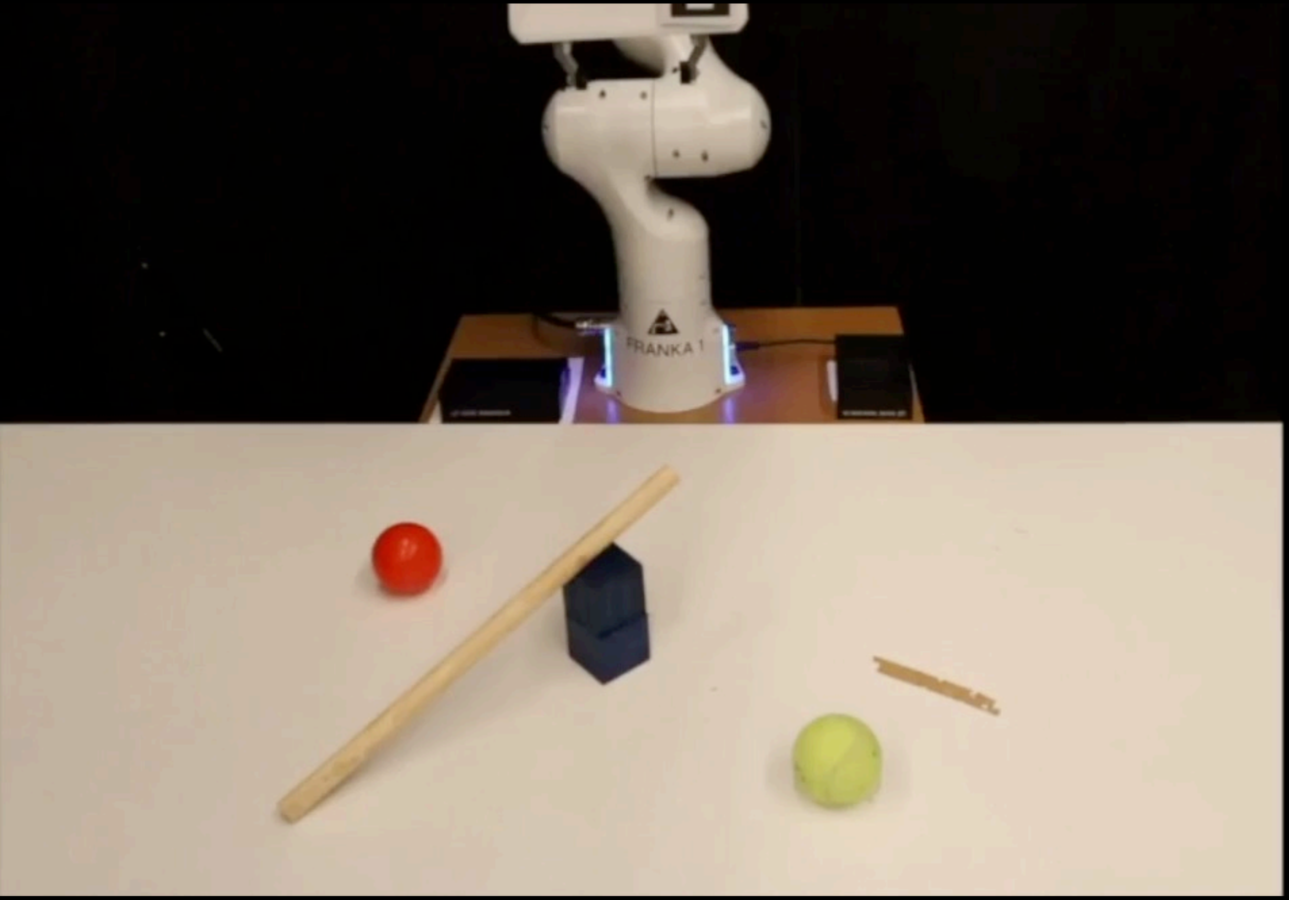
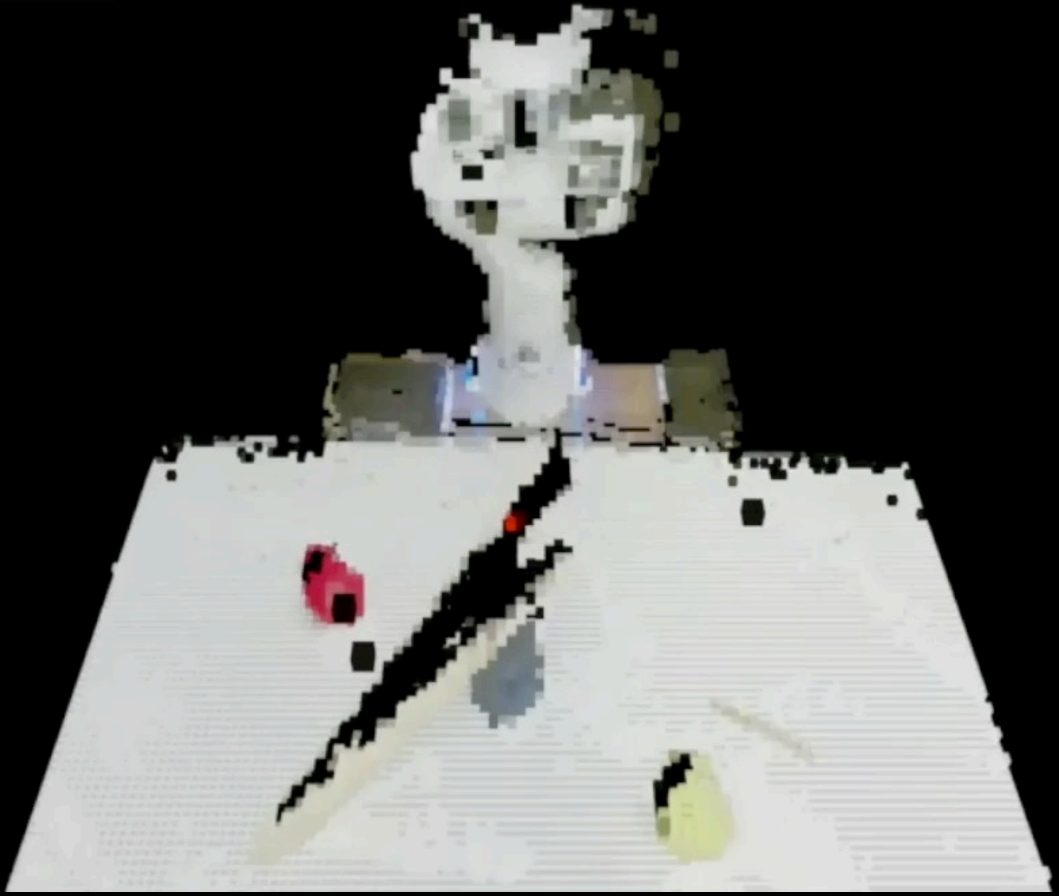
Heuristic for Keyframe Extraction:

