LLM-Driven Mobile Manipulation

Speaker: Liu Dai

Collaborator: Gireesh Nandiraju, Jiazhao Zhang

Advisor: He Wang

April 11, 2023

CONTENT

- Task Introduction and Works from META

  - Introduction

  - Adaptive Skill Coordination (ASC)

- Selected Papers from Google Robotics

  - **SayCan** (2022, April)

  - **RT-1** (2022, December)

  - **MOO** (2023, March) & CoW

  - **PaLM-E** (2023, March)

SayCan

Do As I Can, Not As I Say:
Grounding Language in Robotic Affordances

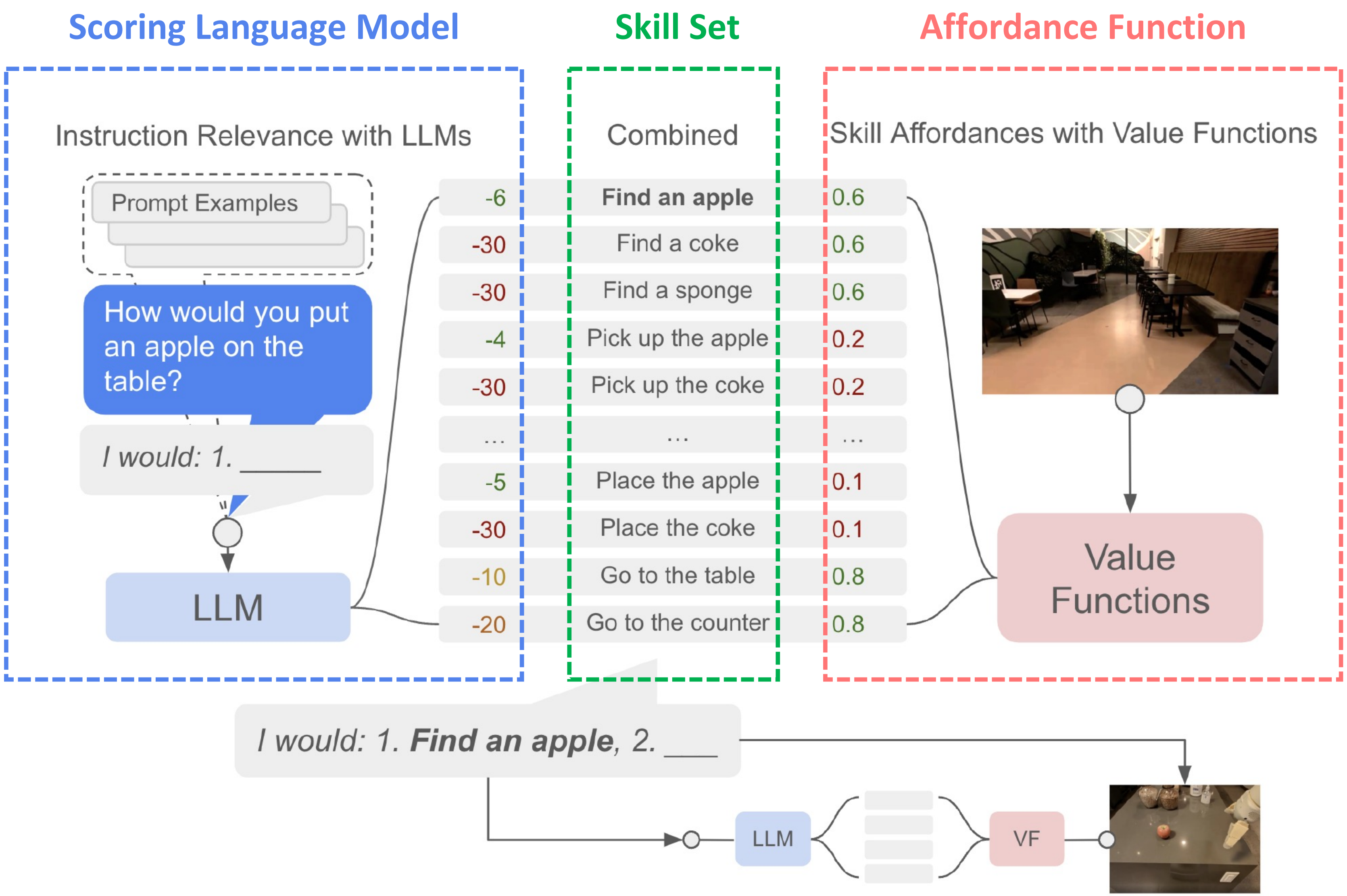
CoRL 2022 Oral

Michael Ahn* Anthony Brohan* Noah Brown* Yevgen Chebotar* Omar Cortes* Byron David* Chelsea Finn*
Chuyuan Fu* Keerthana Gopalakrishnan* Karol Hausman* Alex Herzog* Daniel Ho* Jasmine Hsu* Julian Ibarz*
Brian Ichter* Alex Irpan* Eric Jang* Rosario Jauregui Ruano* Kyle Jeffrey* Sally Jesmonth* Nikhil Joshi*
Ryan Julian* Dmitry Kalashnikov* Yuheng Kuang* Kuang-Huei Lee* Sergey Levine* Yao Lu* Linda Luu* Carolina Parada*
Peter Pastor* Jornell Quiambao* Kanishka Rao* Jarek Rettinghouse* Diego Reyes* Pierre Sermanet* Nicolas Sievers*
Clayton Tan* Alexander Toshev* Vincent Vanhoucke* Fei Xia* Ted Xiao* Peng Xu* Sichun Xu* Mengyuan Yan* Andy Zeng*

SayCan

LLM-Driven Long-Horizon Tasks

SayCan



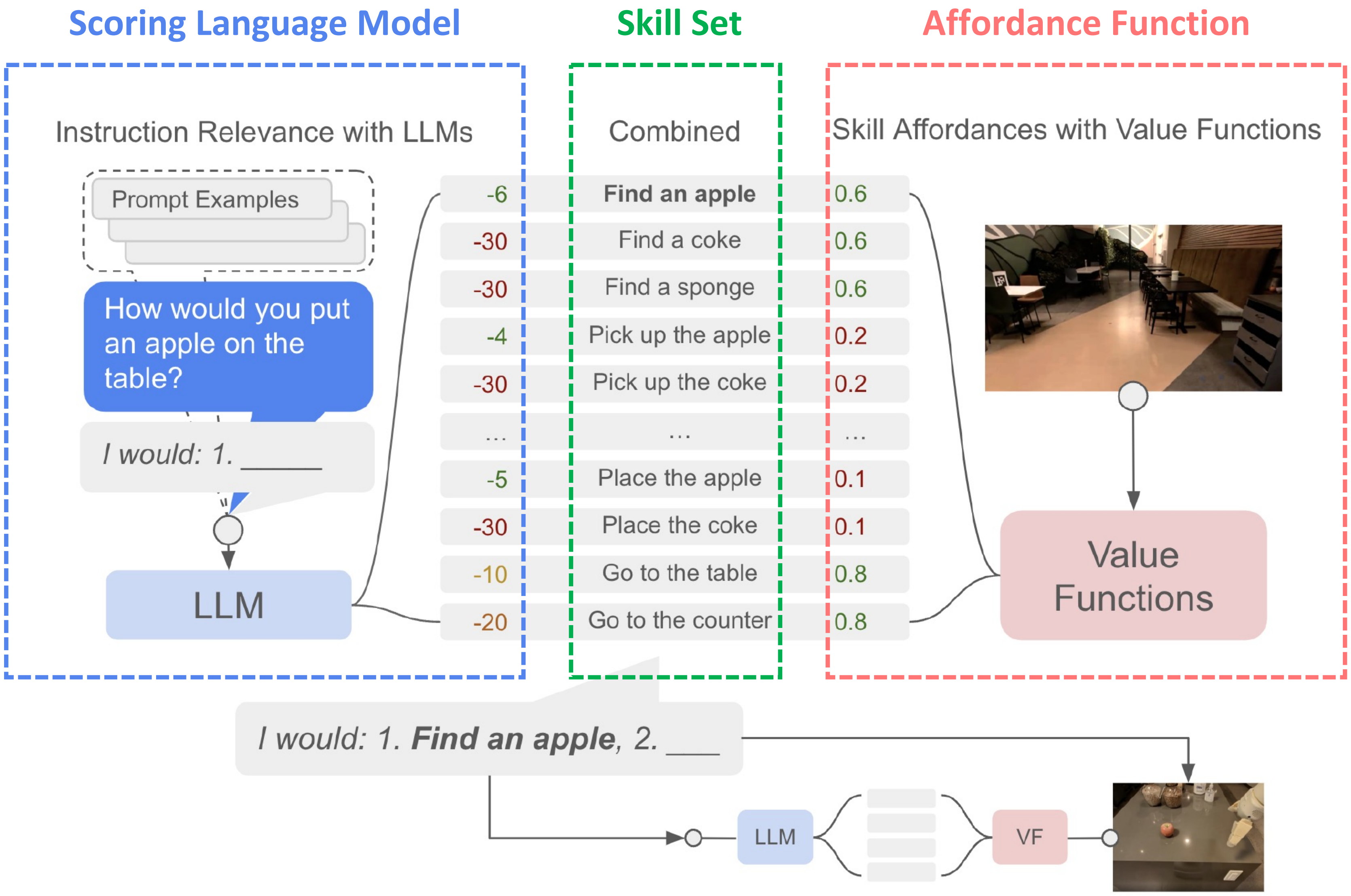
Scoring Language Model | **Skill Set** | **Affordance Function**

Instruction Relevance with LLMs	Combined	Skill Affordances with Value Functions
-6	**Find an apple**	0.6
-30	Find a coke	0.6
-30	Find a sponge	0.6
-4	Pick up the apple	0.2
-30	Pick up the coke	0.2
…	…	…
-5	Place the apple	0.1
-30	Place the coke	0.1
-10	Go to the table	0.8
-20	Go to the counter	0.8

Prompt Examples
How would you put an apple on the table?
I would: 1. _____
LLM
Value Functions
I would: 1. ***Find an apple****, 2. ____*
LLM
VF

SayCan

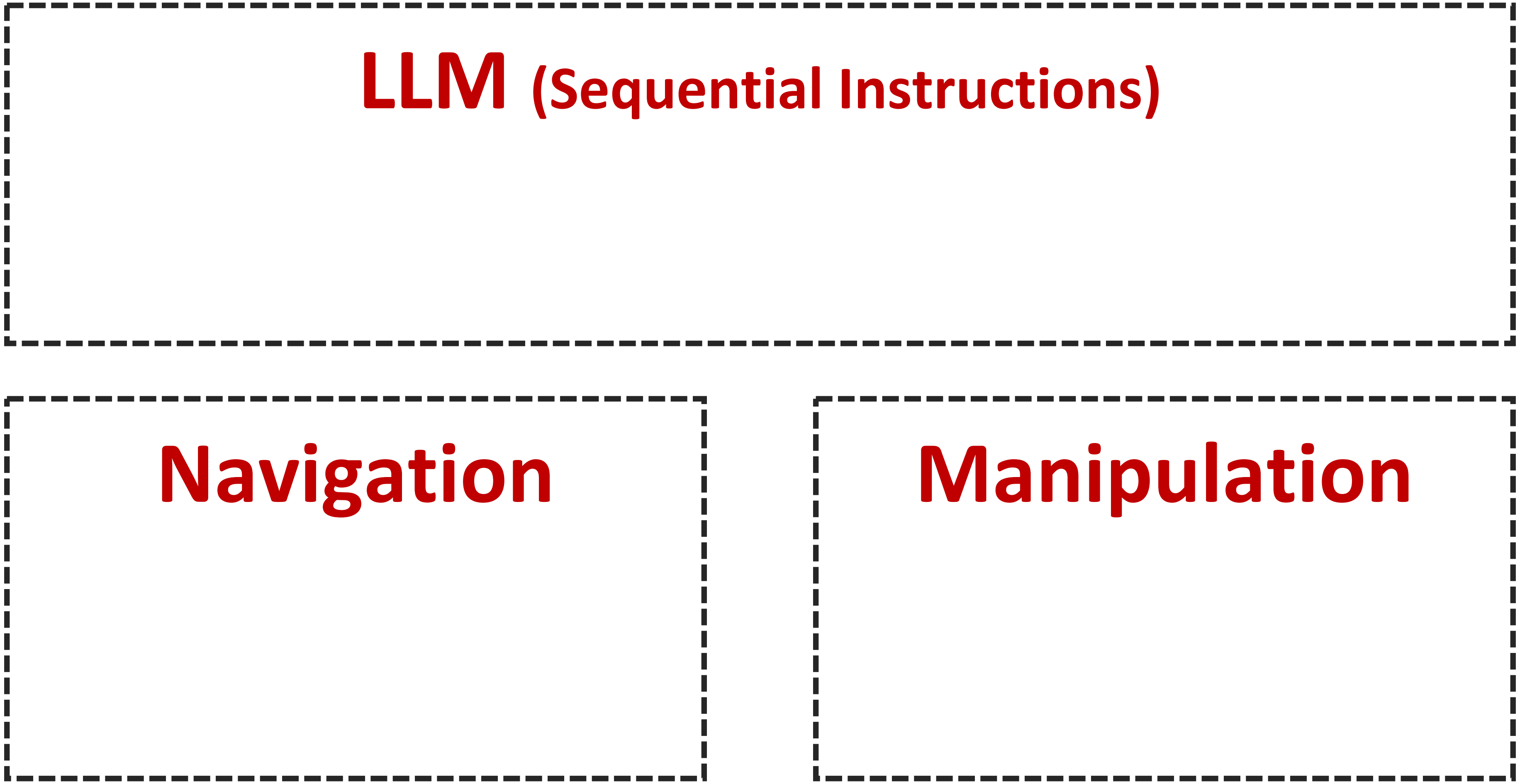
Limit: Have to Train Every Skill Case by Case & Limited Objects


Scoring Language Model
Skill Set
Affordance Function
Instruction Relevance with LLMs
Combined
Skill Affordances with Value Functions
Prompt Examples
How would you put an apple on the table?
I would: 1. ______
LLM
Value Functions
I would: 1. ***Find an apple****, 2. ____*
LLM
VF

Instruction Relevance with LLMs	Combined	Skill Affordances with Value Functions
-6	**Find an apple**	0.6
-30	Find a coke	0.6
-30	Find a sponge	0.6
-4	Pick up the apple	0.2
-30	Pick up the coke	0.2
…	…	…
-5	Place the apple	0.1
-30	Place the coke	0.1
-10	Go to the table	0.8
-20	Go to the counter	0.8

Where are we for now?

