RT-1: ROBOTICS TRANSFORMER FOR REAL-WORLD CONTROL AT SCALE

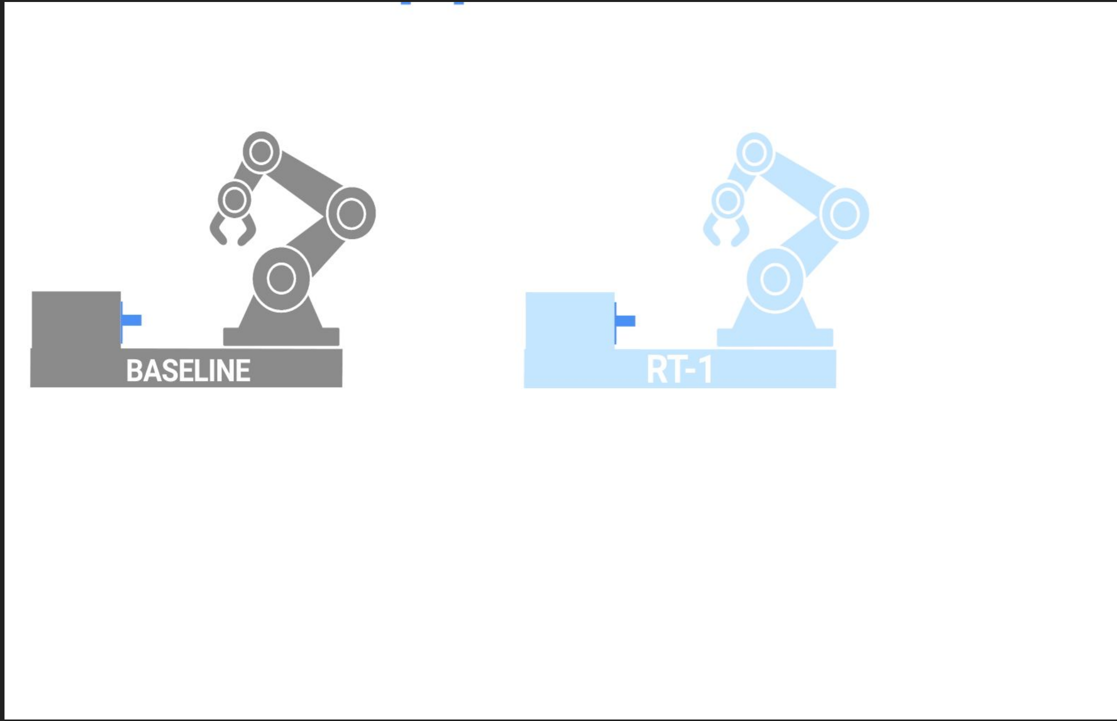
Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, 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

Presented by: Nathan Holmes, Pan Lu

Motivation

Single, multi-task backbone model

- Generalization
  - Zero-shot generalization
- Performance
  - Inference Time



Courtesy of:
https://blog.research.google/2022/12/rt-1-robotics-transformer-for-real.html

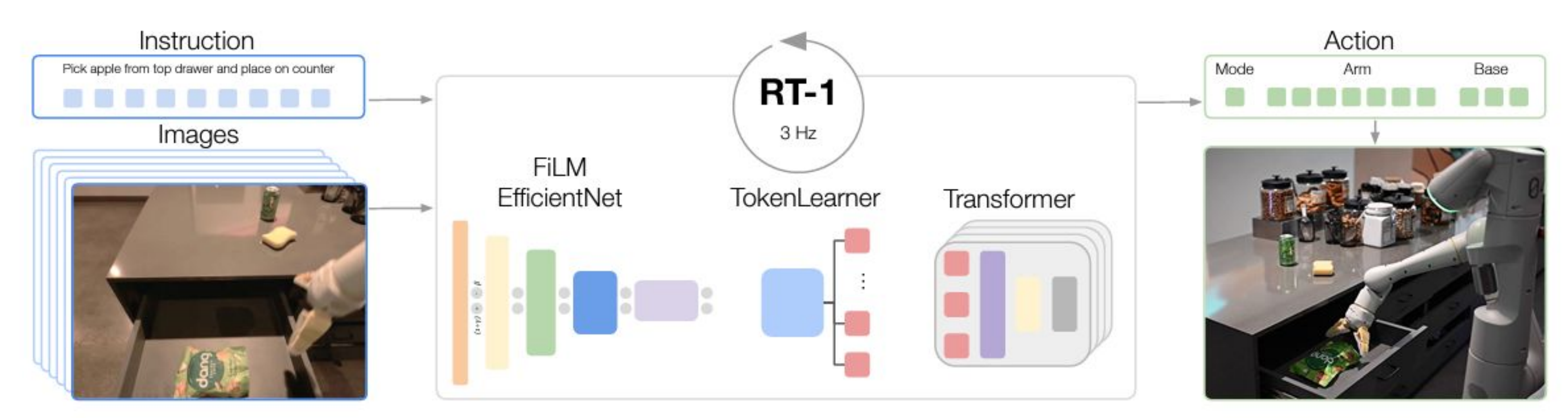
Key Components

- Robot Learning
- Imitation Learning
- Correct scope of training data
  - Scale
  - Breadth
- High capacity, real-time inference
  - Image Tokenization
  - Action Tokenization
  - Token Compression