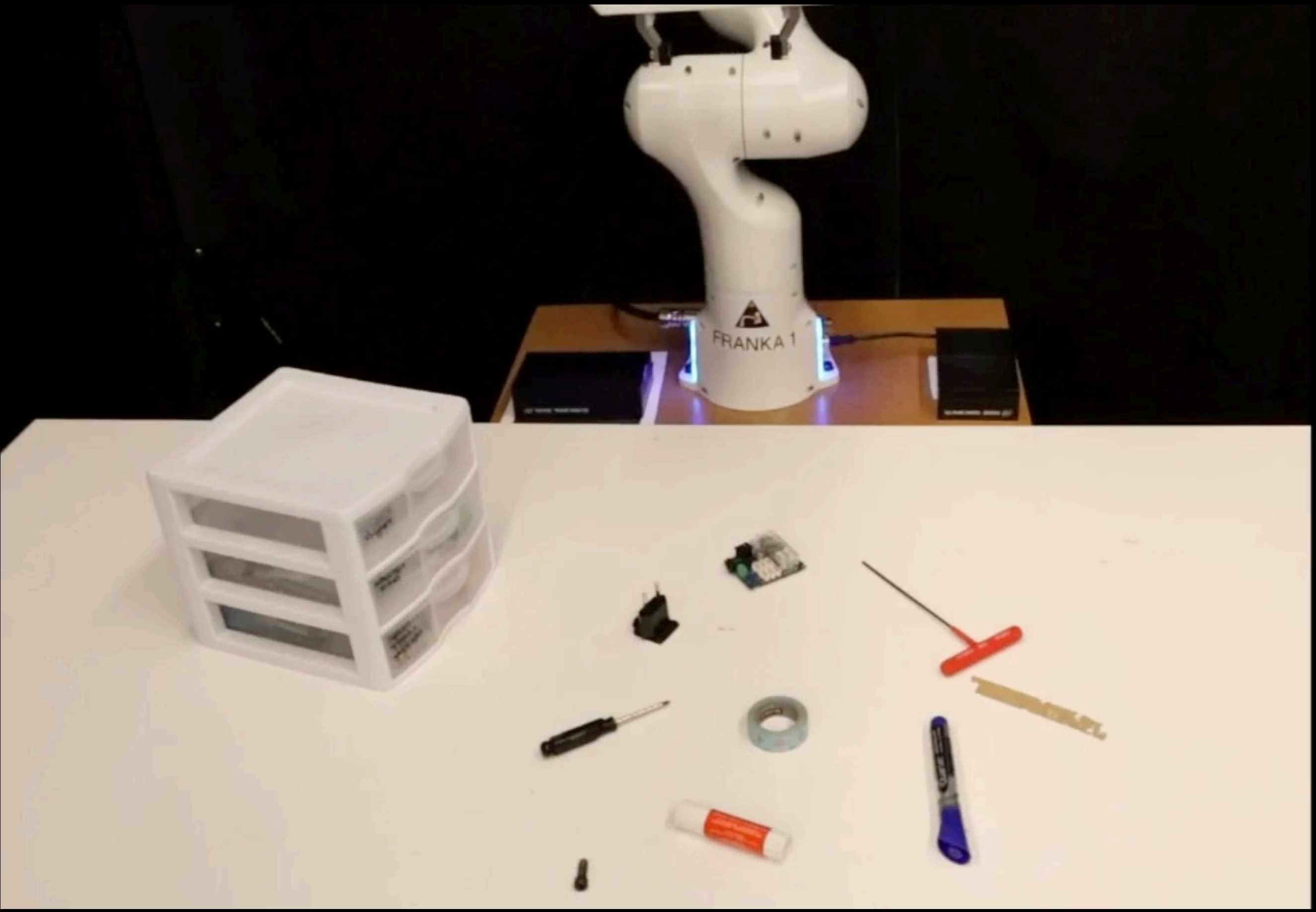
(1) Joint velocities are near zero &

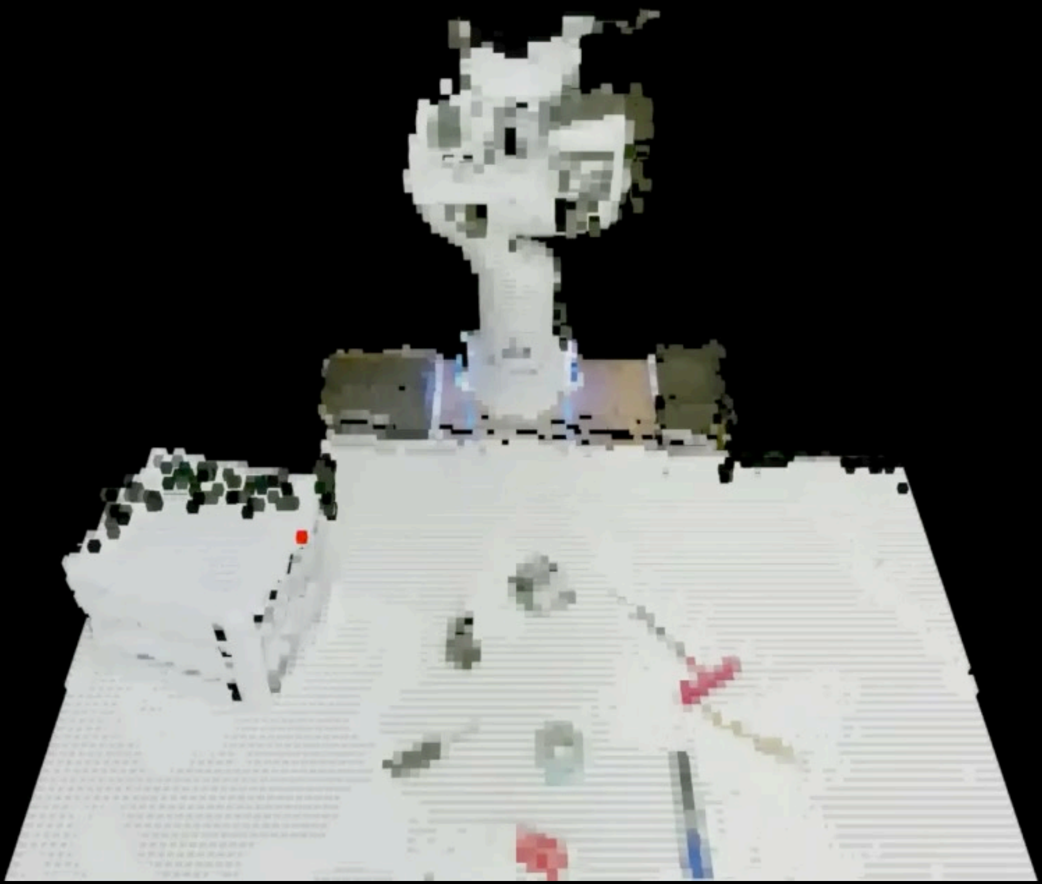
(2) Gripper open state has not changed


2x


FRANKA 1


FRANKA 1



FRANKA 1



PerAct Takeaways

Multi-task paradigm (from vision & NLP) might also work for robotics

PerAct Takeaways

Multi-task paradigm (from vision & NLP) might also work for robotics

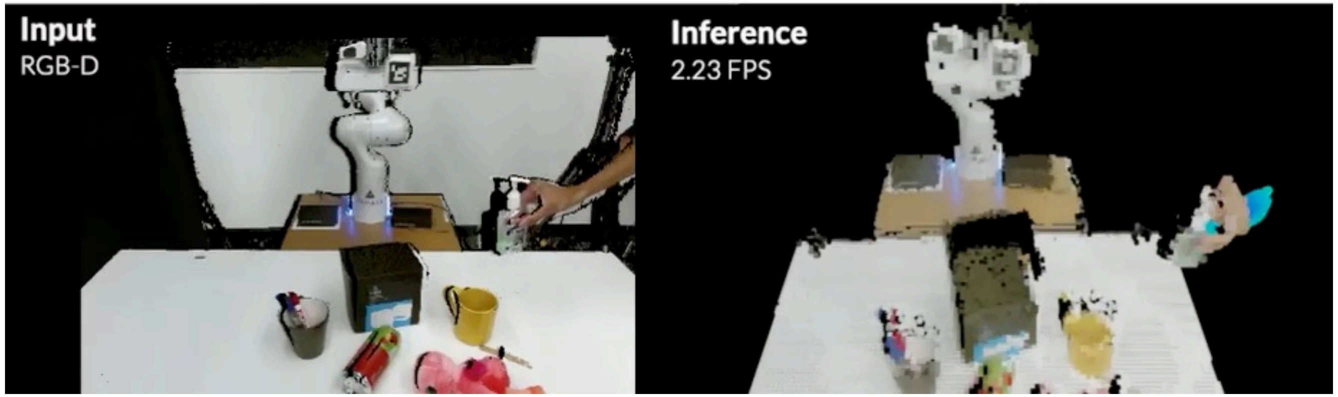
The right problem formulation can make a huge difference for scaling Transformers for robotics