LLM-Driven Long-Horizon Tasks



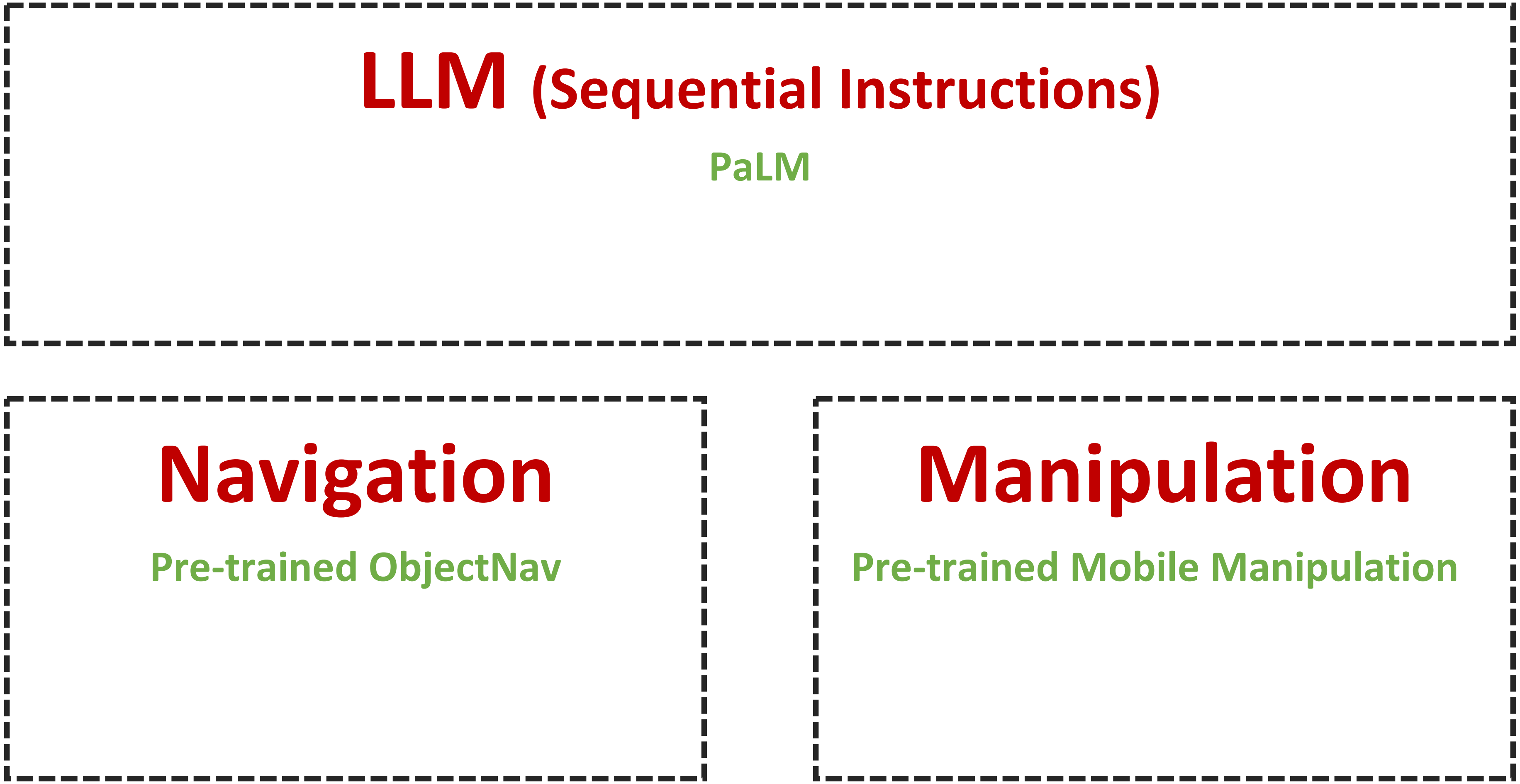
LLM (Sequential Instructions)

Navigation

Manipulation

Where are we for now?

Say-Can



LLM (Sequential Instructions)

PaLM

Navigation

Pre-trained ObjectNav

Manipulation

Pre-trained Mobile Manipulation

RT-1

RT-1: Robotics Transformer

for Real-World Control at Scale

Anthony Brohan Noah Brown Justice Carbajal Yevgen Chebotar Joseph Dabis Chelsea Finn Keerthana Gopalakrishnan
Karol Hausman Alex Herzog Jasmine Hsu Julian Ibarz Brian Ichter Alex Irpan Tomas Jackson
Sally Jesmonth Nikhil Joshi Ryan Julian Dmitry Kalashnikov Yuheng Kuang Isabel Leal Kuang-Huei Lee
Sergey Levine Yao Lu Utsav Malla Deeksha Manjunath Igor Mordatch Ofir Nachum Carolina Parada
Jodilyn Peralta Emily Perez Karl Pertsch Jornell Quiambao Kanishka Rao Michael Ryoo Grecia Salazar
Pannag Sanketi Kevin Sayed Jaspiar Singh Sumedh Sontakke Austin Stone Clayton Tan Huong Tran
Vincent Vanhoucke Steve Vega Quan Vuong Fei Xia Ted Xiao Peng Xu Sichun Xu Tianhe Yu Brianna Zitkovich

Skill	Count	Description	Example Instruction
Pick Object	130	Lift the object off the surface	pick iced tea can
Move Object Near Object	337	Move the first object near the second	move pepsi can near rxbar blueberry
Place Object Upright	8	Place an elongated object upright	place water bottle upright
Knock Object Over	8	Knock an elongated object over	knock redbull can over
Open / Close Drawer	6	Open or close any of the cabinet drawers	open the top drawer
Place Object into Receptacle	84	Place an object into a receptacle	place brown chip bag into white bowl
Pick Object from Receptacle and Place on the Counter	162	Pick an object up from a location and then place it on the counter	pick green jalapeno chip bag from paper bowl and place on counter
Additional tasks	9	Skills trained for realistic, long instructions	pull napkin out of dispenser
Total	744		

RT-1

DATA

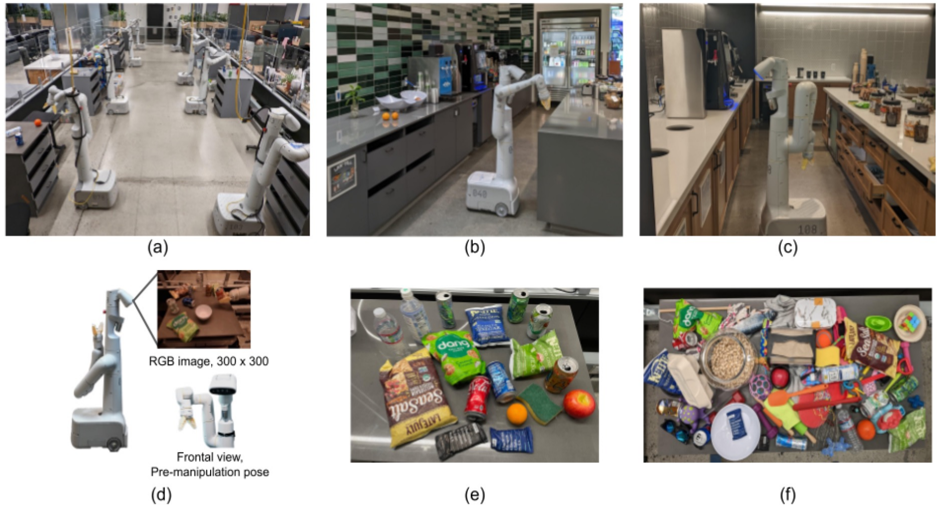
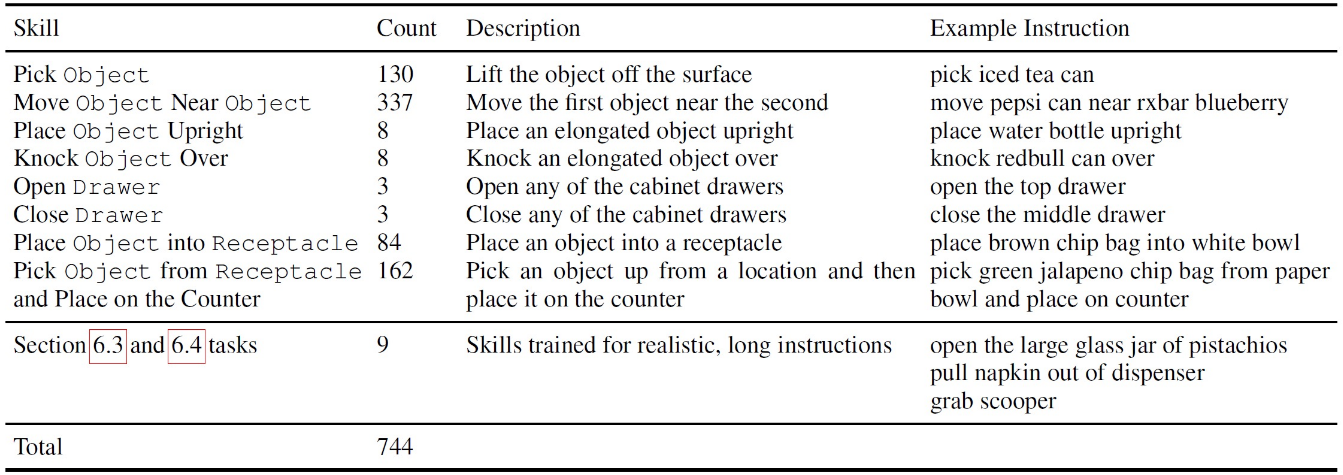


Figure 2: (a) Robot classroom where we collect data at scale; (b) a real office kitchen, one of the two realistic environments used for evaluation (named Kitchen1 in the rest of the paper); (c) a different office kitchen used for evaluation (named Kitchen2 in the rest of the paper); (d) mobile manipulator used throughout the paper; (e) a set of objects used for most of the skills to expand skill diversity; (f) a more diverse set of objects used mostly to expand object diversity of the picking skill.



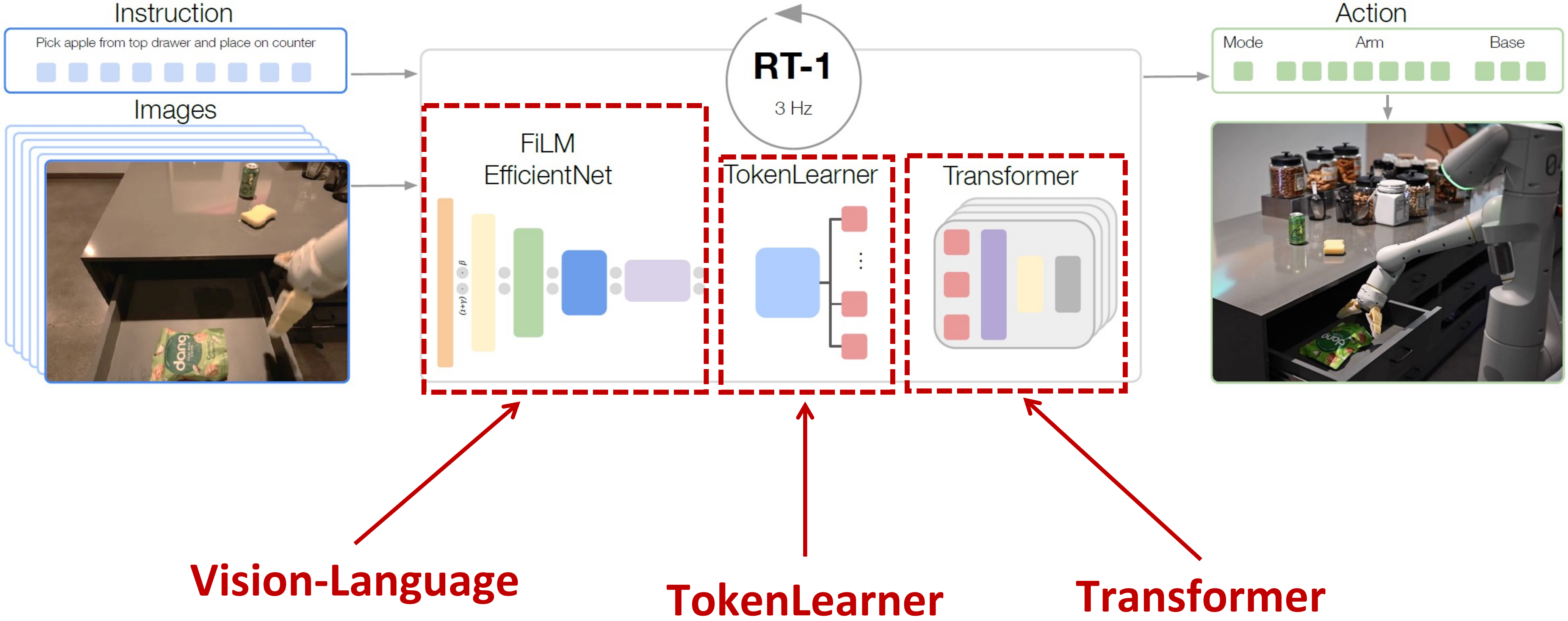
Skill	Count	Description	Example Instruction
Pick `Object`	130	Lift the object off the surface	pick iced tea can
Move `Object` Near `Object`	337	Move the first object near the second	move pepsi can near rxbar blueberry
Place `Object` Upright	8	Place an elongated object upright	place water bottle upright
Knock `Object` Over	8	Knock an elongated object over	knock redbull can over
Open `Drawer`	3	Open any of the cabinet drawers	open the top drawer
Close `Drawer`	3	Close any of the cabinet drawers	close the middle drawer
Place `Object` into `Receptacle`	84	Place an object into a receptacle	place brown chip bag into white bowl
Pick `Object` from `Receptacle` and Place on the Counter	162	Pick an object up from a location and then place it on the counter	pick green jalapeno chip bag from paper bowl and place on counter
Section 6.3 and 6.4 tasks	9	Skills trained for realistic, long instructions	open the large glass jar of pistachios pull napkin out of dispenser grab scooper
Total	744		