Courtesy of: https://community.libretranslate.com/t/rt-1-robotics-transformer/441

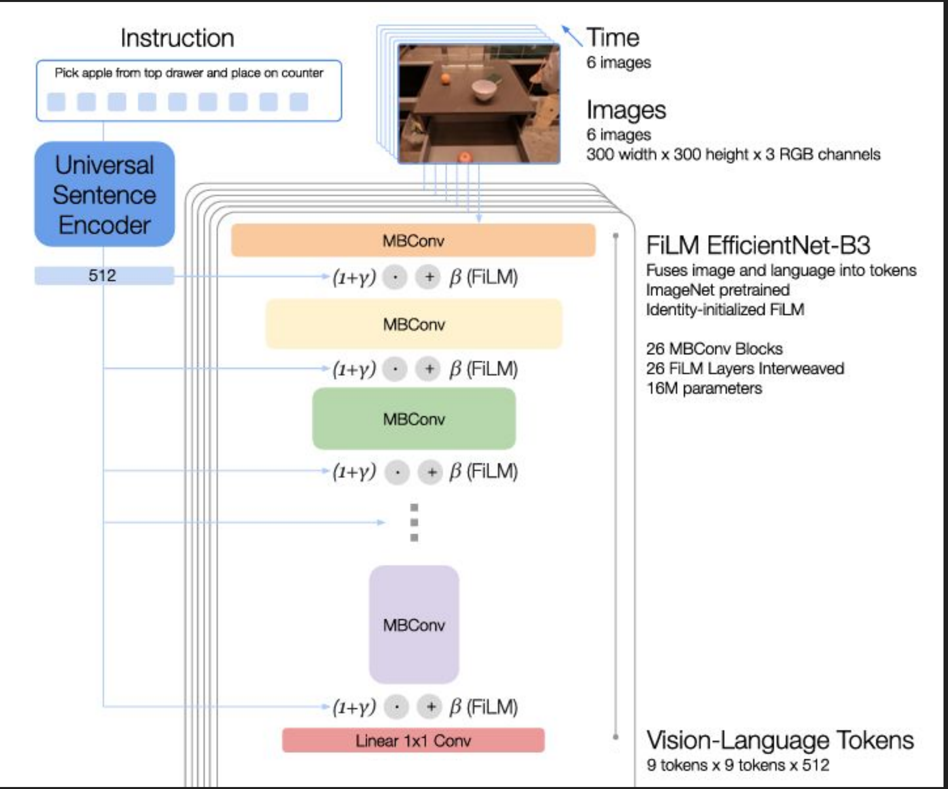
Architecture

- FiLM Conditioned EfficentNet
- TokenLearner
- Transformer



Architecture: FiLM Conditioned EfficientNet

- Input: 6 images, 300×300 resolution
- Fuse image and instruction into tokens
  - Pretrained on ImageNet
- Output: 9×9×512 spatial feature map