PerAct Takeaways

Multi-task paradigm (from vision & NLP) might also work for robotics

The right problem formulation can make a huge difference for scaling Transformers for robotics

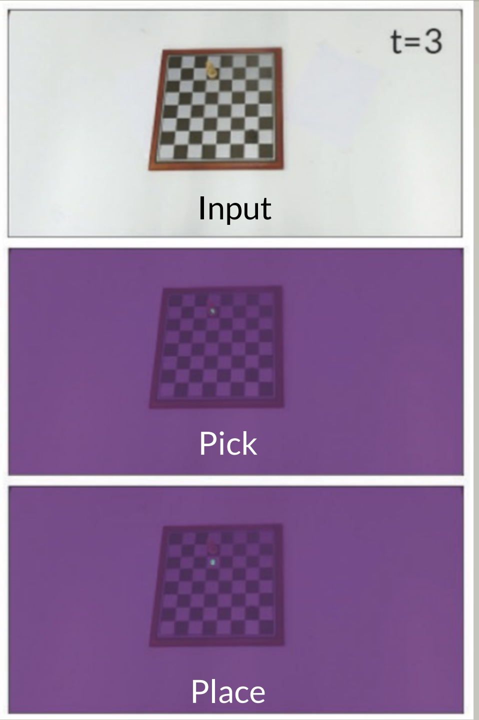
Traditional robotic perception (detecting, pose estimating, grasping) might fallout out of action-centric models