- **Human Demonstrations**
- **Description & Instructions**

700 Tasks, 130K Episodes, 13 Robots, 17 Months

RT-1

METHOD : data-absorbent model



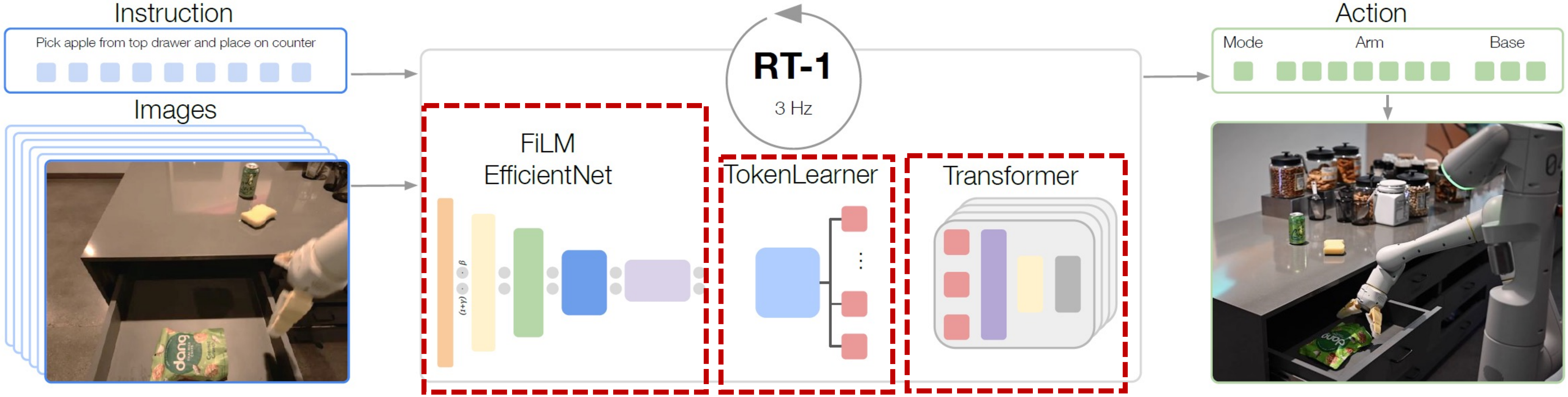
Vision-Language Tokenization

TokenLearner

Transformer

RT-1

METHOD : data-absorbent model



Vision-Language Tokenization

16M

TokenLearner

? M

Transformer

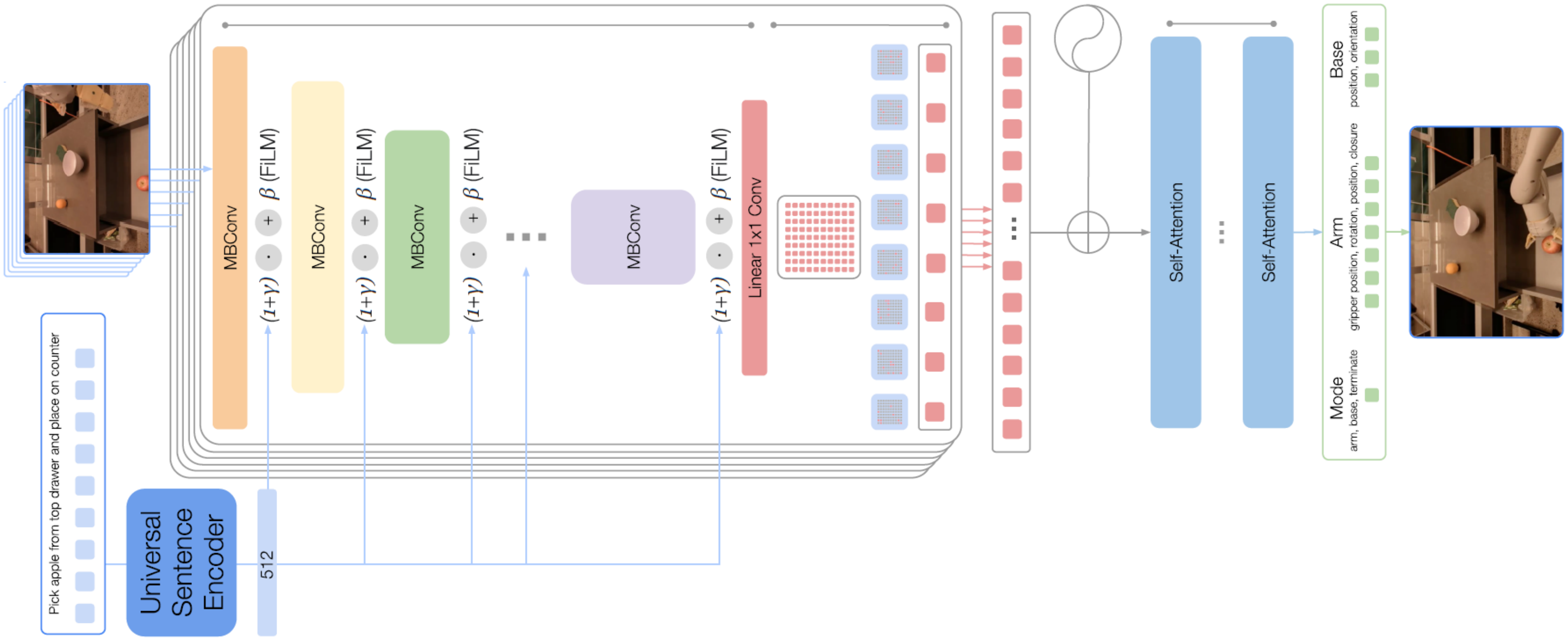
19M

RT-1

METHOD : data-absorbent model

6 x Images

Instruction

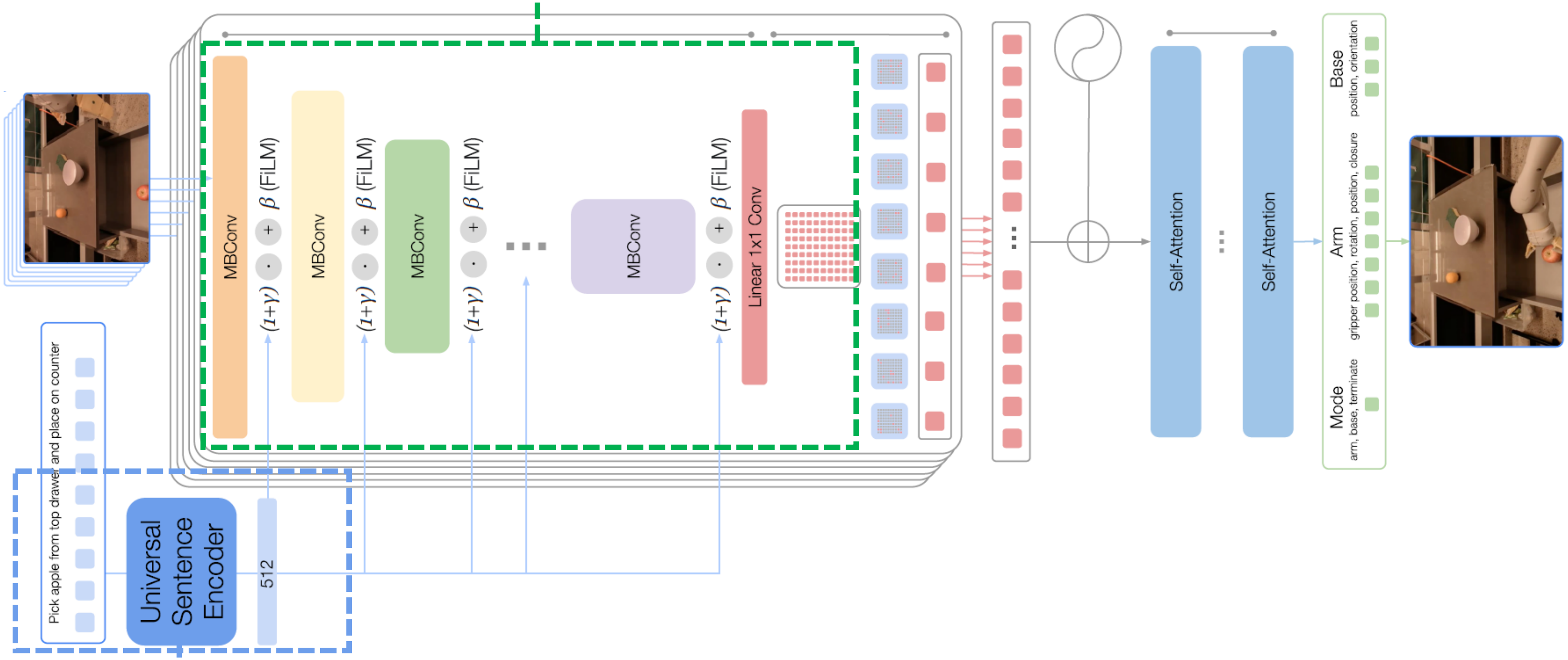


RT-1

METHOD : data-absorbent model

EfficientNET-B3

6 x Images


MBConv
$(1+\gamma)$ · + β (FiLM)
MBConv
$(1+\gamma)$ · + β (FiLM)
MBConv
$(1+\gamma)$ · + β (FiLM)
...
MBConv
$(1+\gamma)$ · + β (FiLM)
Linear 1x1 Conv
Self-Attention
Self-Attention
Base
position, orientation
Arm
gripper position, rotation, position, closure
Mode
arm, base, terminate
Pick apple from top drawer and place on counter
Universal Sentence Encoder
512