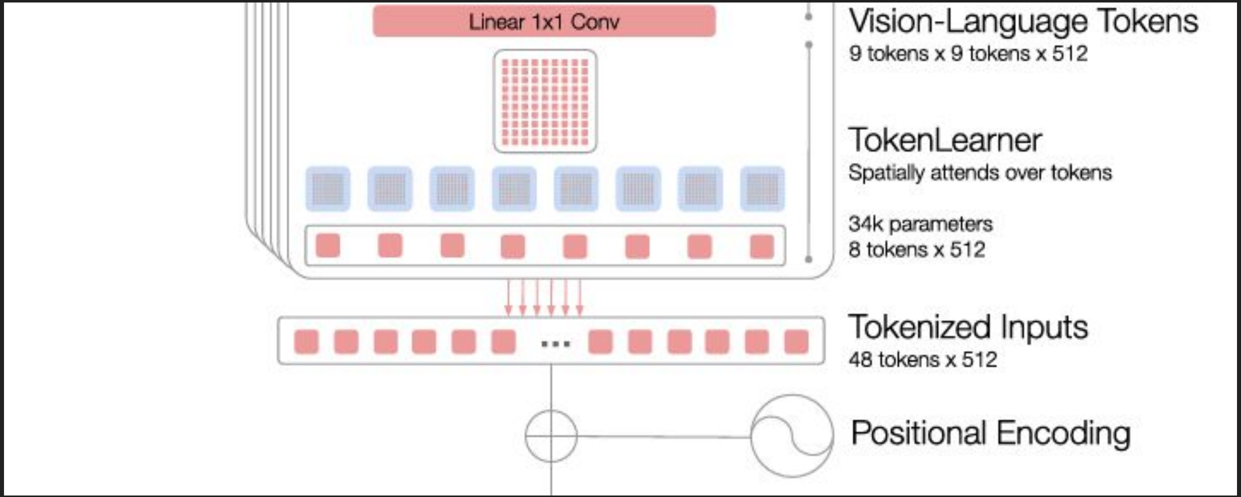


Architecture: TokenLearner

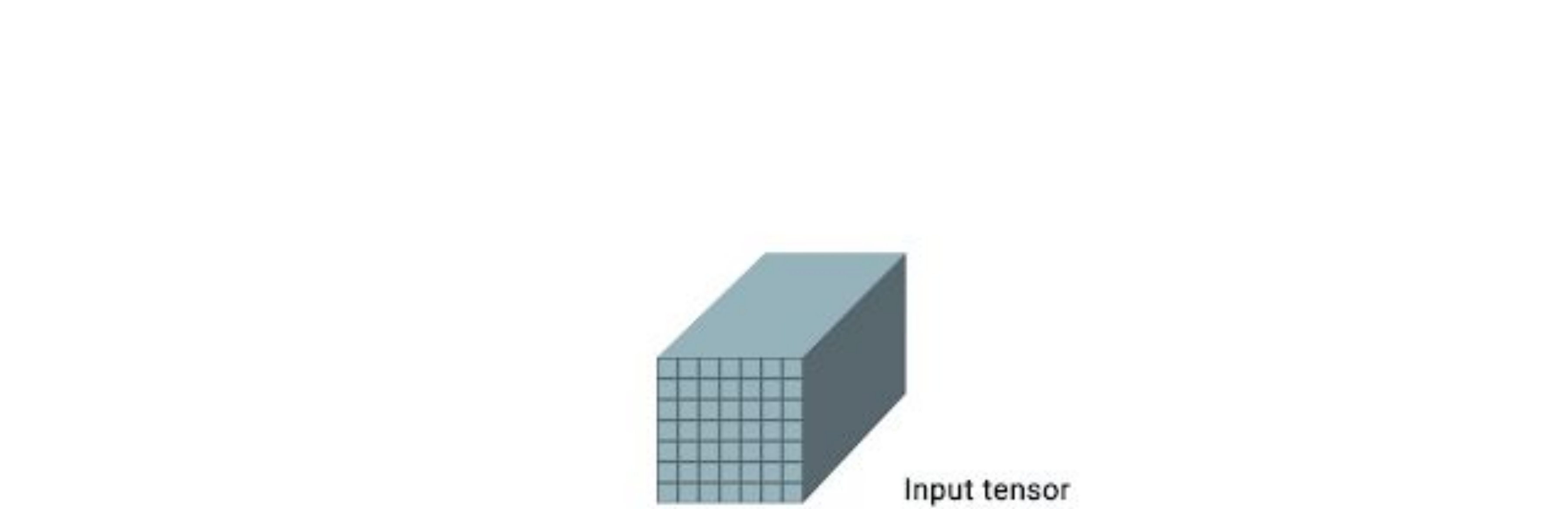
- Input: 9 x 9 x 512 Spatial Map

  = 81 visual tokens

- Element-wise attention model compresses tokens
- Output: 8 visual tokens per image