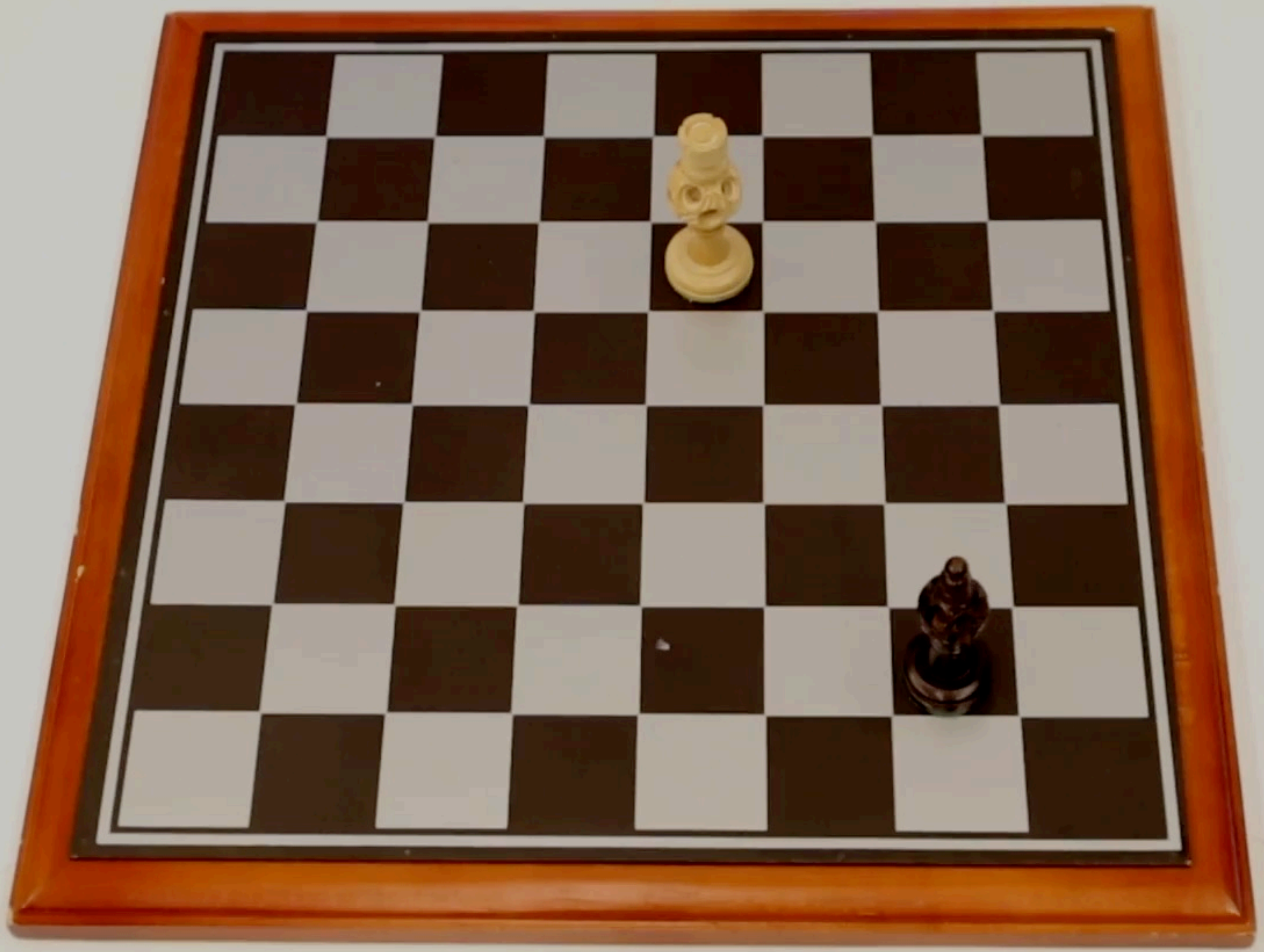


More 2D action detections

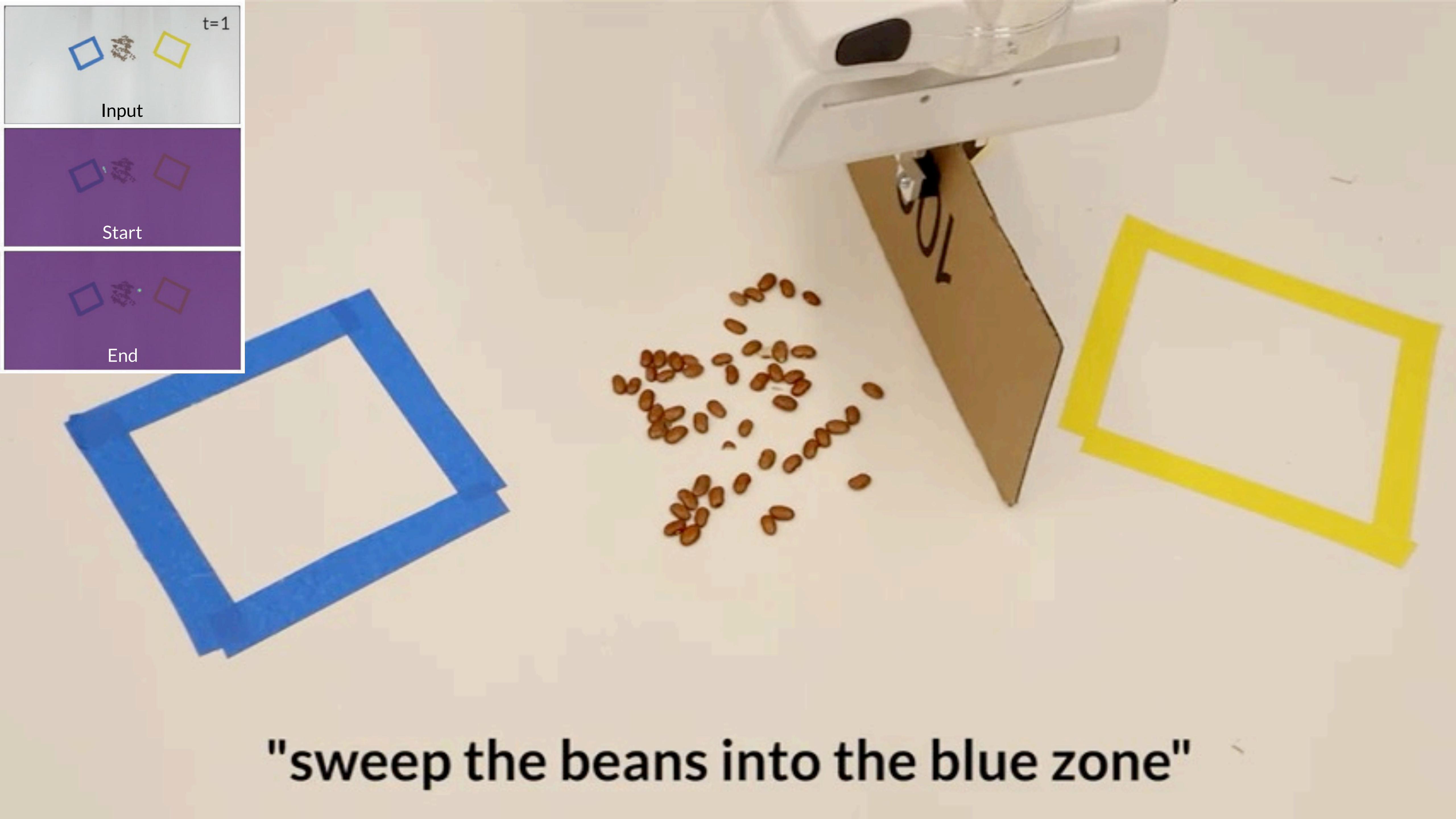
from **CLIPort**


t=1
Input
Pick
Place
"fold the cloth in half"





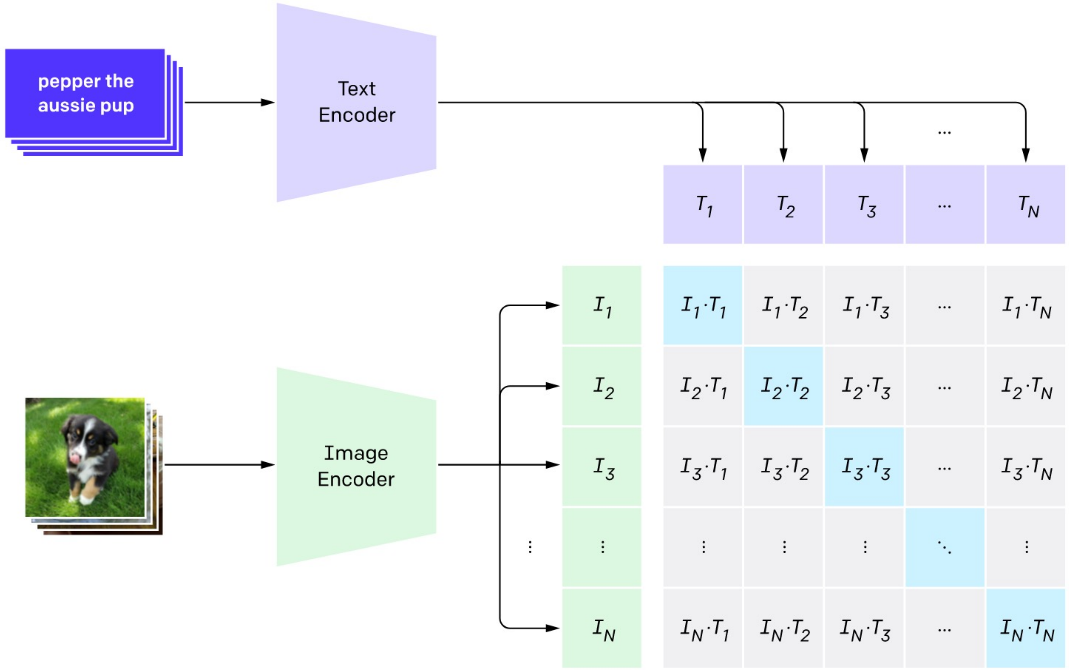
"move the rook one block right"


t=1
Input
Start
End
"sweep the beans into the blue zone"