Instruction

FiLM Language Embedding

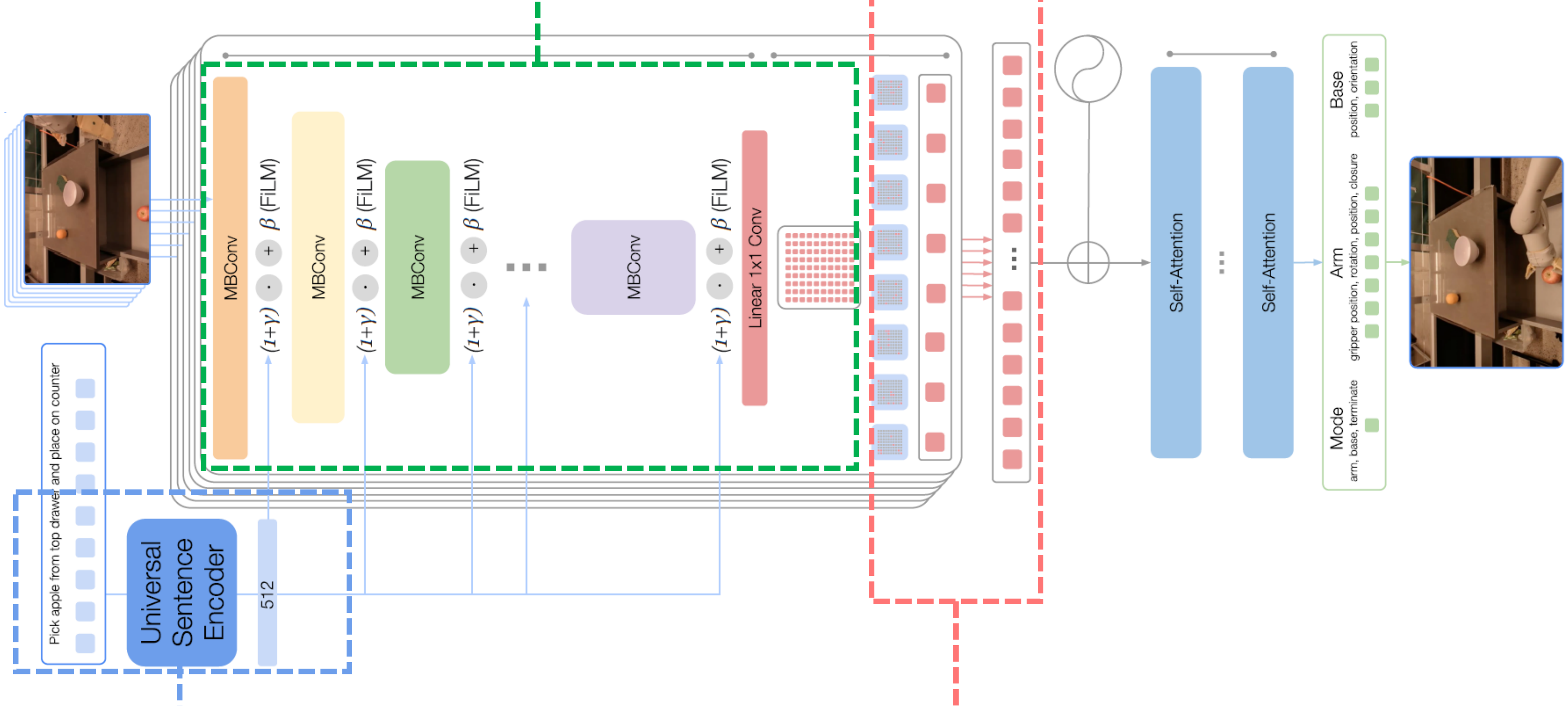
RT-1

METHOD : data-absorbent model

EfficientNET-B3

6 x Images

Instruction


MBConv
$(1+\gamma)$ · + β (FiLM)
MBConv
$(1+\gamma)$ · + β (FiLM)
MBConv
$(1+\gamma)$ · + β (FiLM)
MBConv
$(1+\gamma)$ · + β (FiLM)
Linear 1x1 Conv
Pick apple from top drawer and place on counter
Universal Sentence Encoder
512
Self-Attention
Self-Attention
Base
position, orientation
Arm
gripper position, rotation, position, closure
Mode
arm, base, terminate

FiLM Language Embedding

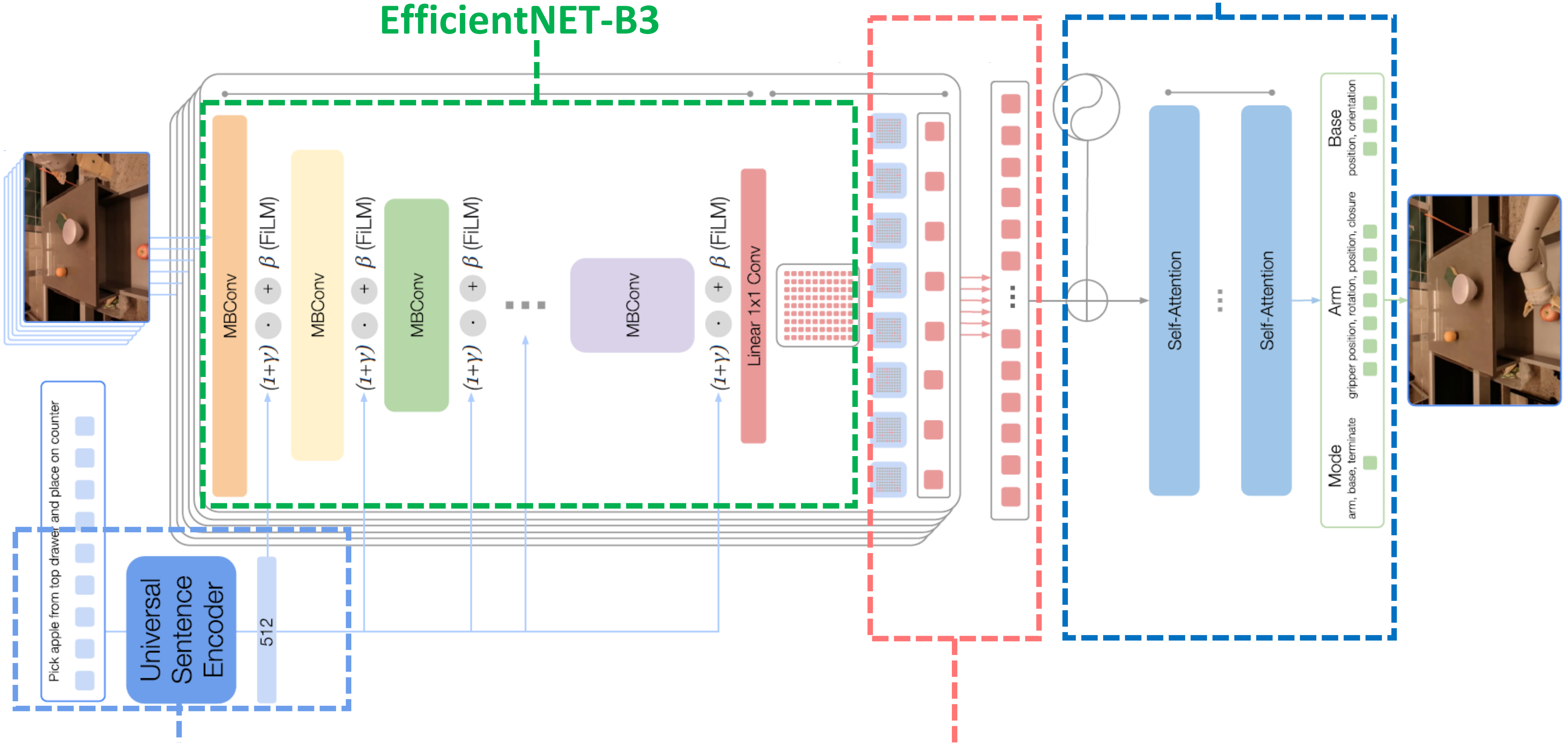
TokenLearner: 6 x 9 x 9 x 512 → 6 x 8 x 512

RT-1

METHOD : data-absorbent model

6 x Images

Instruction



EfficientNET-B3

Transformer

FiLM Language Embedding

TokenLearner: 6 x 9 x 9 x 512 → 6 x 8 x 512

Output:

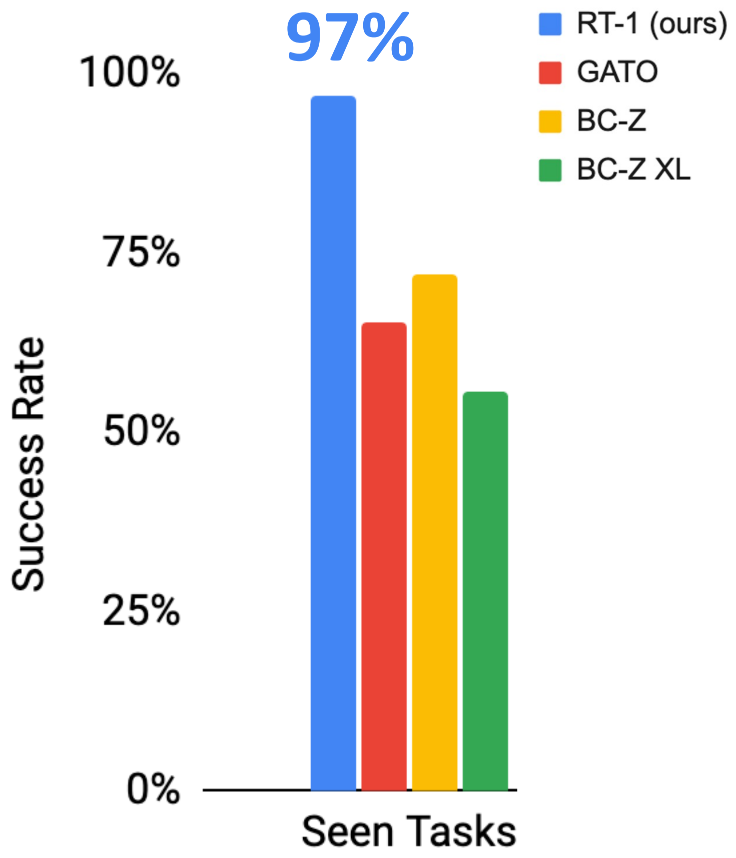
1D Mode
Arm, Base, Stop

3D Base
X, Y, Yaw

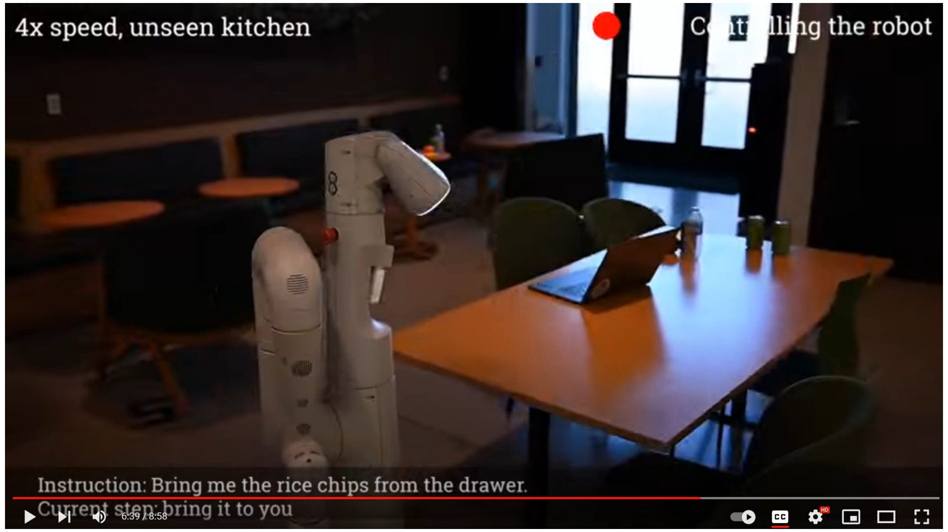
7D Arm
X, Y, Z
Roll, Pitch, Yaw
Gripper

RT-1

Performance

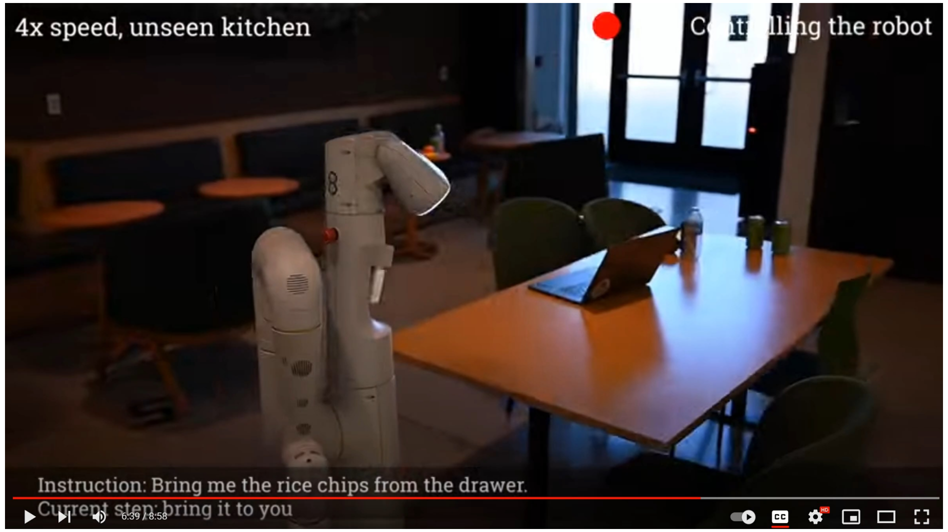


Articulated Object:



Instruction: Bring me the rice chips from the drawer.
Current step: bring it to you