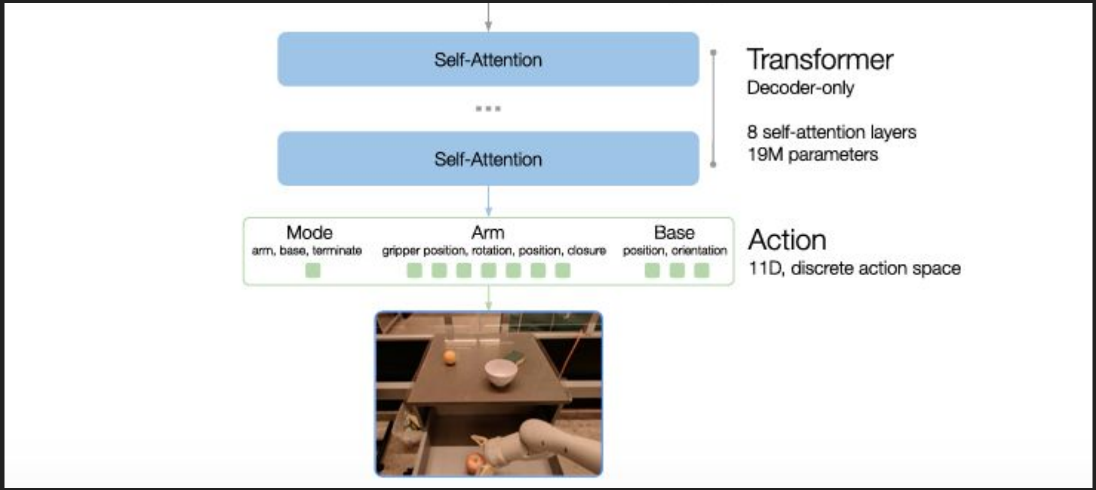
Architecture: TokenLearner



Courtesy of https://blog.research.google/2021/12/improving-vision-transformer-efficiency.html

Architecture: Transformer

- Input: 8 tokens per-image x 6 images = 48 total tokens
  - Added position encoding
  - Fed into the Transformer
- Transformer is a **decoder-only sequence model**
  - 8 self-attention layers
  - 19M total parameters
- Output: Action tokens

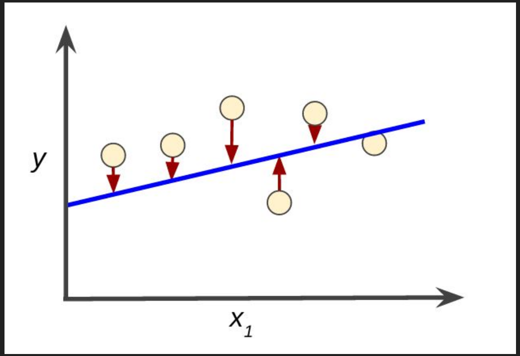


Action Tokens

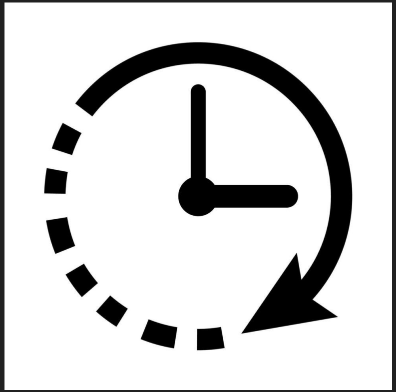
- 7 variables for arm movement
  - x, y, z, roll, pitch, yaw, gripper opening
- 3 variables for base movement
  - x, y, yaw
- Extra variable to switch between three modes:
  - controlling arm
  - controlling base
  - terminating the episode
- Each action dimension is discretized into 256 bins
  - 11 variables x 256 bins

Other Architectural Components

- Loss function:
  - Standard categorical cross-entropy entropy objective
    - Classification
  - Causal masking
    - Predictions conditioned on preceding elements



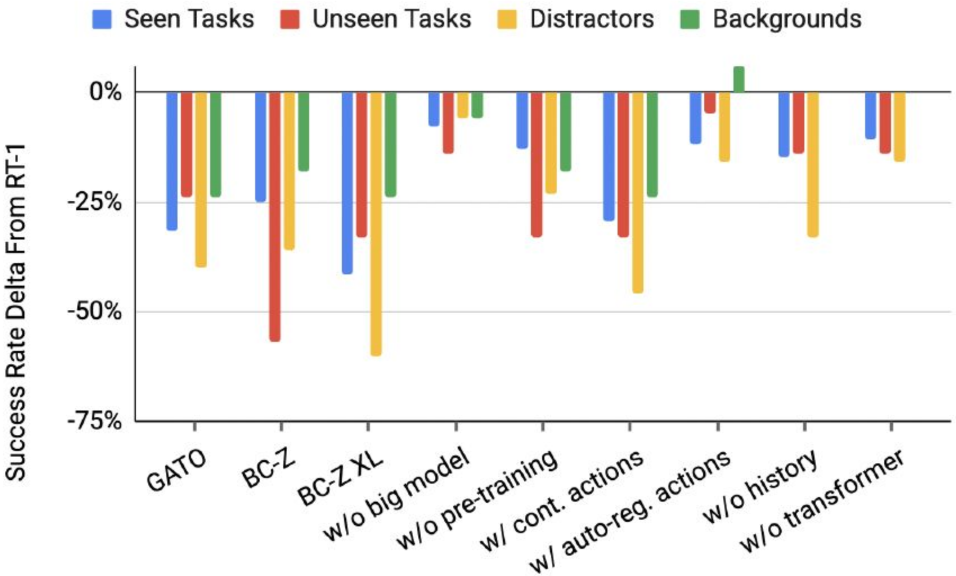
- Inference Speed Limitations:
  - Human speeds of 2-4 seconds
  - 100ms inference time
  - At least 3Hz control frequency (rate)