t=2
Input
Pick
Place

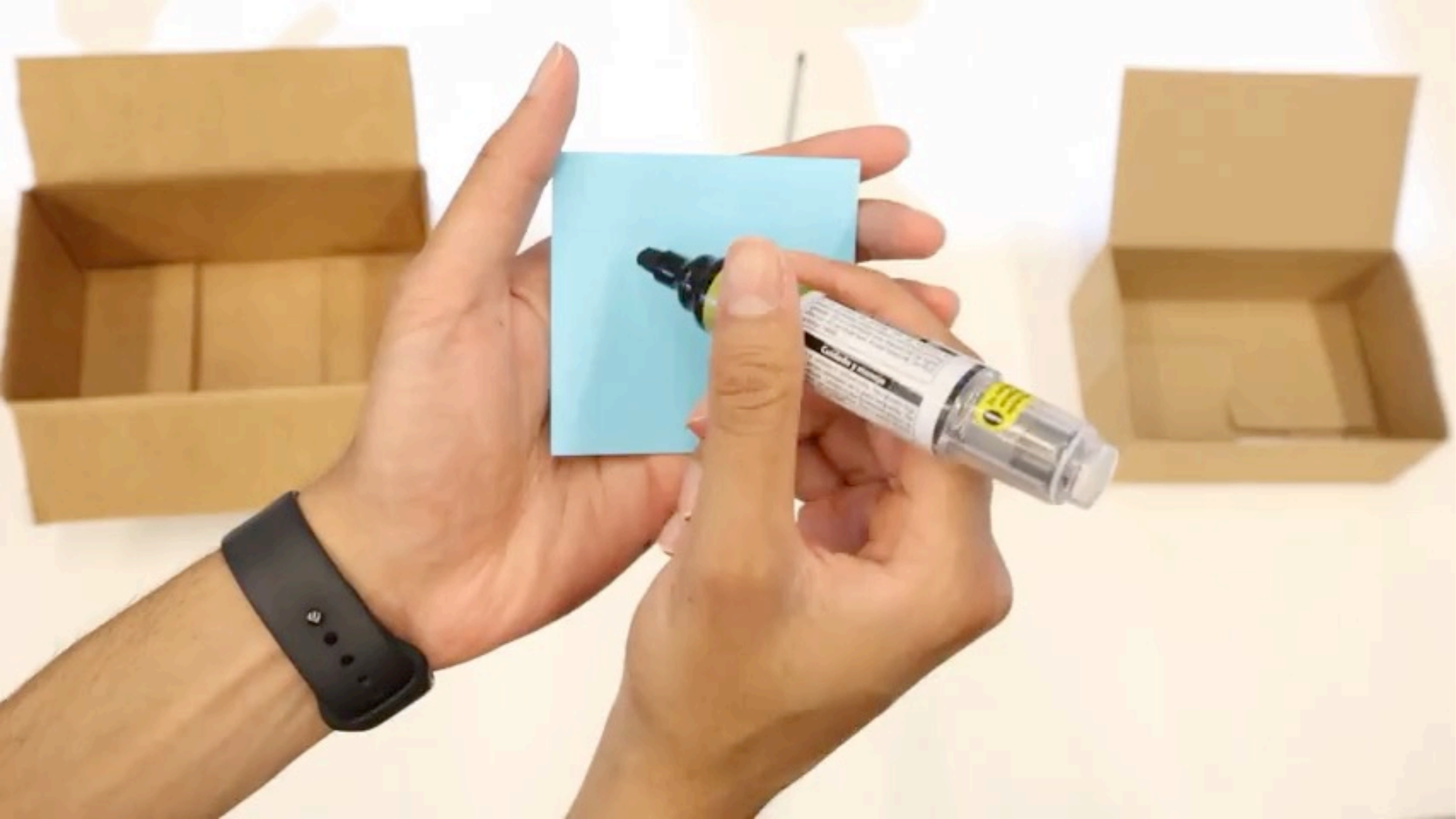
David Marr

“Vision is a computational process that transforms the retinal image into an objective representation of 3D shape.”



Actionless priors (like pretrained CLIP)

can speed up learning



Three Perspectives on Vision

David Marr

Image
Processing

Vision Priors
like CLIP

James Gibson

Action
Directed

Action
Affordances

Jan Koenderink

Goal &
Embodiment
Specific

Language
Goals

Credit: Lana Lazebnik

Summary

Summary

👁️

Learning **visual** representations

Summary

👁️ 👋🏻

Learning **visual** representations of **actions**

Summary

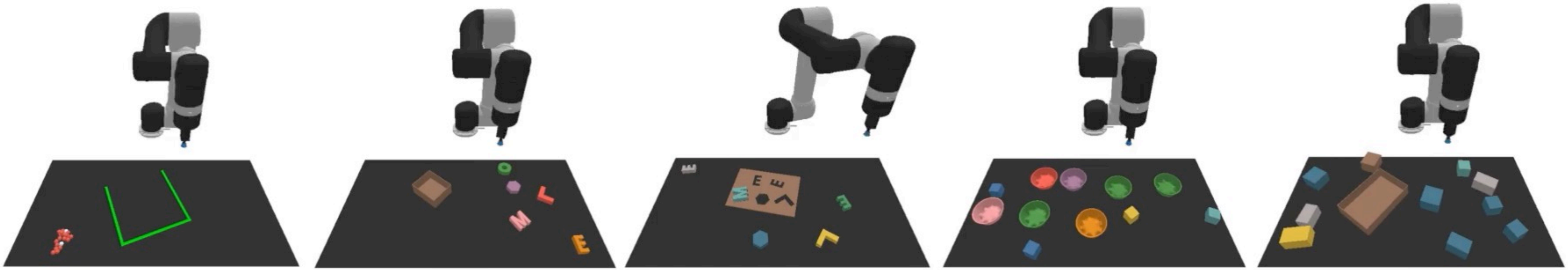
Learning **visual** representations of **actions** conditioned on **language**

PerAct

Paper, videos, Colab, code:
peract.github.io

CLIPort

Paper, videos, code, models:
cliport.github.io



"align the rope from back right corner to back left corner" | "pack the hexagon in the brown box" | "put the green letter E in the right letter E shape hole" | "put the blue blocks in a green bowl" | "pack all the yellow and blue blocks in the brown box"

Limitations

Hard to extend to dynamic and dexterous manipulation

Struggles with unseen objects

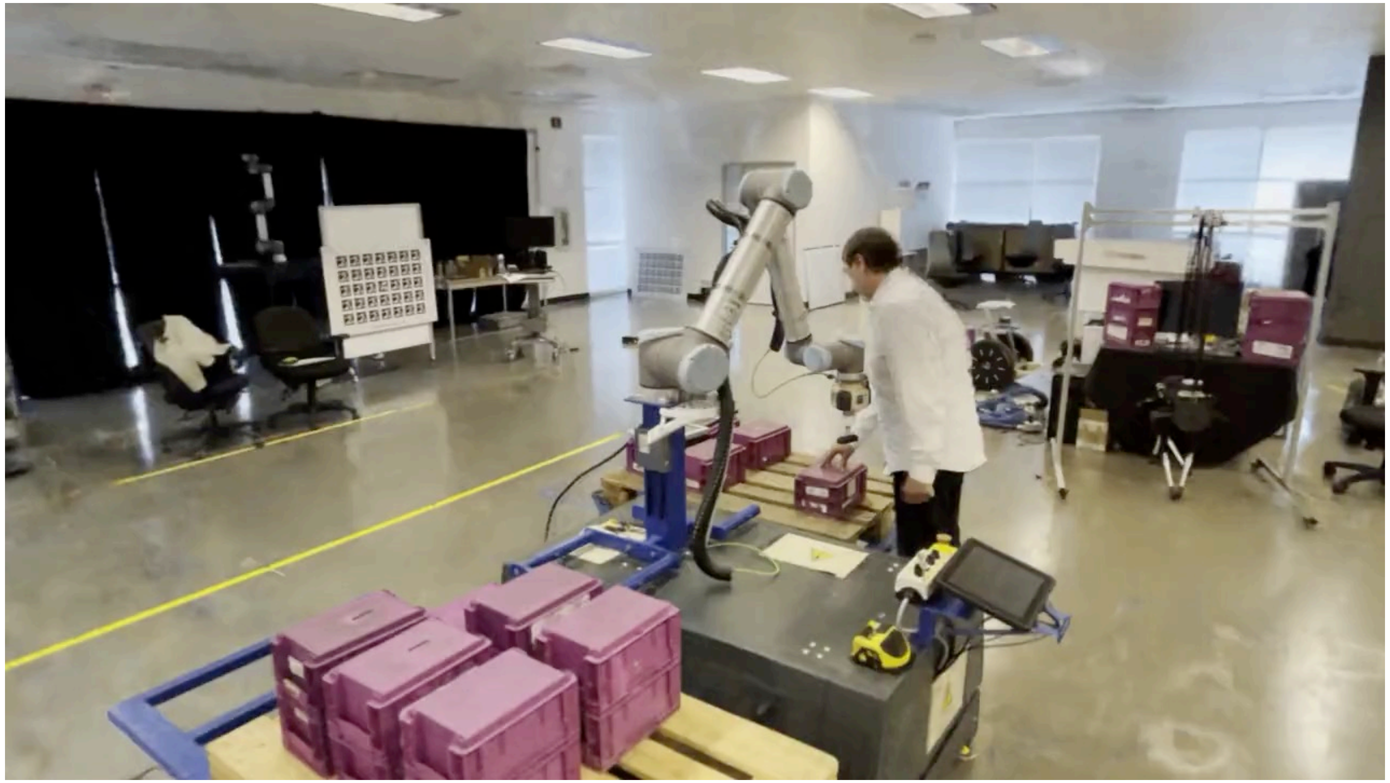
Does not predict task-completion

Struggles with complex spatial relationships

Needs good hand-eye

Scope of language (especially verbs) is mostly limited to the training distribution

What next?

