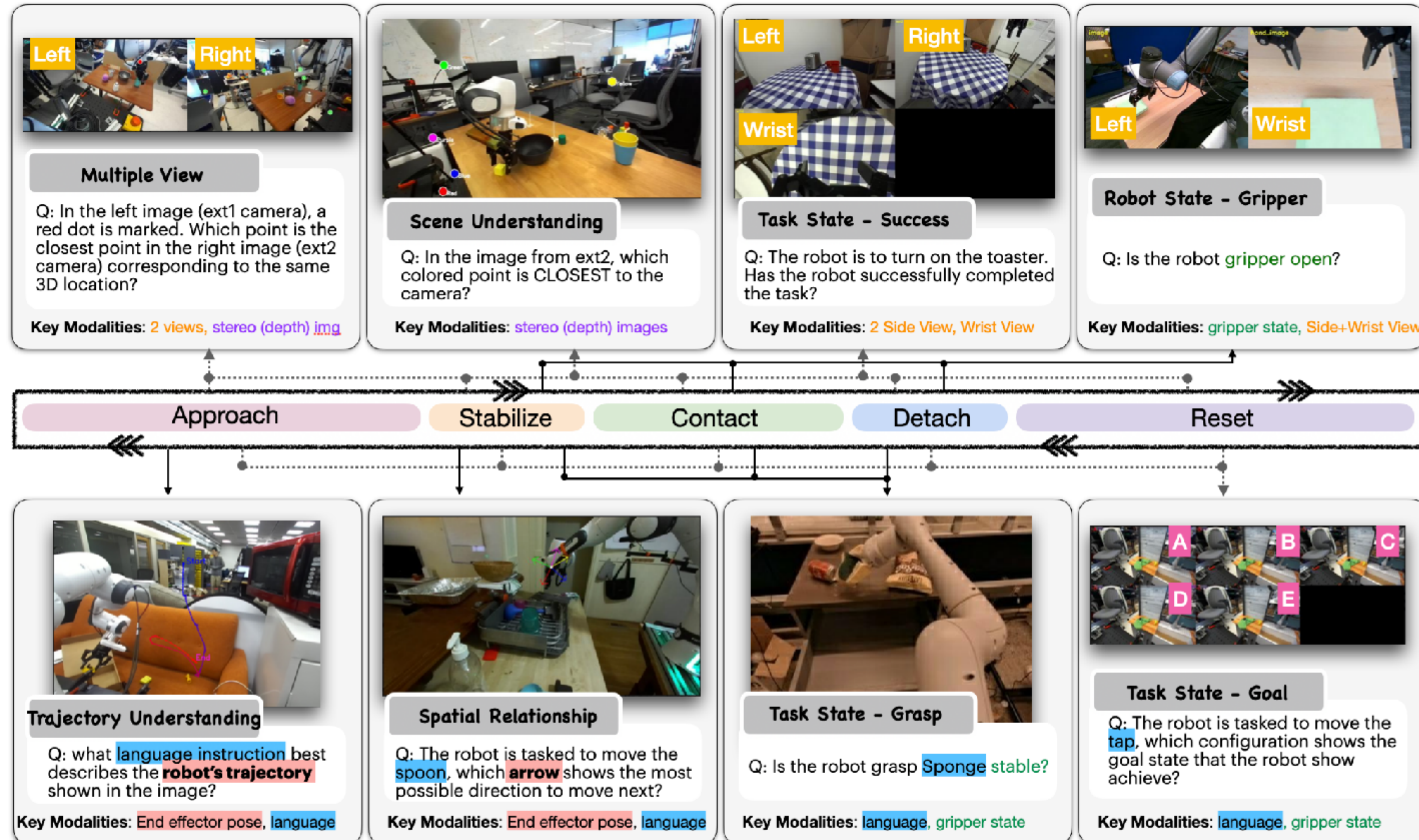


Robo2VLM: Visual Question Answering from Large-Scale In-the-Wild Robot Manipulation Datasets