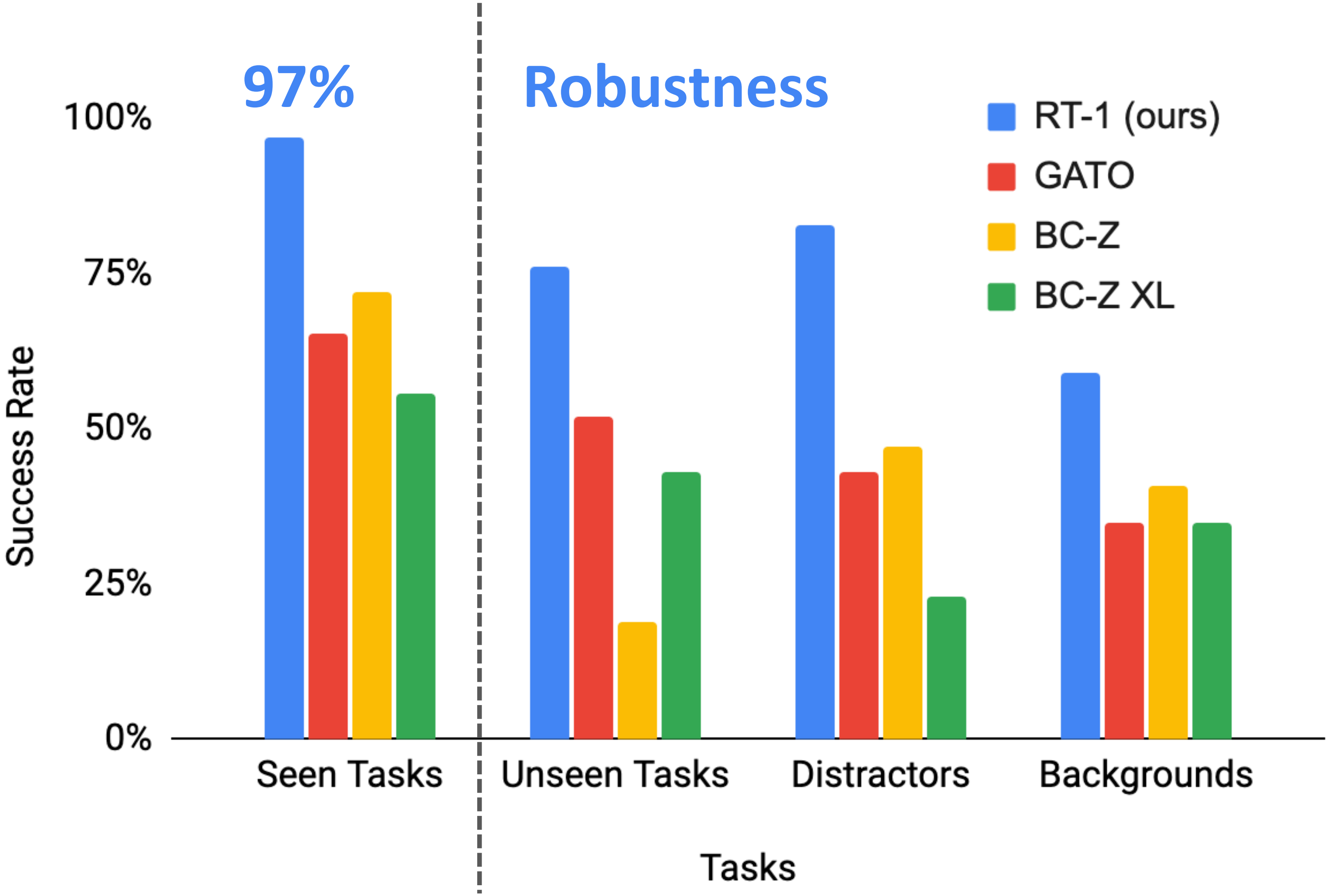
Deformable Object:



Instruction: Bring me the rice chips from the drawer.
Current step: bring it to you

RT-1

Performance



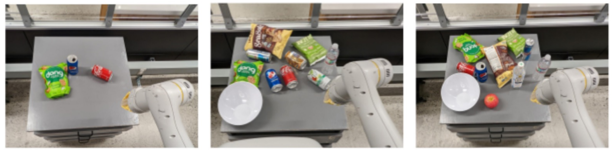
Unseen Tasks:

New Instructions

(Combination of Known Concepts)

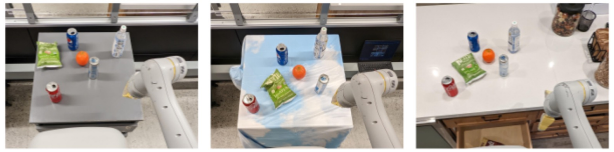
Distractors:

Distract Objects



Backgrounds:

New Environments



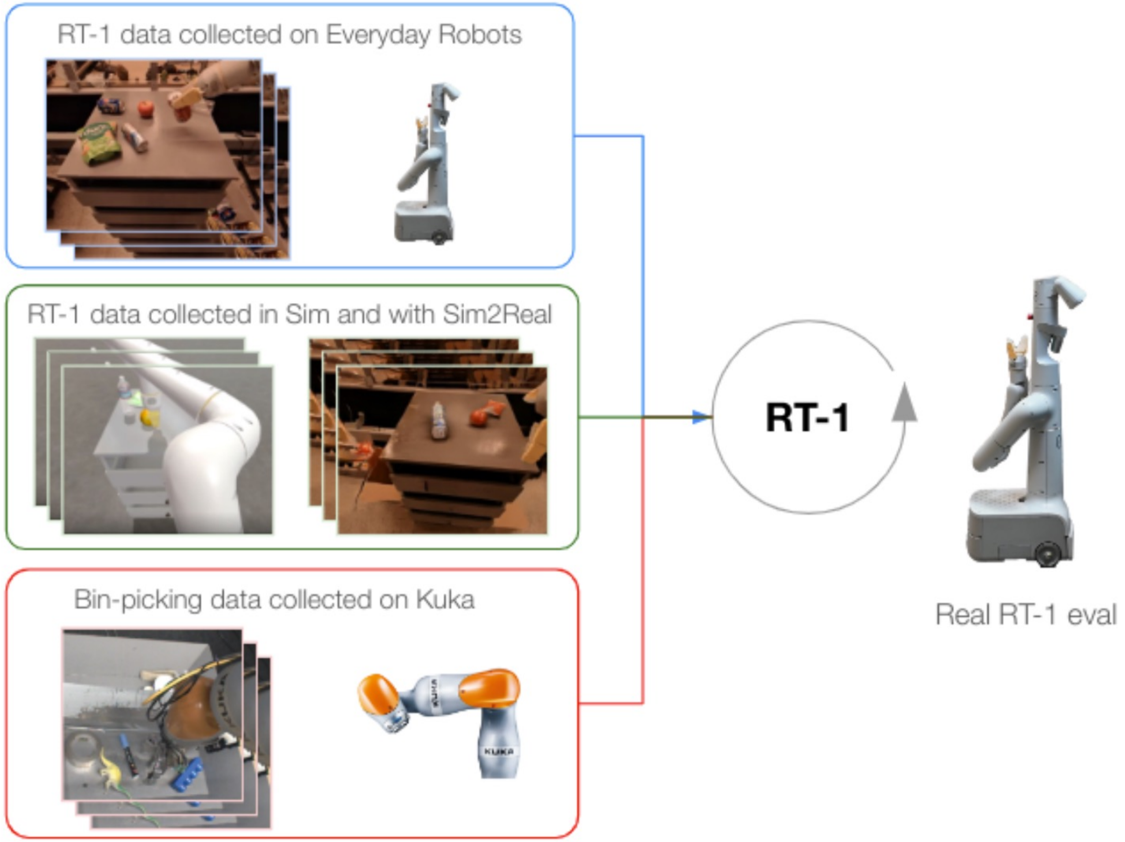
RT-1

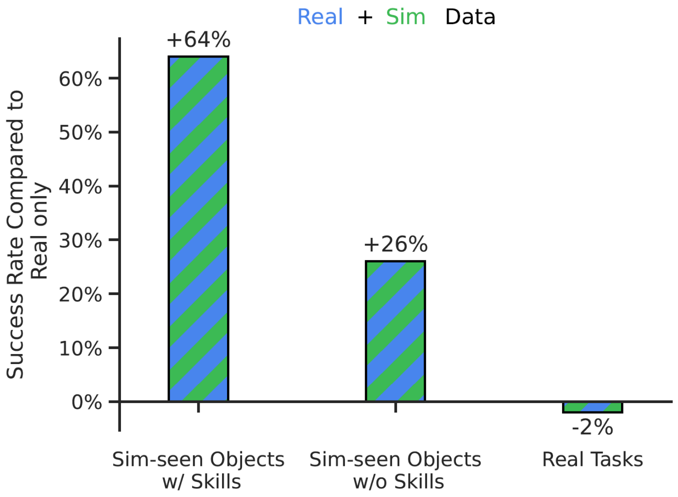
Generalizability for Data

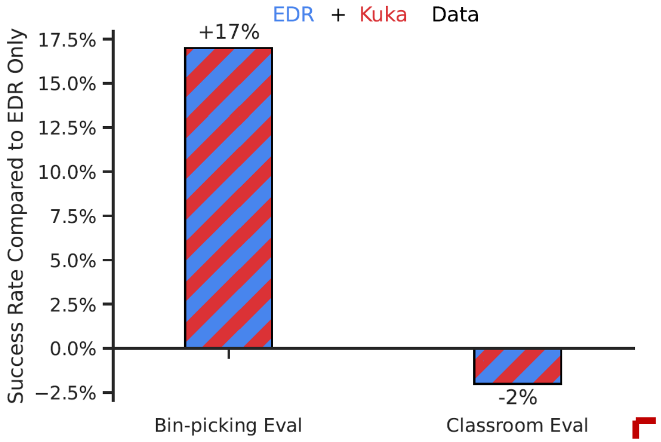
Origin Data

New Sim Data

New Robot Data (Padding Action Space)


RT-1 data collected on Everyday Robots
RT-1 data collected in Sim and with Sim2Real
Bin-picking data collected on Kuka
RT-1
Real RT-1 eval



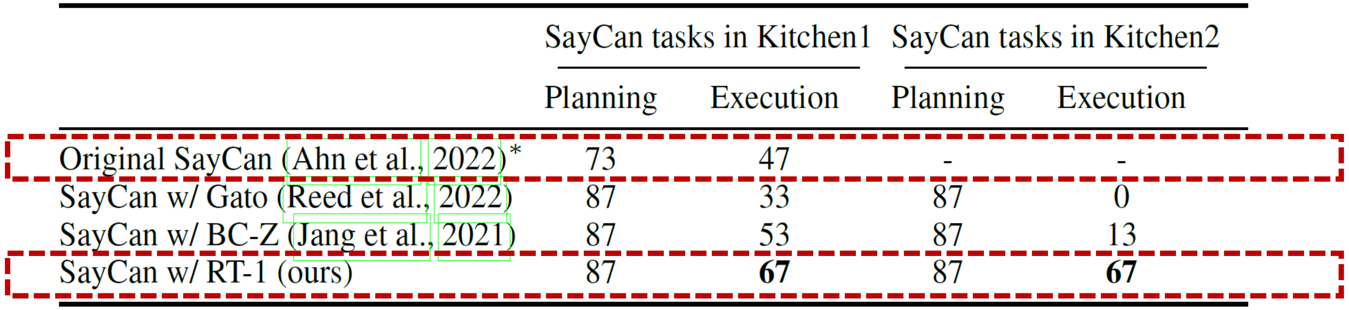


Models	Training Data	Classroom eval	Bin-picking eval
RT-1	Kuka bin-picking data + EDR data	90(-2)	**39(+17)**
RT-1	EDR only data	92	22
RT-1	Kuka bin-picking only data	0	0