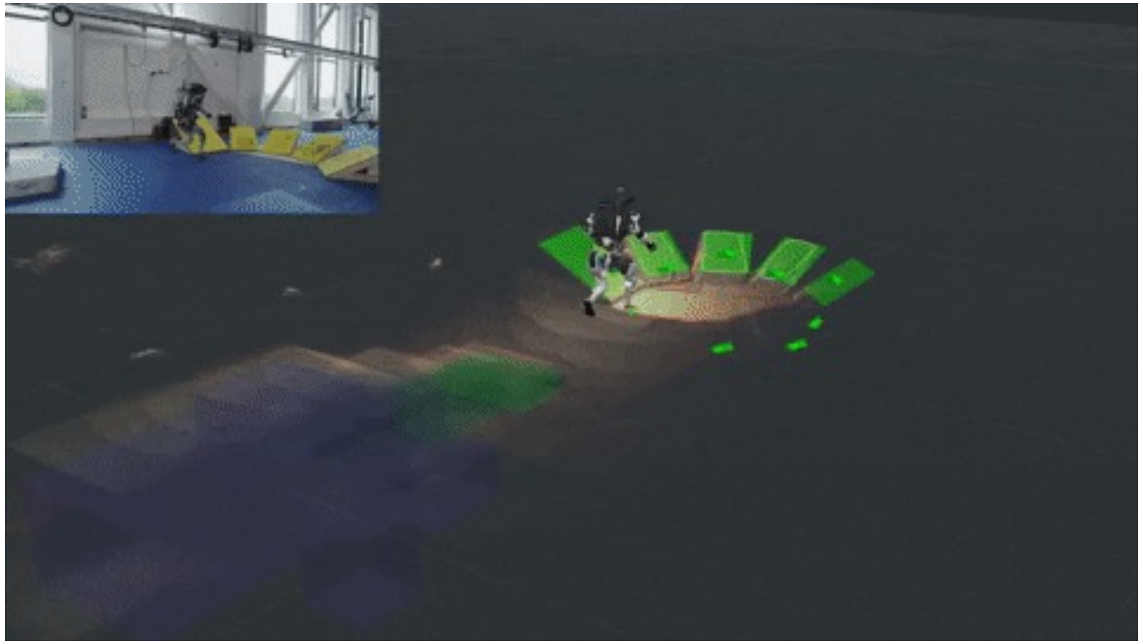


NeRF voxel features?

What next?

NeRF voxel features?

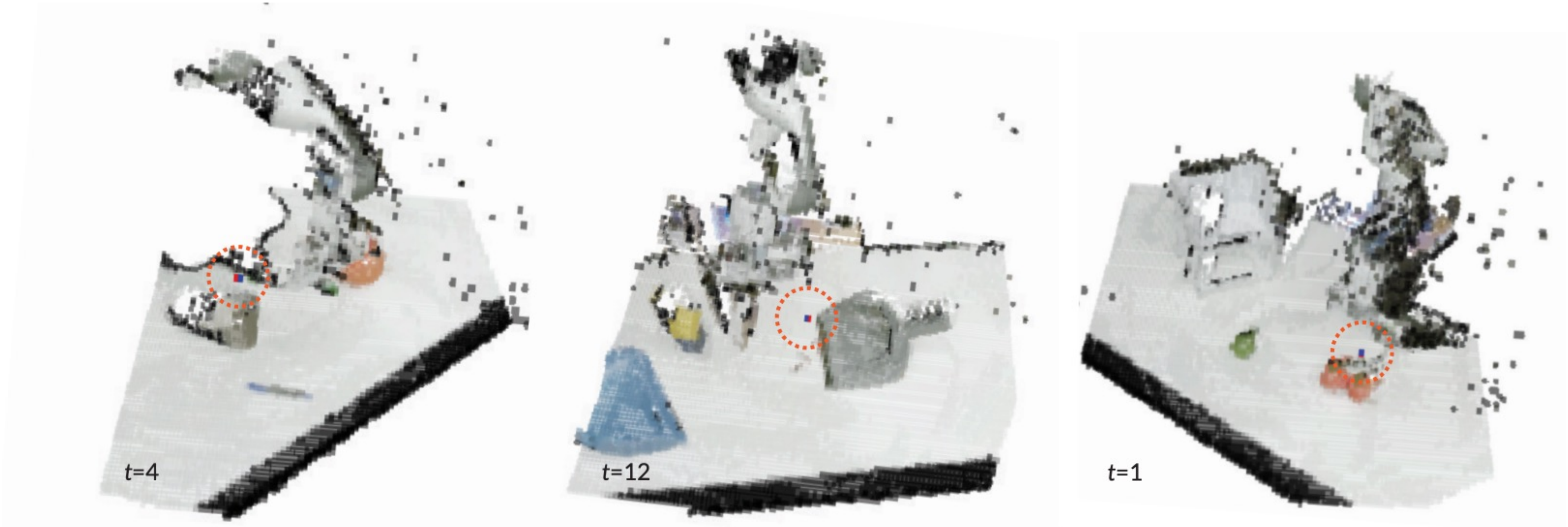
What next?



Detect footprint poses? bimanual grippers poses? finger tips?

Appendix

Data Augmentation



"put the green whiteboard marker in the mug"

"sweep the beans into the gray dustpan"

"put the tomatoes in the top bin"

Data Augmentation

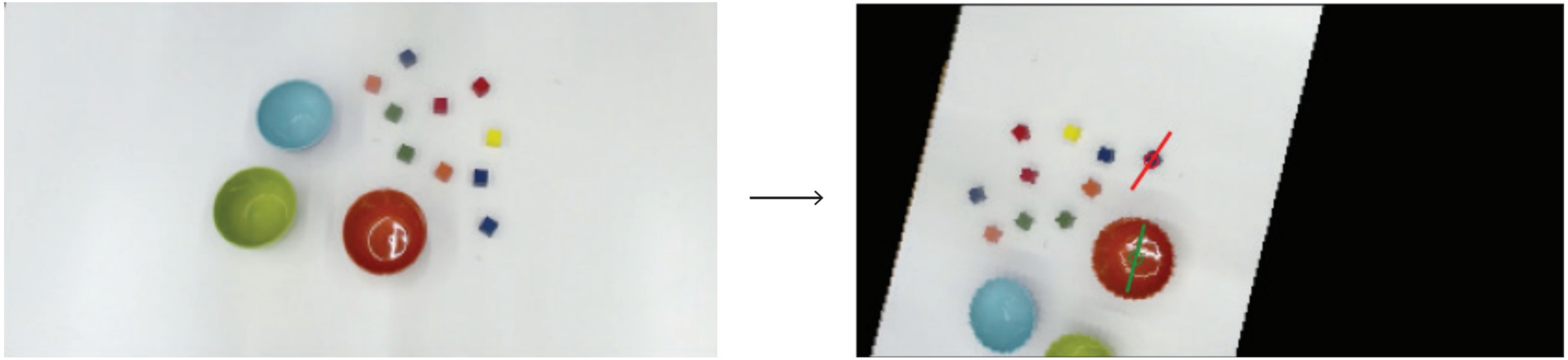


Figure 9. **Data Augmentation**: $\mathbf{SE}(2)$ transform applied to RGB-D input. The left image shows the original input, and the right image shows the transformed input along with expert $\mathcal{T}_{\text{pick}}$ (red) and $\mathcal{T}_{\text{place}}$ (green) actions.