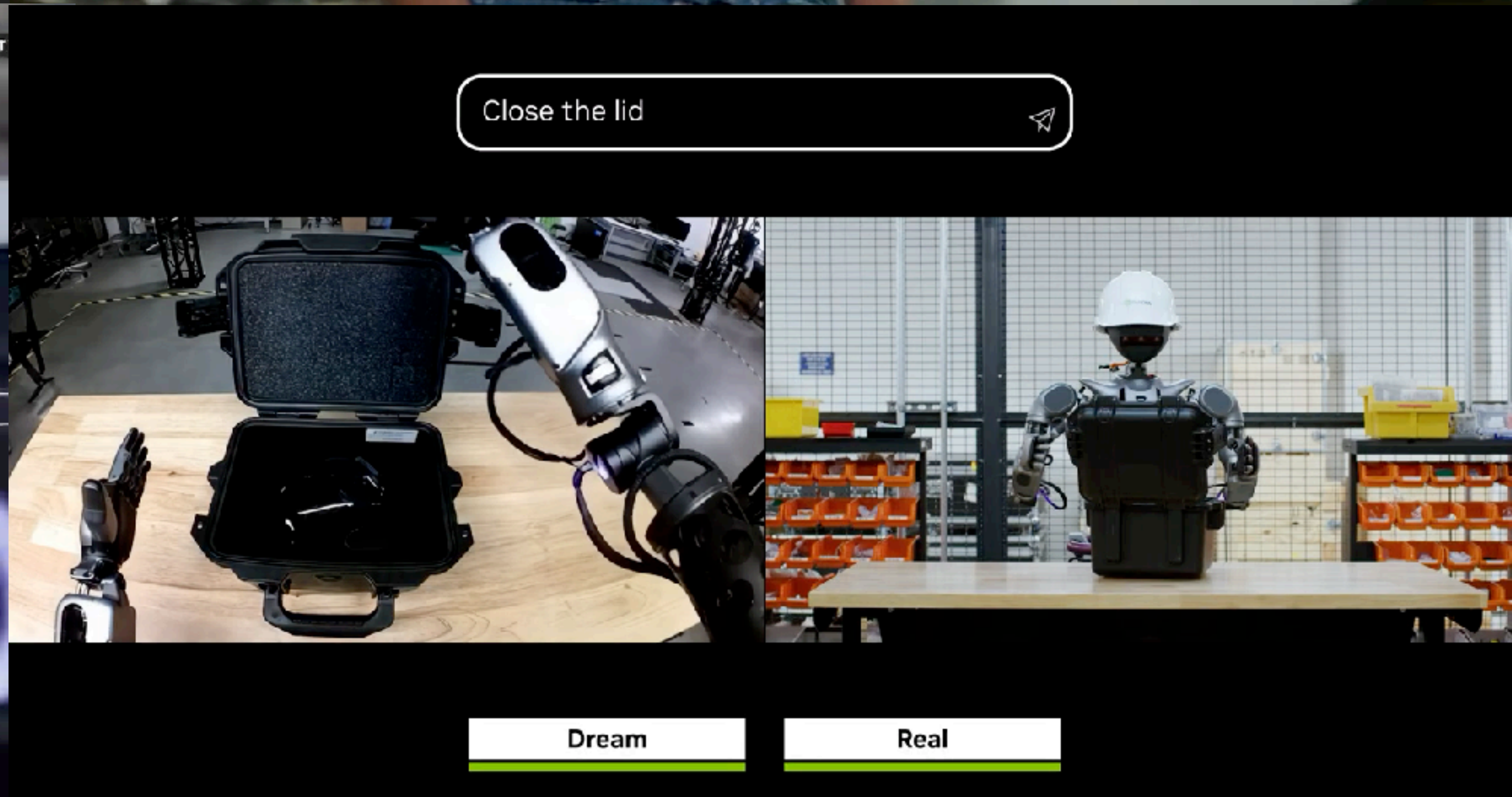
Website:



Robo2VLM: Visual Question Answering from Large-Scale In-the- Wild Robot Manipulation Datasets

Kaiyuan Chen*, Shuangyu Xie*, Zehan Ma, Pannag Sanketi, Ken Goldberg





**How can we evaluate and enhance
spatial intelligence of VLMs?**

Robo2VLM

🔥 Features

- 684,710 VQA questions from 176K real robot trajectories
- 463 distinct scenes across diverse environments (office, lab, kitchen)
- 3,396 manipulation tasks with ground-truth from robot sensors
- Multi-modal reasoning using spatial, goal-conditioned, and interaction templates