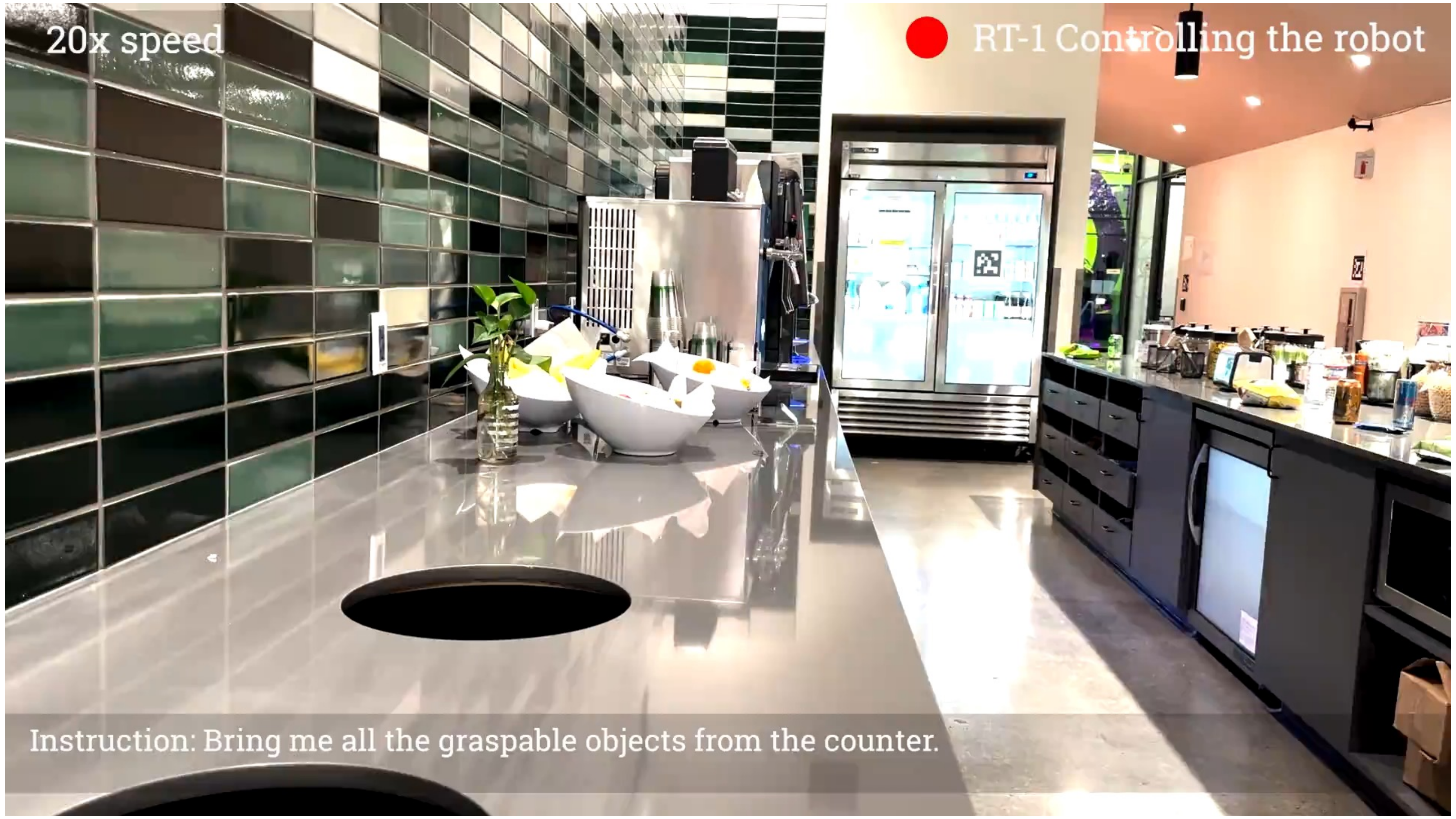
EDR: EveryDay Robot

RT-1

SayCan + RT-1



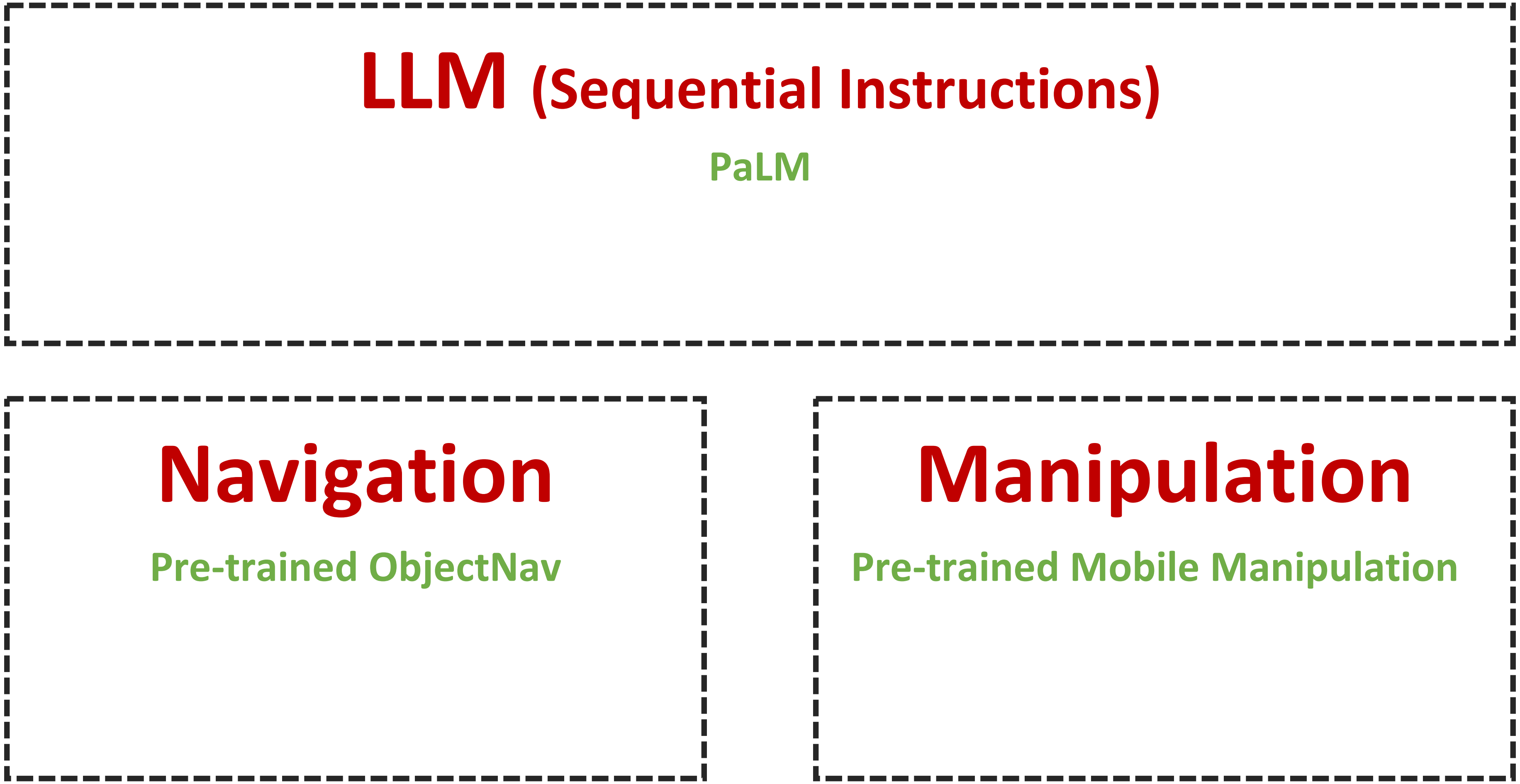
	SayCan tasks in Kitchen1		SayCan tasks in Kitchen2	
	Planning	Execution	Planning	Execution
Original SayCan (Ahn et al., 2022)*	73	47	-	-
SayCan w/ Gato (Reed et al., 2022)	87	33	87	0
SayCan w/ BC-Z (Jang et al., 2021)	87	53	87	13
SayCan w/ RT-1 (ours)	87	**67**	87	**67**



Instruction: Bring me all the graspable objects from the counter.

Where are we for now?

Say-Can



LLM (Sequential Instructions)

PaLM

Navigation

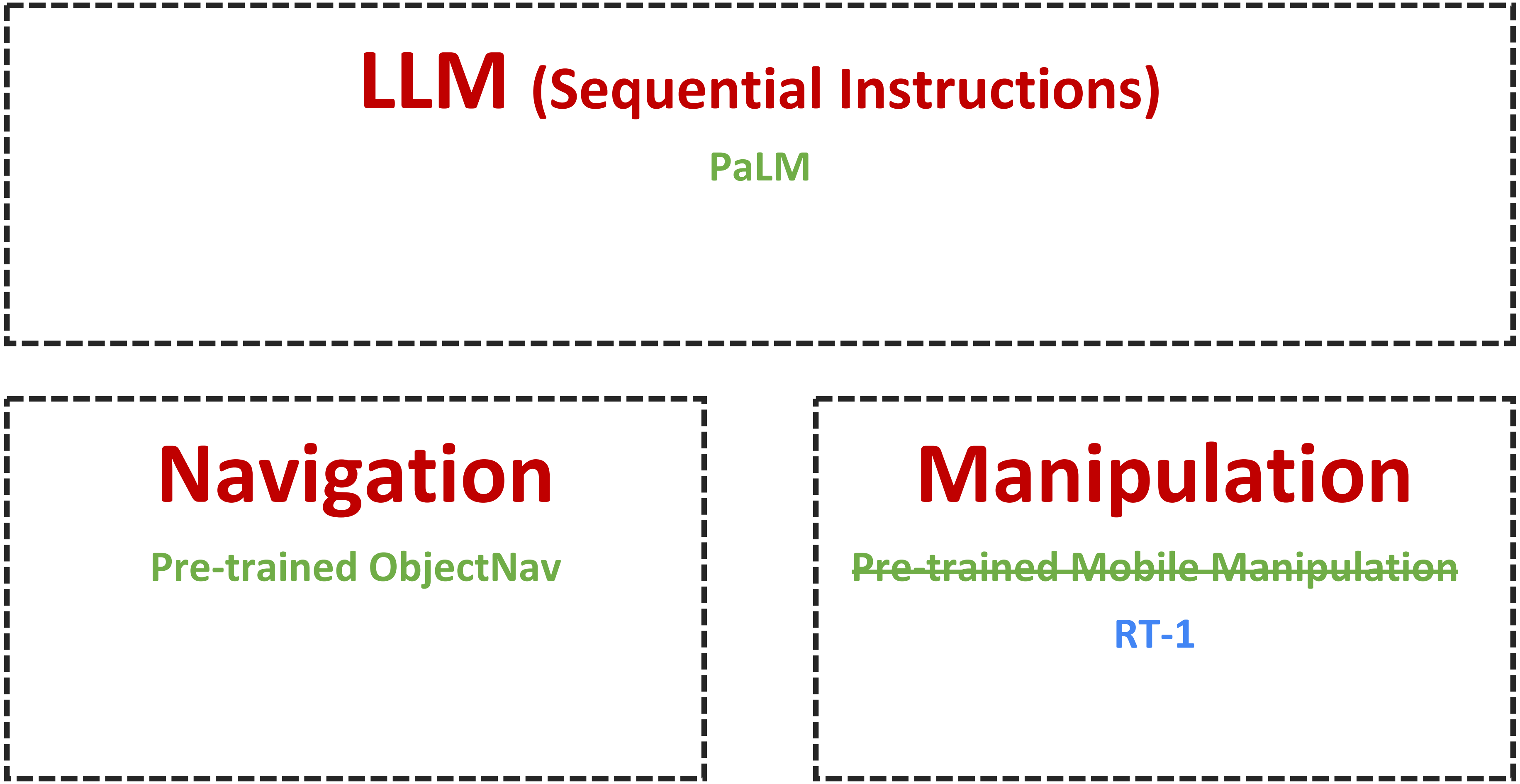
Pre-trained ObjectNav

Manipulation

Pre-trained Mobile Manipulation

Where are we for now?

Say-Can + RT-1



LLM (Sequential Instructions)

PaLM

Navigation

Pre-trained ObjectNav

Manipulation

~~Pre-trained Mobile Manipulation~~

RT-1

Where are we for now?

Say-Can + RT-1 + Open-World?

LLM (Sequential Instructions)

PaLM

Navigation

Pre-trained ObjectNav

Manipulation

~~Pre-trained Mobile Manipulation~~

RT-1

Where are we for now?

Say-Can + RT-1 + Open-World?

LLM (Sequential Instructions)

PaLM

Navigation

~~Pre-trained ObjectNav~~

Open-World ObjectNav
(CLIP on Wheels)

Manipulation

~~Pre-trained Mobile Manipulation~~

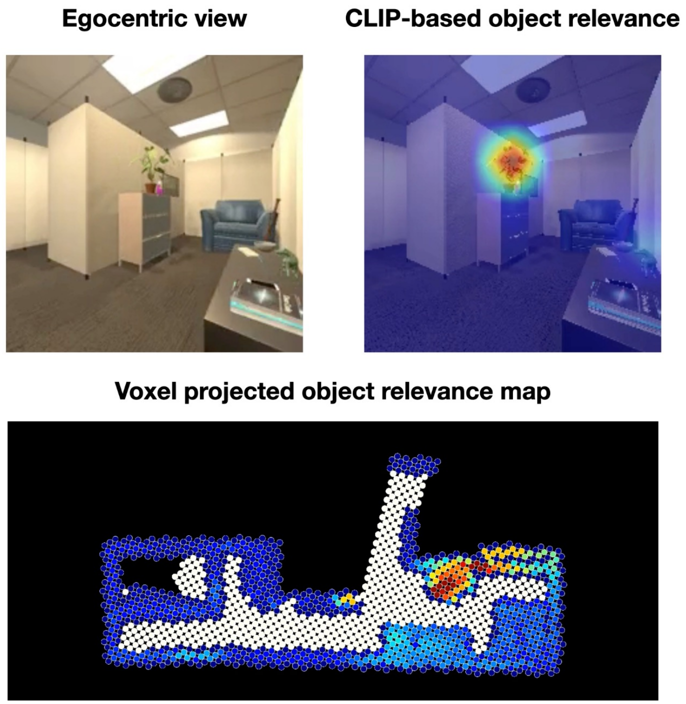
RT-1

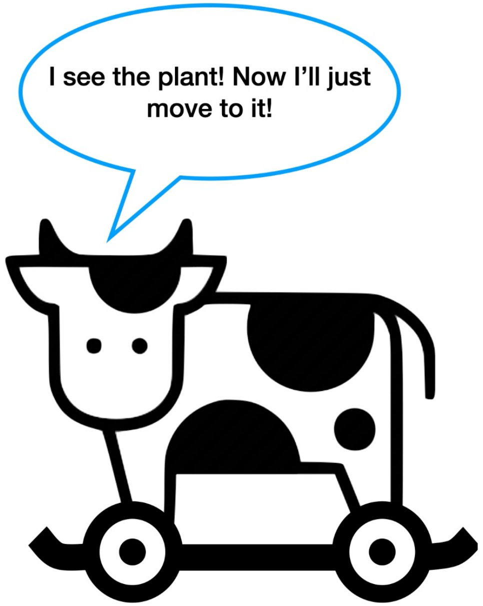
CoW (CLIP on Wheels)

CoWs on PASTURE: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation

(CVPR 2022)

Samir Yitzhak Gadre[◇] Mitchell Wortsman[†] Gabriel Ilharco[†] Ludwig Schmidt[†] Shuran Song[◇]