Model Ablations

- Justifies current architectural choices

Model	Seen Tasks	Unseen Tasks	Distractors				Backgrounds	Inference Time (ms)
			All	Easy	Medium	Hard	All	
Gato (Reed et al., 2022)	65 (-32)	52 (-24)	43 (-40)	71	44	29	35 (-24)	129
BC-Z (Jang et al., 2021)	72 (-25)	19 (-57)	47 (-36)	100	67	7	41 (-18)	5.3
BC-Z XL	56 (-41)	43 (-33)	23 (-60)	57	33	0	35 (-24)	5.9
RT-1 (ours)	**97**	**76**	**83**	100	100	64	**59**	15
RT-1 w/o big model	89 (-8)	62 (-14)	77 (-6)	100	100	50	53 (-6)	13.5
RT-1 w/o pre-training	84 (-13)	43 (-33)	60 (-23)	100	67	36	41 (-18)	15
RT-1 w/ continuous actions	68 (-29)	43 (-33)	37 (-46)	71	67	0	35 (-24)	16
RT-1 w/ auto-regressive actions	85 (-12)	71 (-5)	67 (-16)	100	78	43	**65 (+6)**	36
RT-1 w/o history	82 (-15)	62 (-14)	50 (-33)	71	89	14	**59 (+0)**	15
RT-1 w/o Transformer	86 (-13)	62 (-14)	67 (-16)	100	100	29	**59 (+0)**	26



Data

- Our primary dataset consists of ~130k robot demonstrations, collected with a fleet of 13 robots over the course of 17 months
- Definitions of Instructions and skills
  - Instruction(aka tasks): a verb surrounded by one or multiple
    - Eg. “place water bottle upright”
  - Skill: instructions grouped by the verbs

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		

Table 1: The list of skills collected for RT-1 together with their descriptions and example instructions.

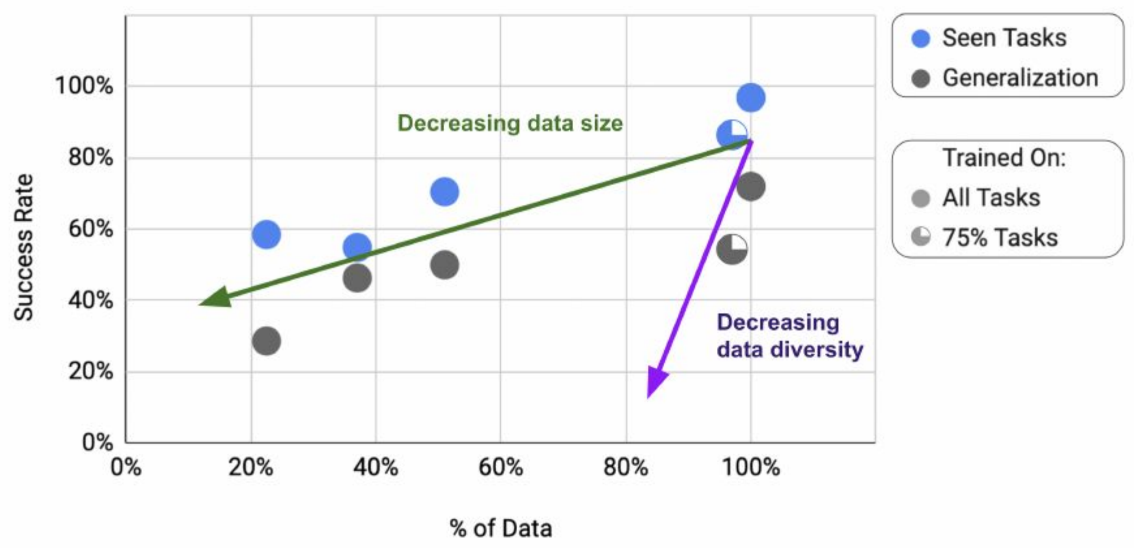
(e)

(f)