Multi-modality Real Data



176k Manipulator Episode
from 463 Real Scenes

Robo2VLM

Scene-interaction Understanding

Semantic Segmentation

Manipulation Phase Classification

Object info,
current phase

Keyframe selection

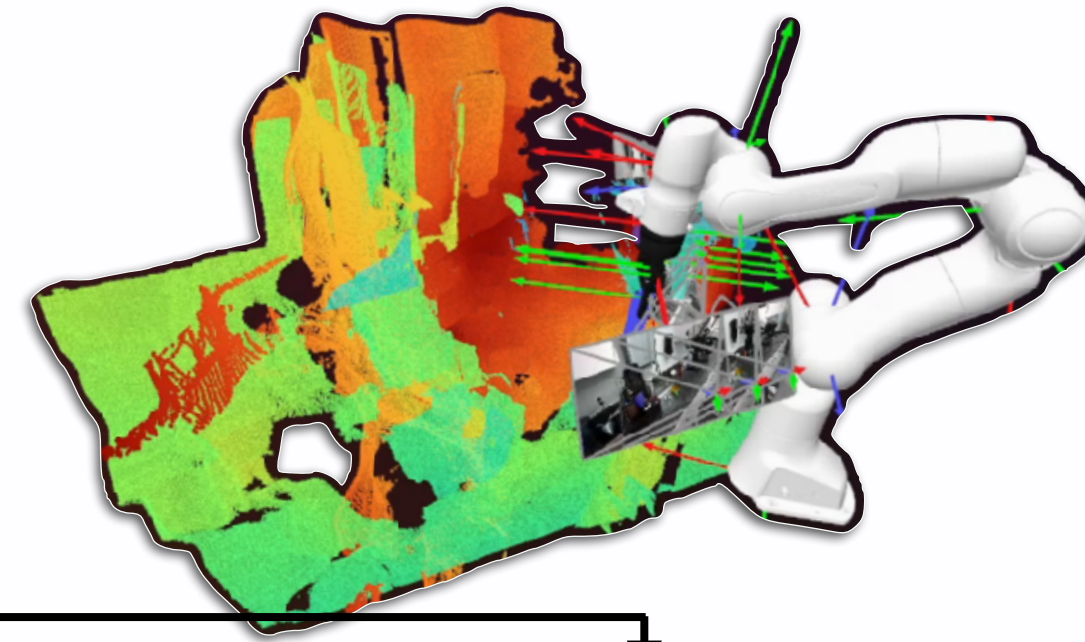
Embodied Question Template

Query