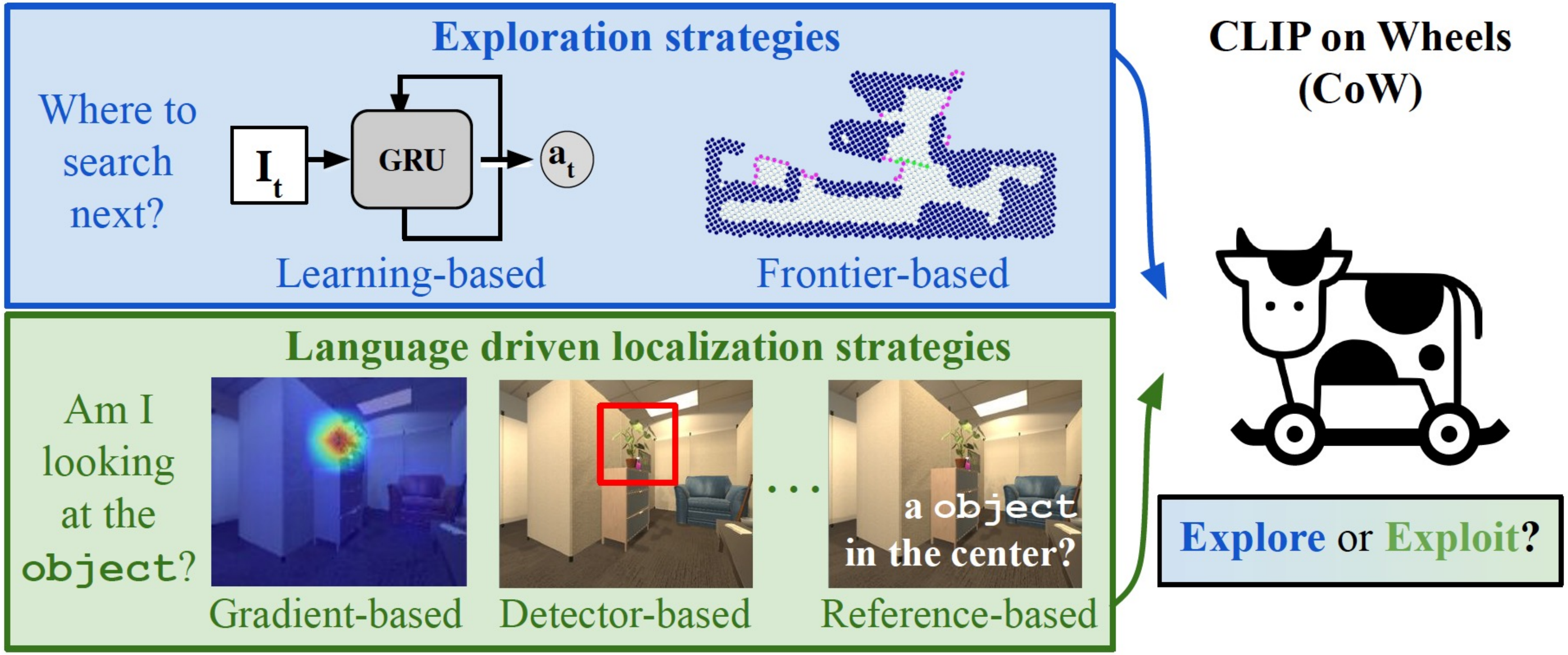

Egocentric view
CLIP-based object relevance
Voxel projected object relevance map



Target: Plant!

CoW (CLIP on Wheels)

METHOD



Exploration strategies

Where to search next?

I_t → GRU → a_t

Learning-based

Frontier-based

Language driven localization strategies

Am I looking at the object?

Gradient-based

Detector-based

. . .

a object in the center?

Reference-based

CLIP on Wheels (CoW)

Explore or Exploit?

Where are we for now?

Say-Can + RT-1 + Open-World?

LLM (Sequential Instructions)

PaLM

Navigation

~~Pre-trained ObjectNav~~

Open-World ObjectNav
(CLIP on Wheels)

Manipulation

~~Pre-trained Mobile Manipulation~~

RT-1

Where are we for now?

Say-Can + RT-1 + Open-World?

LLM (Sequential Instructions)

PaLM

Navigation

~~Pre-trained ObjectNav~~

Open-World ObjectNav
(CLIP on Wheels)

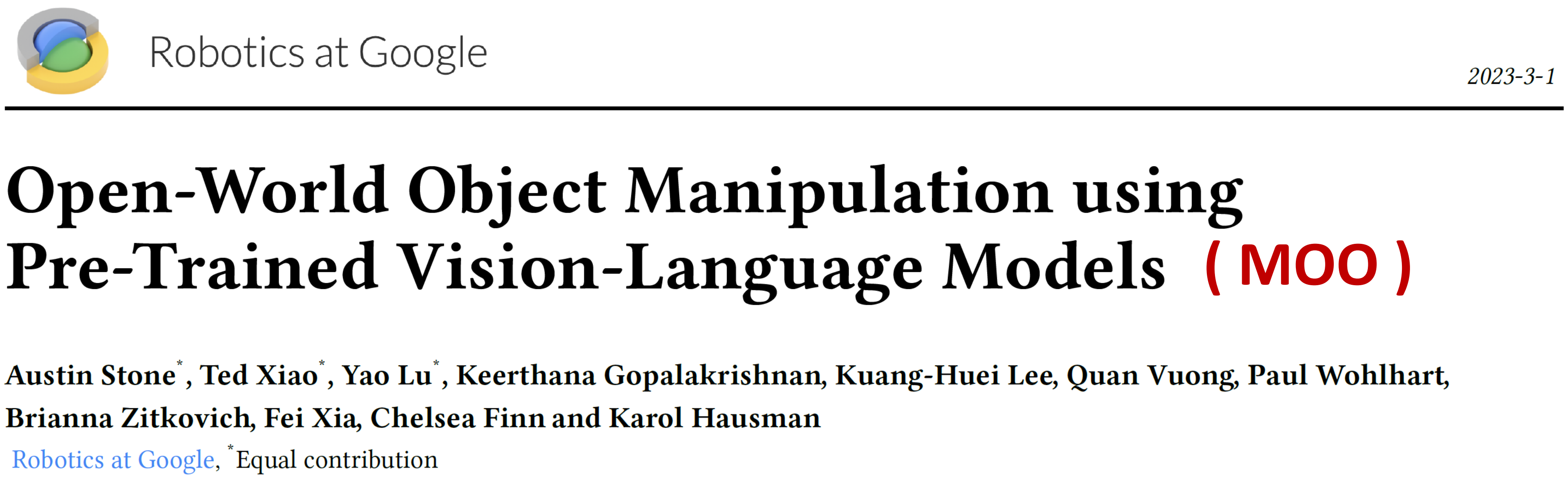
Manipulation

~~Pre-trained Mobile Manipulation~~

~~RT-1~~

Open-World RT-1 -> MOO

MOO



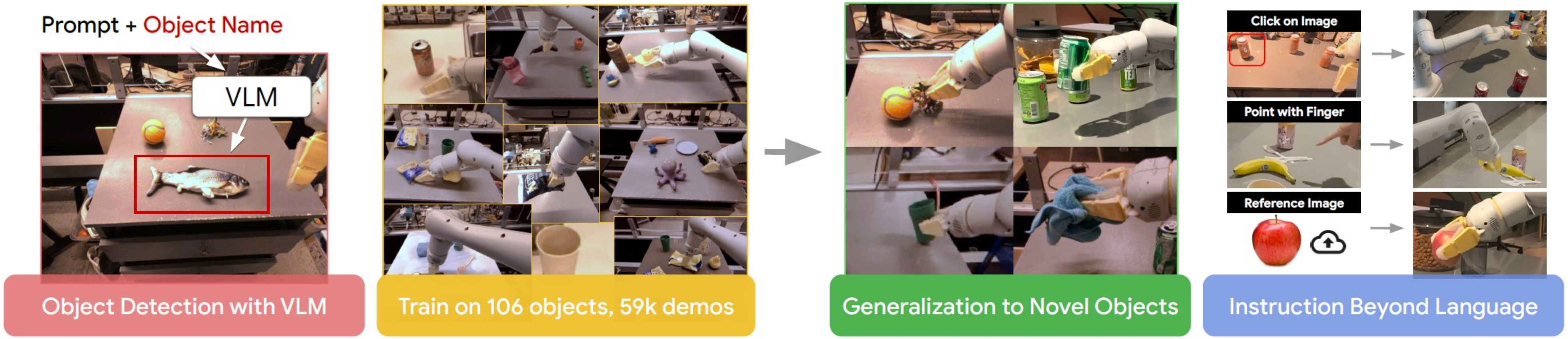
Robotics at Google

2023-3-1

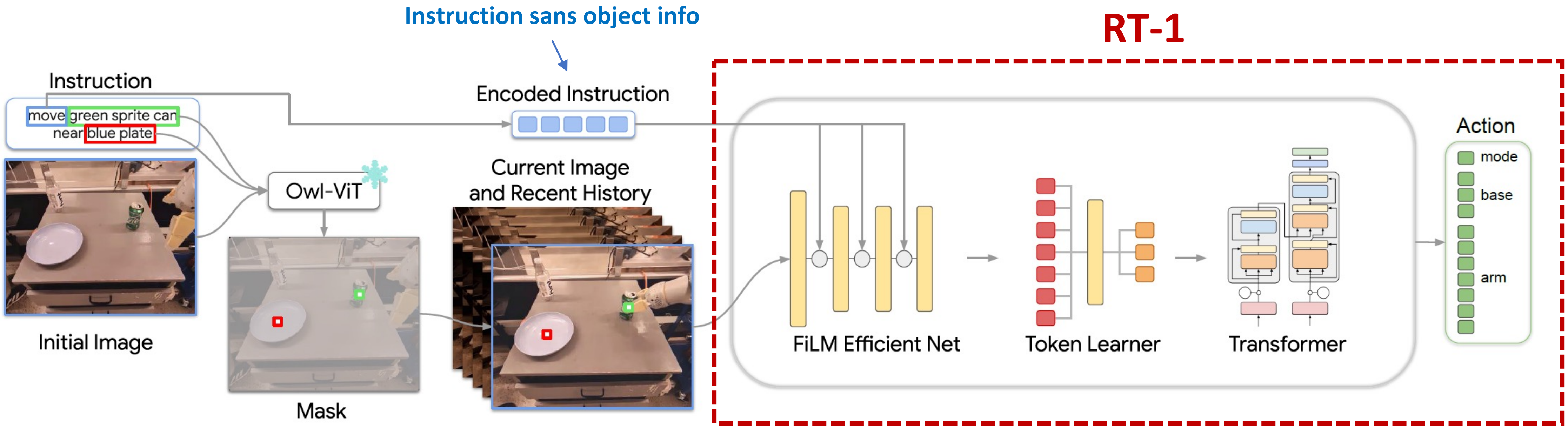
Open-World Object Manipulation using Pre-Trained Vision-Language Models (MOO)

Austin Stone*, Ted Xiao*, Yao Lu*, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Brianna Zitkovich, Fei Xia, Chelsea Finn and Karol Hausman

Robotics at Google, *Equal contribution


Prompt + Object Name
VLM
Object Detection with VLM
Train on 106 objects, 59k demos
Generalization to Novel Objects
Click on Image
Point with Finger
Reference Image
Instruction Beyond Language

METHOD

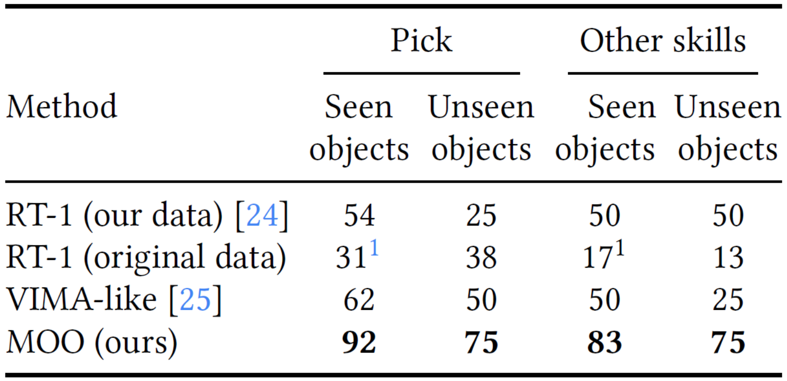


VLM for Open-World Object Detection

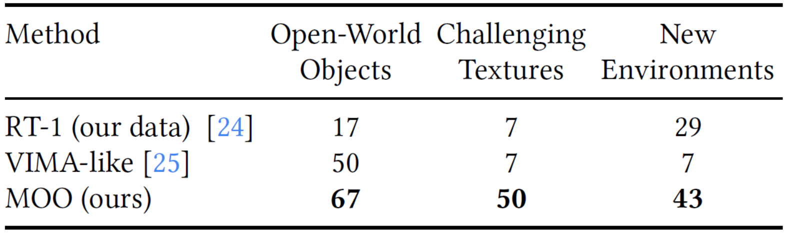
Data : Original RT-1 data (16 objects) + New Human demos for 90 new objects