Data Ablations

- Success impacted more by data diversity than data size

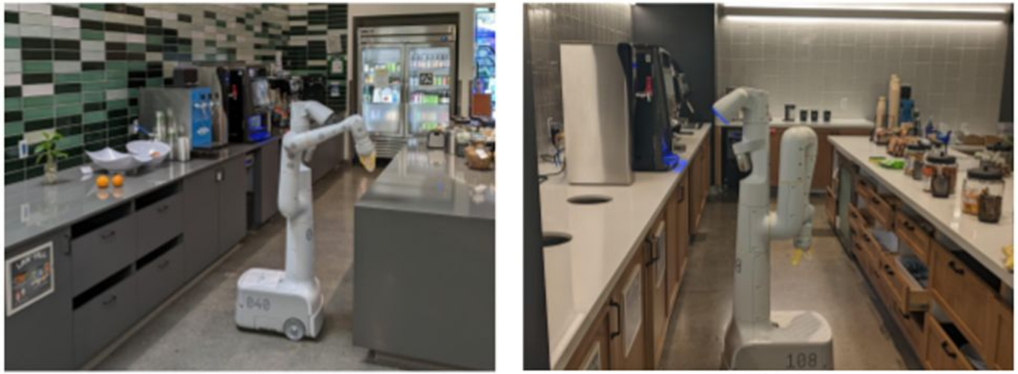
Models	% Tasks	% Data	Seen Tasks	Generalization All	Unseen Tasks	Distractors	Backgrounds
Smaller Data							
RT-1 (ours)	100	100	97	73	76	83	59
RT-1	100	51	71	50	52	39	59
RT-1	100	37	55	46	57	35	47
RT-1	100	22	59	29	14	31	41
Narrower Data							
RT-1 (ours)	100	100	97	73	76	83	59
RT-1	75	97	86	54	67	42	53



Experiments — Experiment Setup

- Equipment:
  - Mobile manipulators from Everyday Robot
- Environments:
  - Two real office kitchens
  - A training environment modelled off these real kitchens

(d)



(b) (c)

(a)

Experiments — Experiment Setup

- Evaluate Performance on Seen instructions
  - Evaluate performance on instructions sampled from the training set
    - Still involves varying the placement of objects and other factors of the setup (e.g., time of day, robot position)

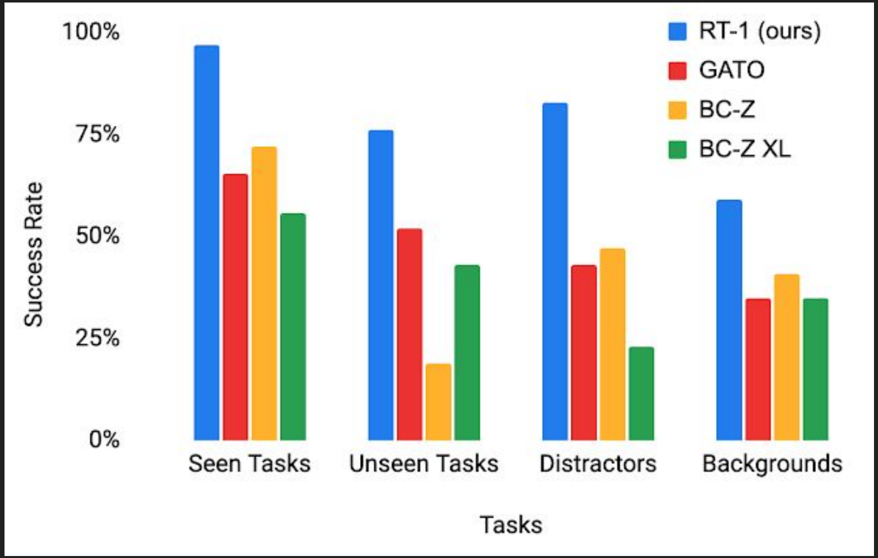
  - Test over 200 tasks in this evaluation in all
    - 36 for picking
    - 35 for knocking objects
    - 35 for placing things upright
    - 48 for moving objects
    - 18 for opening and closing various drawers
    - 36 for picking out of and placing objects into drawers

Experiments — Experiment Setup

- Evaluate generalization to unseen tasks
    - Test 53 novel, unseen instructions
    - Instructions are distributed across skills and objects
    - Eg. if “pick up the apple” is held out, then there are other training instructions that include the apple.
- Evaluate robustness
    - Perform 30 real-world tasks for distractor robustness
    - Perform 22 tasks for background robustness
- Evaluate generalization long-horizon scenarios
    - Require executing a sequence of skills
    - New tasks, objects, environments
    - Eg. “Bring me two different sodas”