Perturbation Tests

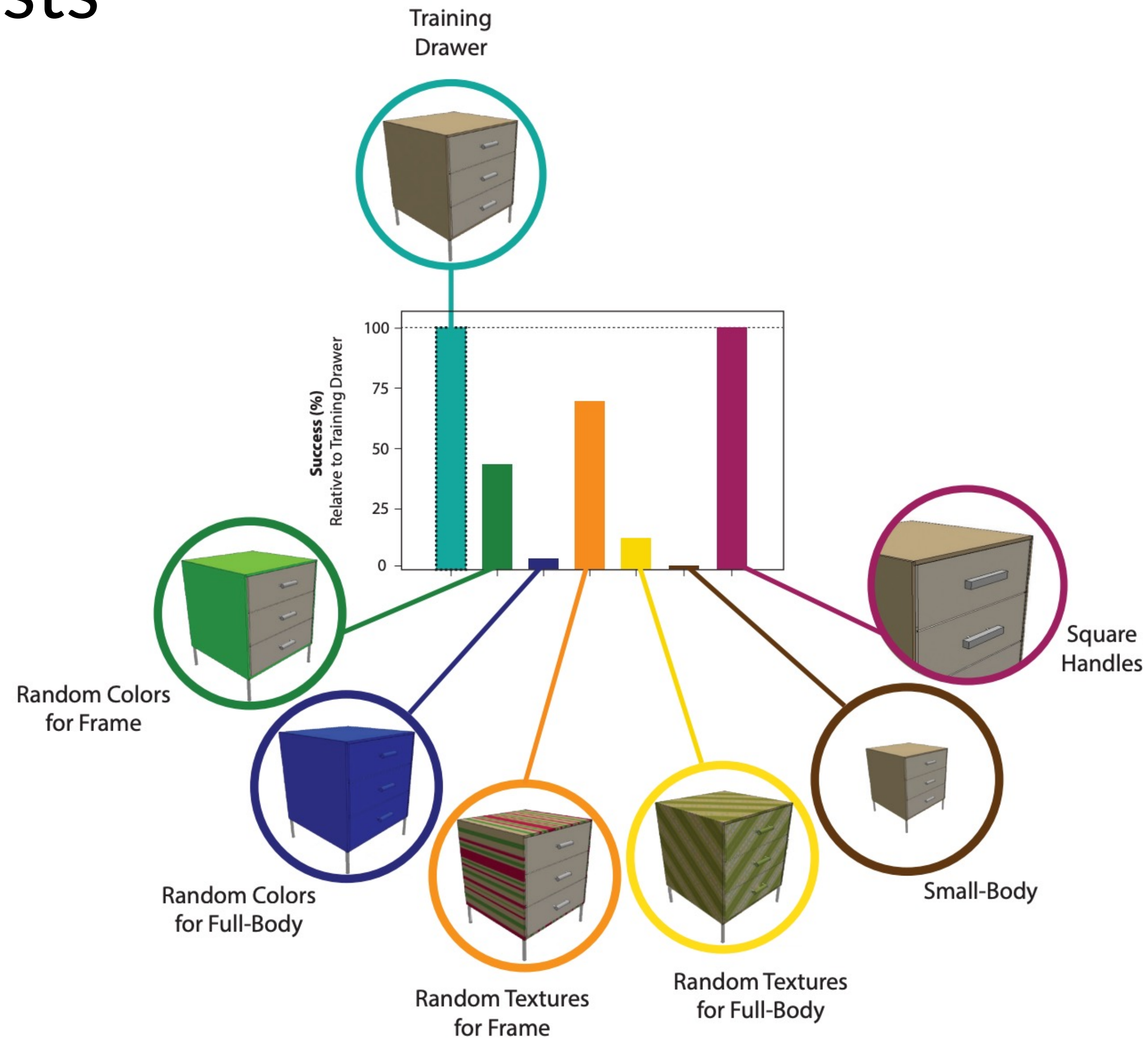
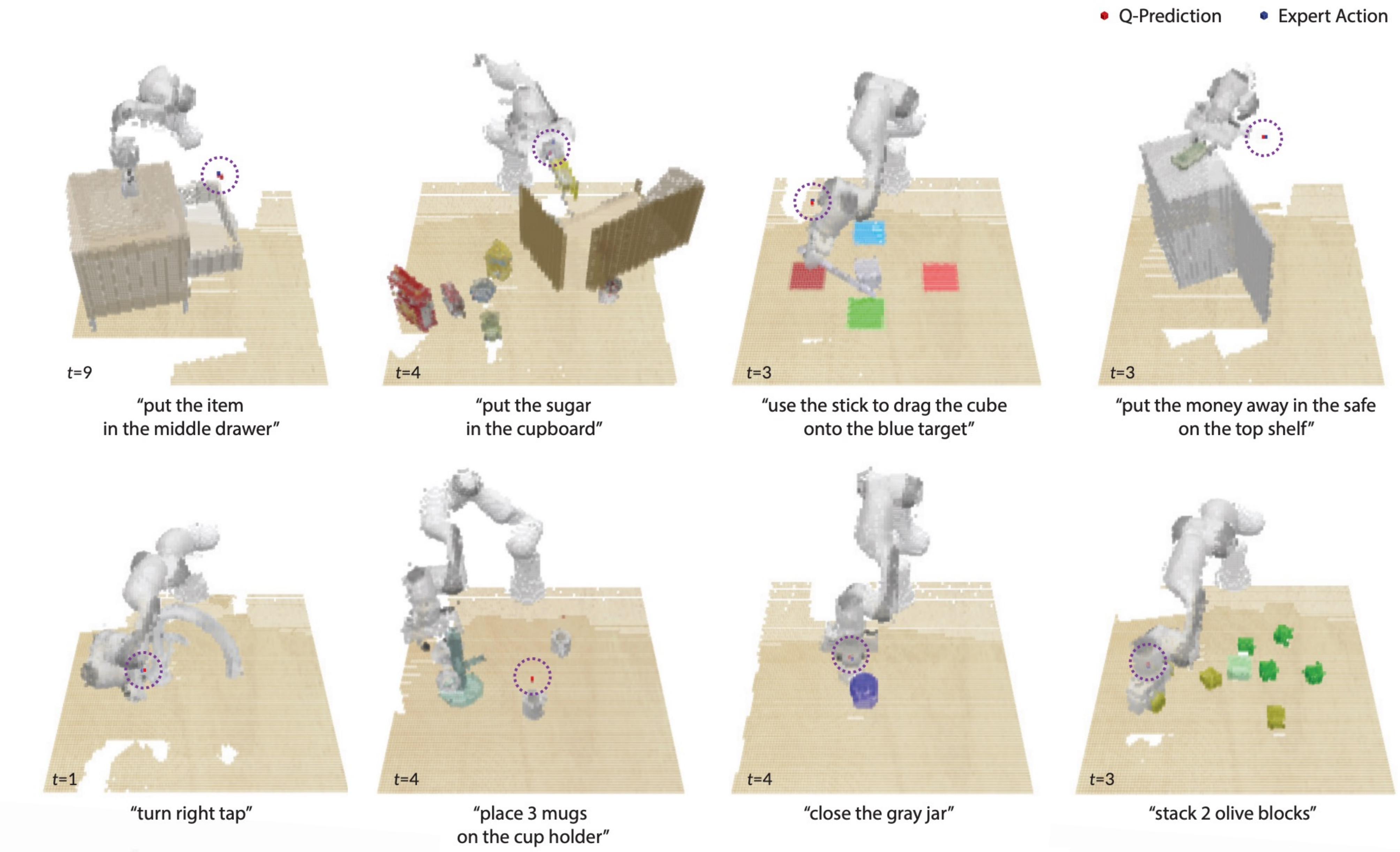
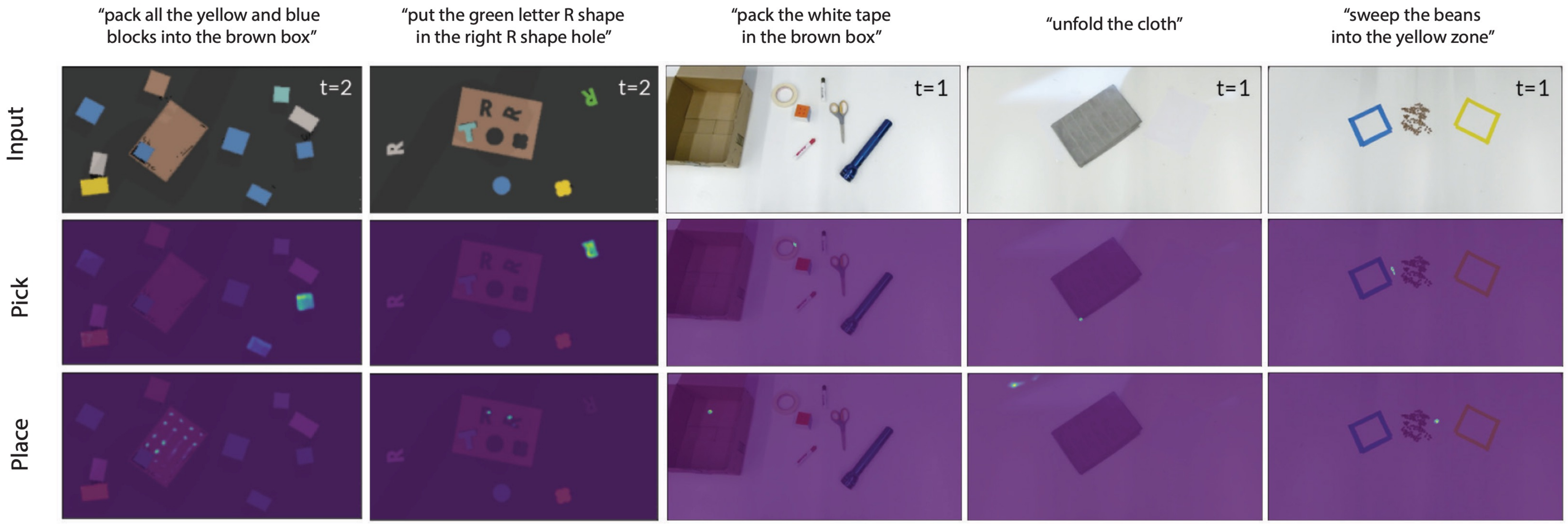