MOO

Performance

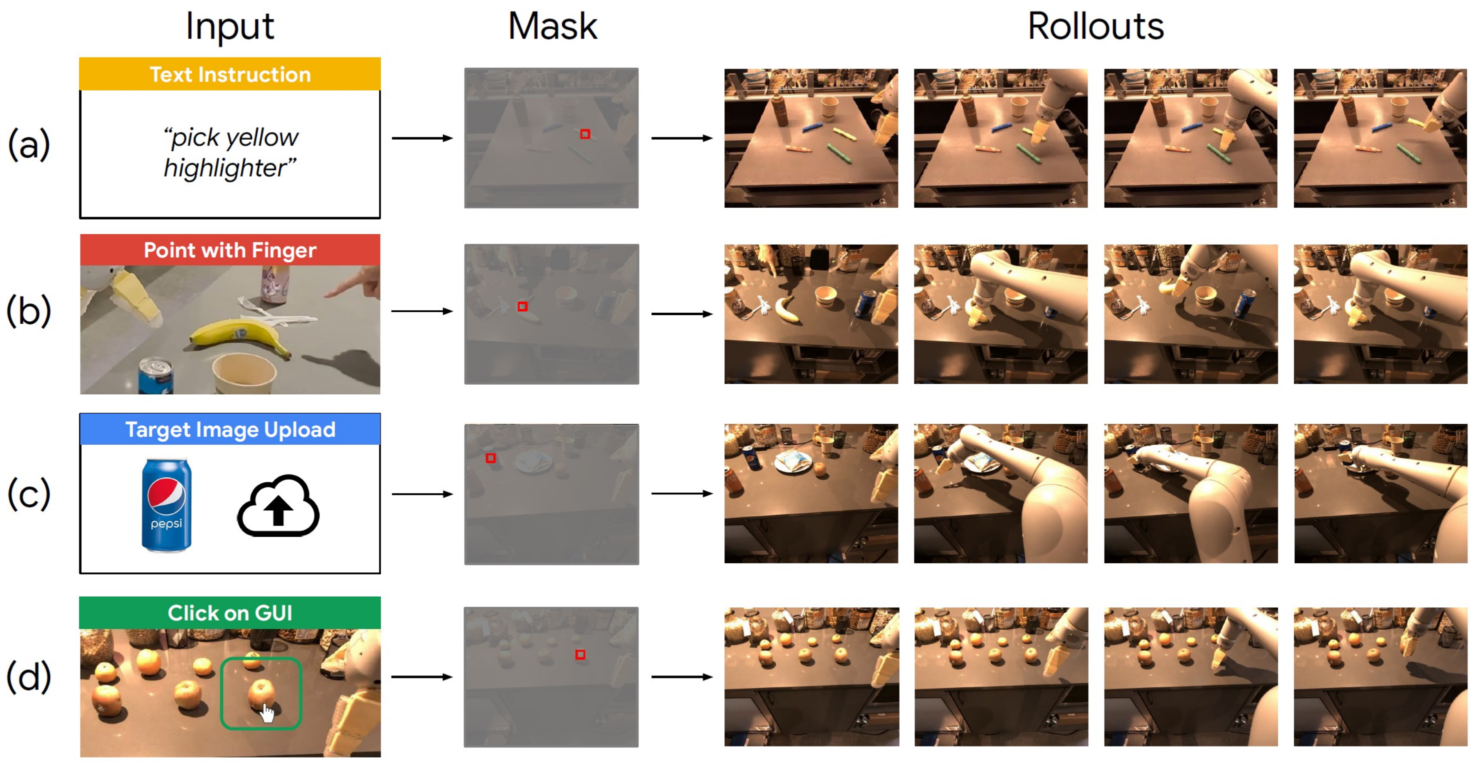


Method	Pick		Other skills	
	Seen objects	Unseen objects	Seen objects	Unseen objects
RT-1 (our data) [24]	54	25	50	50
RT-1 (original data)	31[1]	38	17[1]	13
VIMA-like [25]	62	50	50	25
MOO (ours)	**92**	**75**	**83**	**75**



Method	Open-World Objects	Challenging Textures	New Environments
RT-1 (our data) [24]	17	7	29
VIMA-like [25]	50	7	7
MOO (ours)	**67**	**50**	**43**

New Modality



MOO

MOO + CoW (CLIP on Wheels):

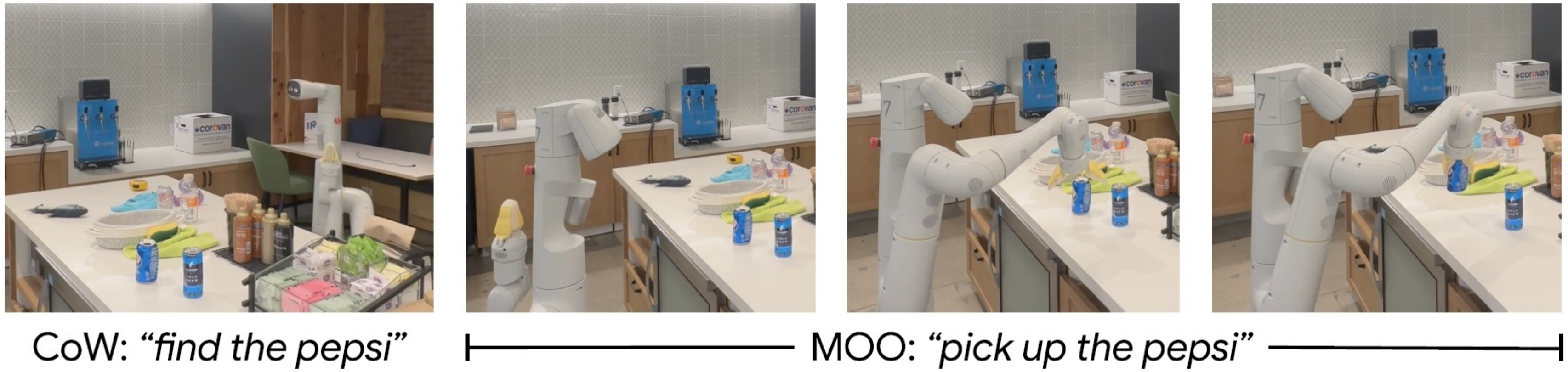
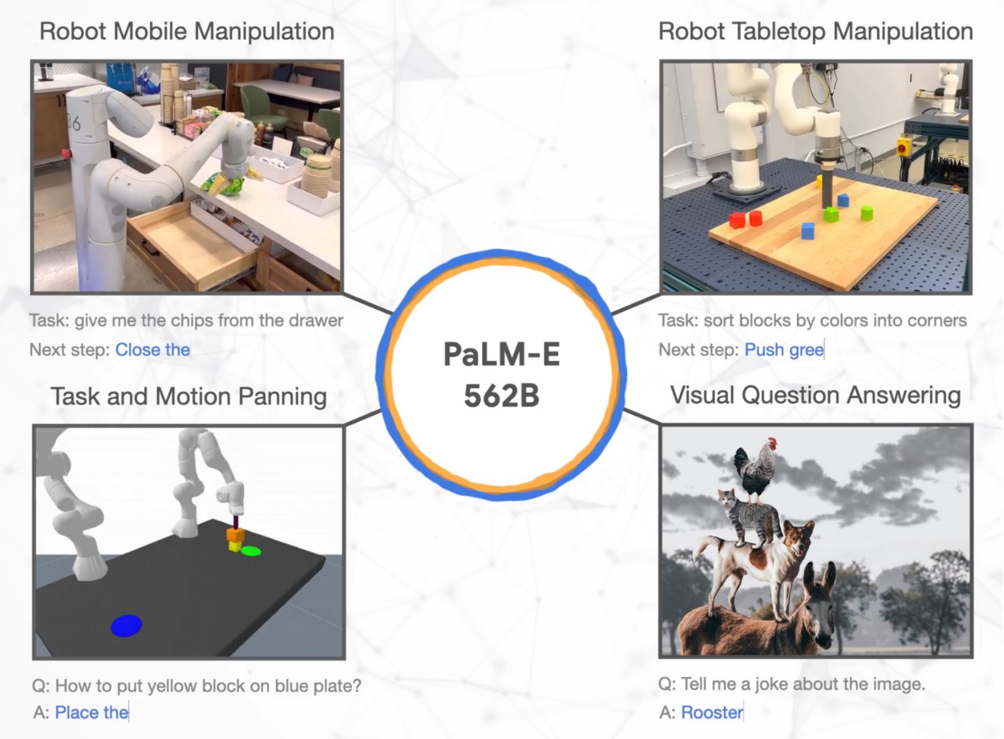


Figure 10: We present CoW-MOO, a system that combines an open-vocabulary object navigation by CoW [54] with open-world manipulation by MOO. Full videos are shown on the project’s website.

PaLM-E

PaLM-E: An Embodied Multimodal Language Model

Danny Driess[1,2] Fei Xia[1] Mehdi S. M. Sajjadi[3] Corey Lynch[1] Aakanksha Chowdhery[3]
Brian Ichter[1] Ayzaan Wahid[1] Jonathan Tompson[1] Quan Vuong[1] Tianhe Yu[1] Wenlong Huang[1]
Yevgen Chebotar[1] Pierre Sermanet[1] Daniel Duckworth[3] Sergey Levine[1] Vincent Vanhoucke[1]
Karol Hausman[1] Marc Toussaint[2] Klaus Greff[3] Andy Zeng[1] Igor Mordatch[3] Pete Florence[1]


Robot Mobile Manipulation
Robot Tabletop Manipulation
PaLM-E
562B
Task and Motion Panning
Visual Question Answering

Task: give me the chips from the drawer
Next step: Close the

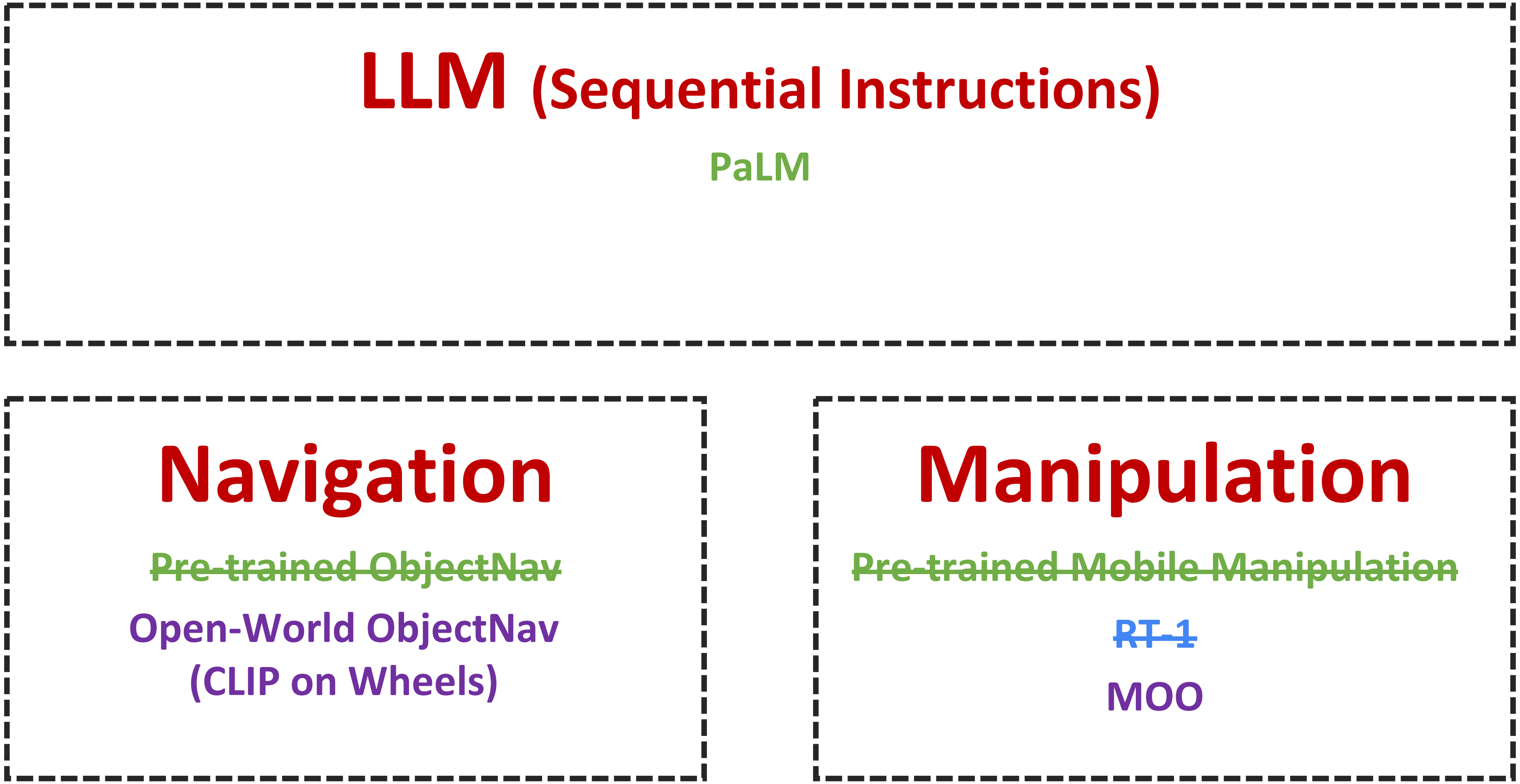
Task: sort blocks by colors into corners
Next step: Push gree

Q: How to put yellow block on blue plate?
A: Place the

Q: Tell me a joke about the image.
A: Rooster

Where are we for now?

Say-Can + RT-1 + Open-World?



LLM (Sequential Instructions)

PaLM

Navigation

~~Pre-trained ObjectNav~~

Open-World ObjectNav
(CLIP on Wheels)

Manipulation

~~Pre-trained Mobile Manipulation~~

~~RT-1~~

MOO

Where are we for now?

~~Say-Can~~ + RT-1 + Open-World + PaLM-E

LLM (Sequential Instructions)

~~PaLM~~

PaLM-E

Navigation

~~Pre-trained ObjectNav~~

Open-World ObjectNav
(CLIP on Wheels)

Manipulation

~~Pre-trained Mobile Manipulation~~

~~RT-1~~

MOO

Embodied Mobile Manipulation

Thanks for Listening!

Any Questions?