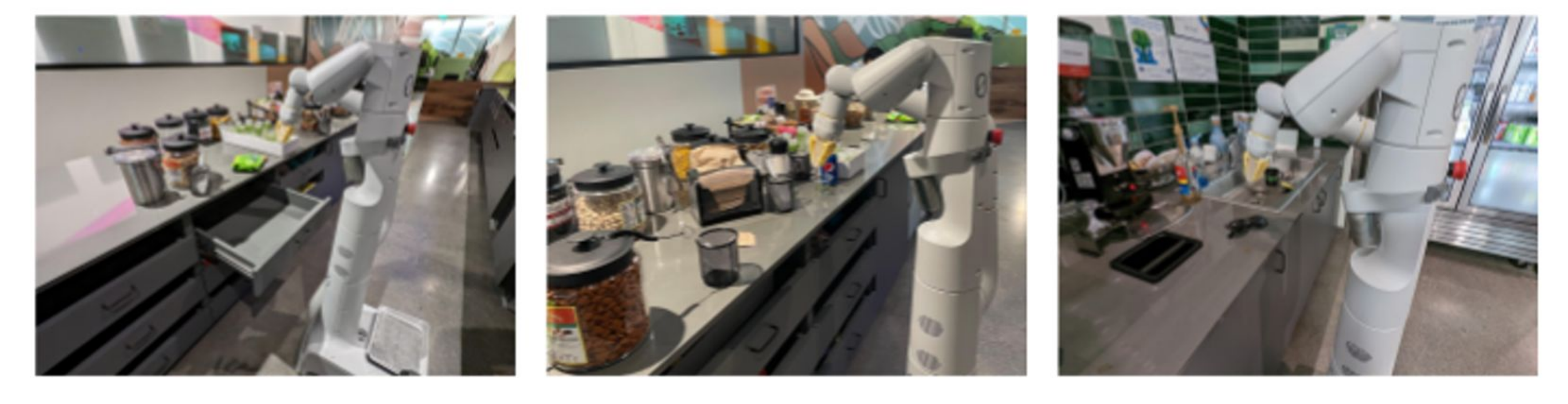
Results — CAN RT-1 LEARN TO PERFORM A LARGE NUMBER OF INSTRUCTIONS, AND TO GENERALIZE TO NEW TASKS, OBJECTS AND ENVIRONMENTS?

Model	Seen Tasks	Unseen Tasks	Distractors	Backgrounds
Gato (Reed et al., 2022)	65	52	43	35
BC-Z (Jang et al., 2021)	72	19	47	41
BC-Z XL	56	43	23	35
RT-1 (ours)	**97**	**76**	**83**	**59**



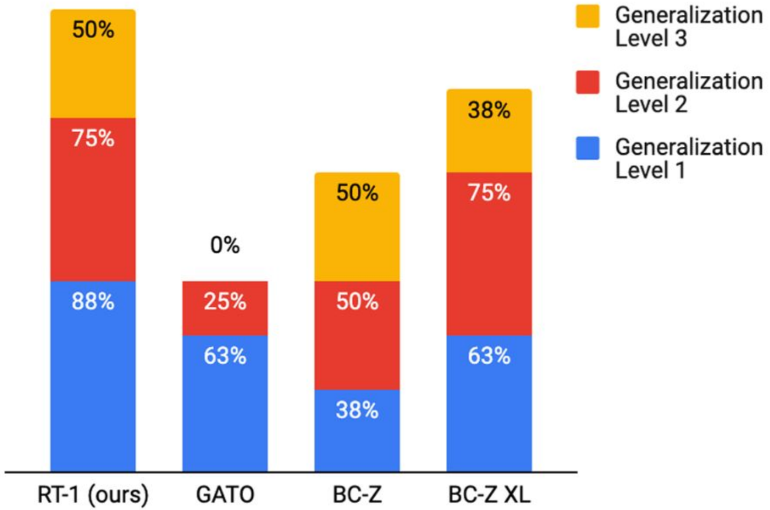
Results — Generalization to realistic instructions

- L1 for generalization to the new counter-top layout and lighting conditions
- L2 for additionally generalization to unseen distractor objects
- L3 for additionally generalization to drastically new task settings, new task objects or in unseen locations like near a sink.



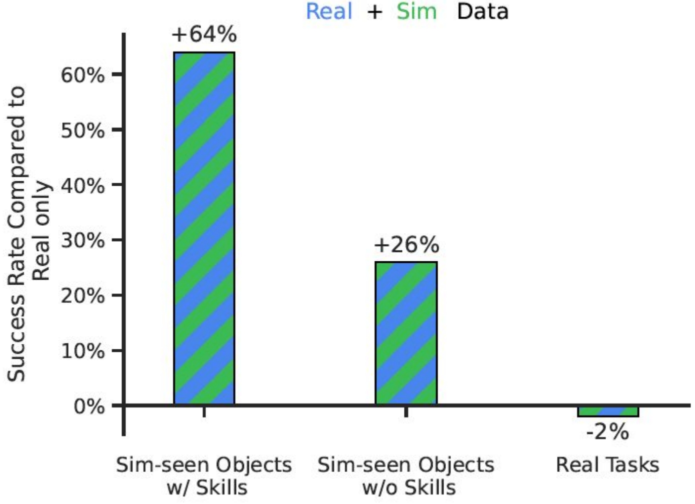
Results — Generalization to realistic instructions

		Generalization Scenario Levels		
Models	All	L1	L2	L3
Gato Reed et al. (2022)	30	63	25	0
BC-Z Jang et al. (2021)	45	38	50	**50**
BC-Z XL	55	63	**75**	38
RT-1 (ours)	**70**	**88**	**75**	**50**



Results — CAN WE PUSH THE RESULTING MODEL FURTHER BY INCORPORATING HETEROGENEOUS DATA SOURCES?