Figure 11. **Perturbation Tests.** Results from a multi-task PERACT agent trained on a single drawer and evaluated on several instances perturbed drawers. Each perturbation consists of 25 evaluation episodes, and reported successes are relative to the training drawer.

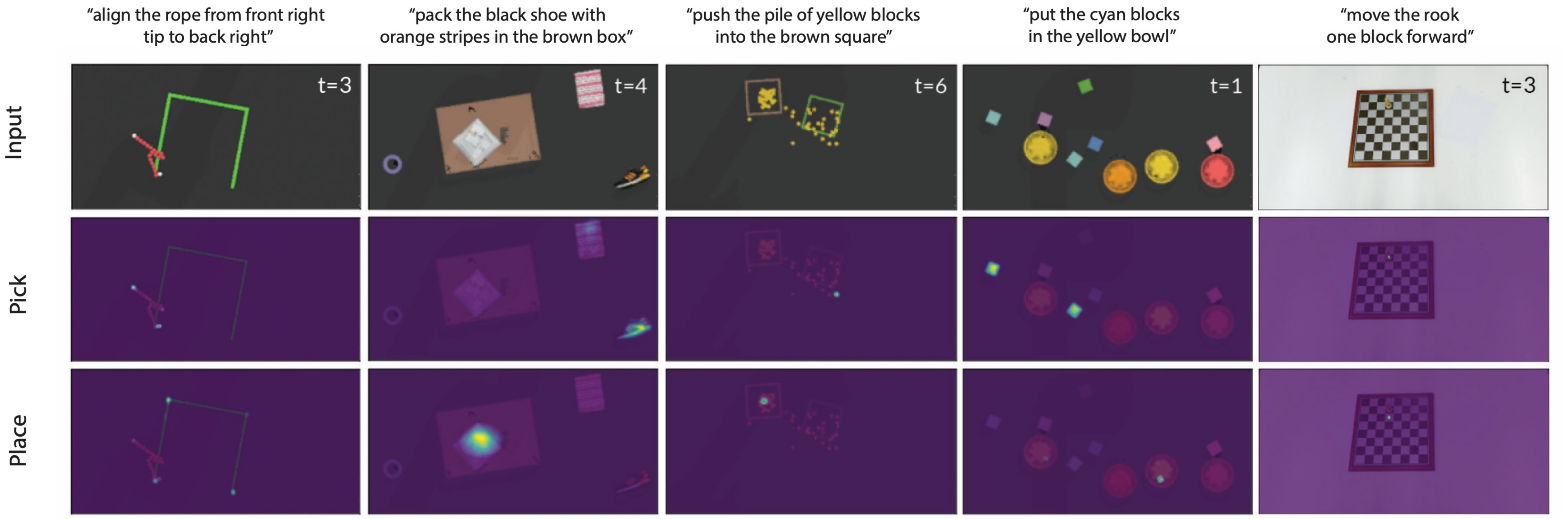
More Affordances



More Affordances



More Affordances



Multi Modal Actions

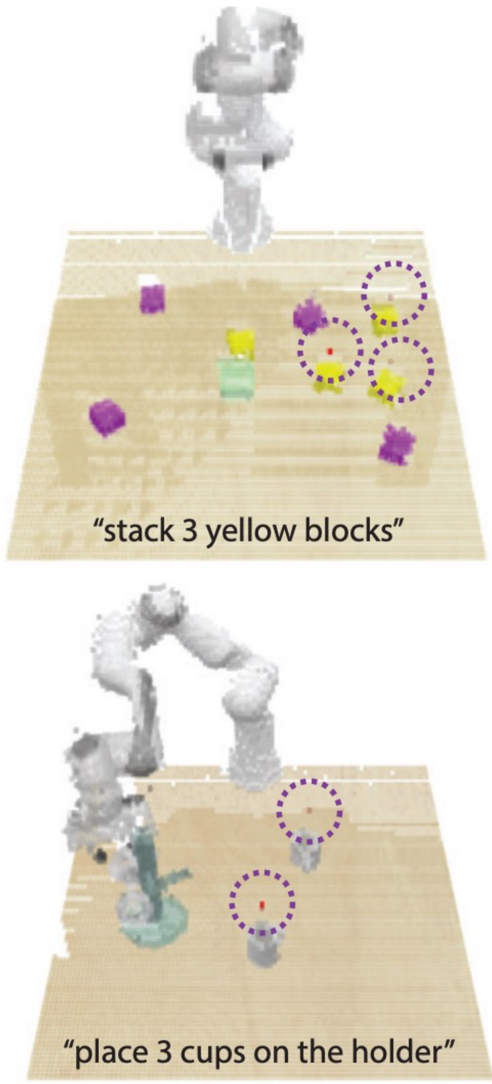


Figure 13. **Examples of Multi-Modal Predictions.**