		Real Objects	Sim Objects (not seen in real)	
Models	Training Data	Seen Skill w/ Objects	Seen Skill w/ Objects	Unseen Skill w/ Objects
RT-1	Real Only	92	23	7
RT-1	Real + Sim	90(-2)	**87(+64)**	**33(+26)**



Results — CAN WE PUSH THE RESULTING MODEL FURTHER BY INCORPORATING HETEROGENEOUS DATA FROM DIFFERENT ROBOTS?

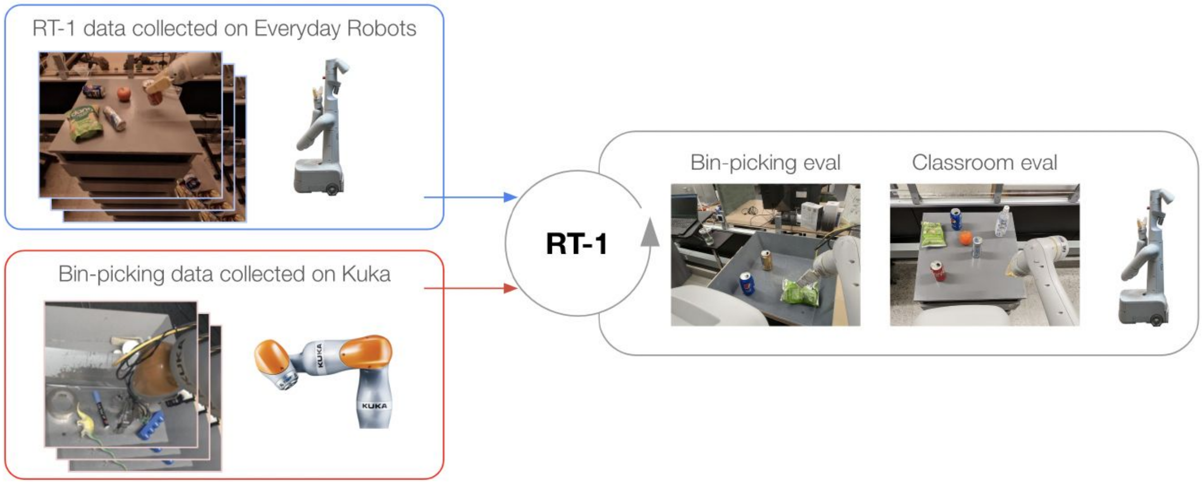
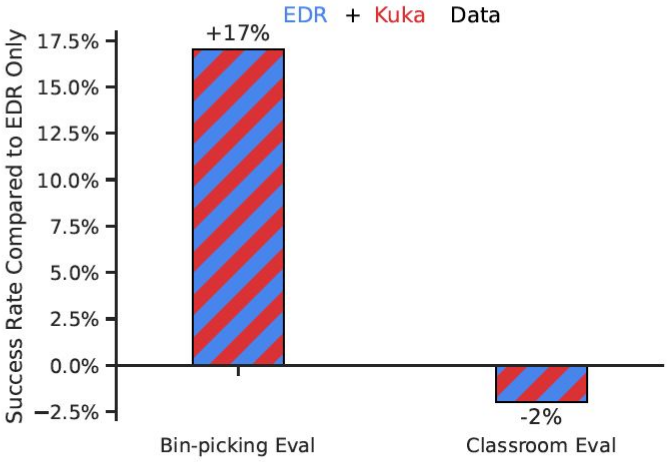


Figure 6: In Table 5, RT-1 is trained with data from two robotics platforms and learns to generalize across them.

Results — CAN WE PUSH THE RESULTING MODEL FURTHER BY INCORPORATING HETEROGENEOUS DATA FROM DIFFERENT ROBOTS?

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



Results — HOW DO VARIOUS METHODS GENERALIZE LONG-HORIZON ROBOTIC SCENARIOS

	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**

Table 6: SayCan style long horizon tasks in Kitchen1 and Kitchen2. (*Original SayCan eval uses a slightly different prompt so the planning success rate is lower.)

Limitations

- Unable to surpass the performance of the demonstrators
- Unable to generalize to a completely new motion that has not been seen before
- Presented on a large but not very dexterous set of manipulation tasks.

Discussion

- Single, multi-task backbone model
- Showed improvements in generalization
  - Unseen tasks, distractors, backgrounds
- Future goals:
  - Faster scaling of robot skills
  - Improve performance on backgrounds
  - New motions