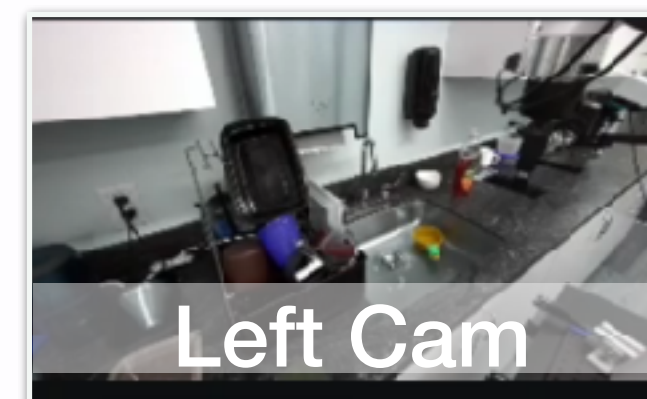
Visual Language Grounding

Embodied Question Conversion

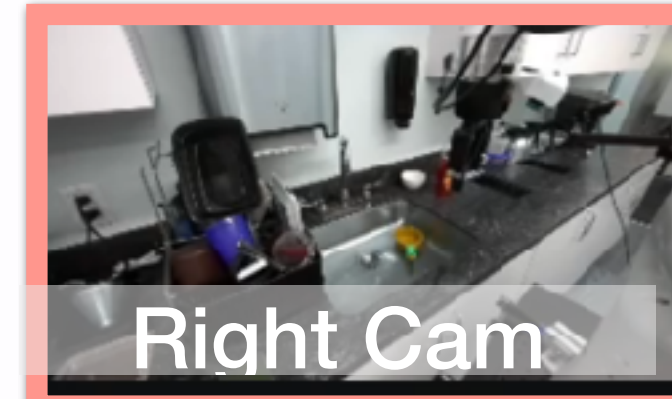
Spatial Query Projection



Wrist Cam

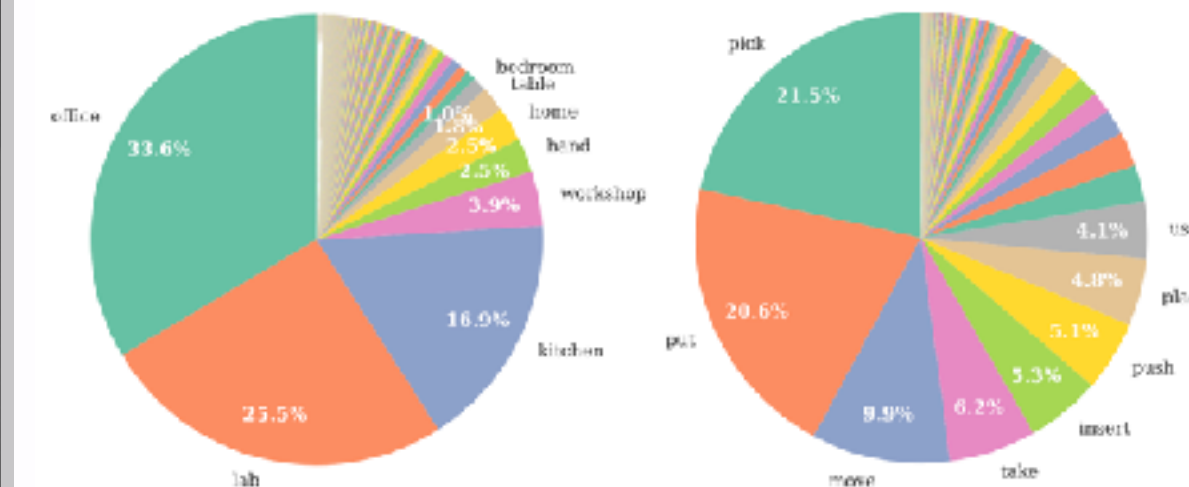


Left Cam

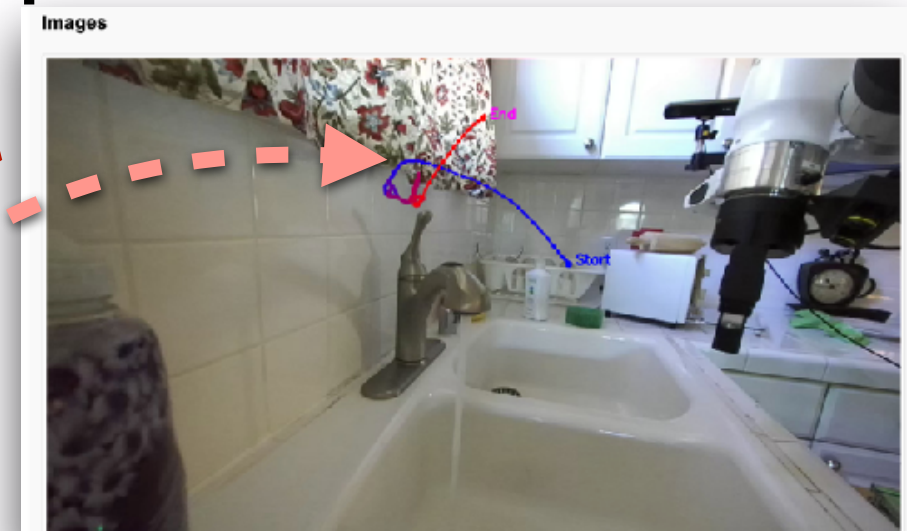


Right Cam

VQA categories

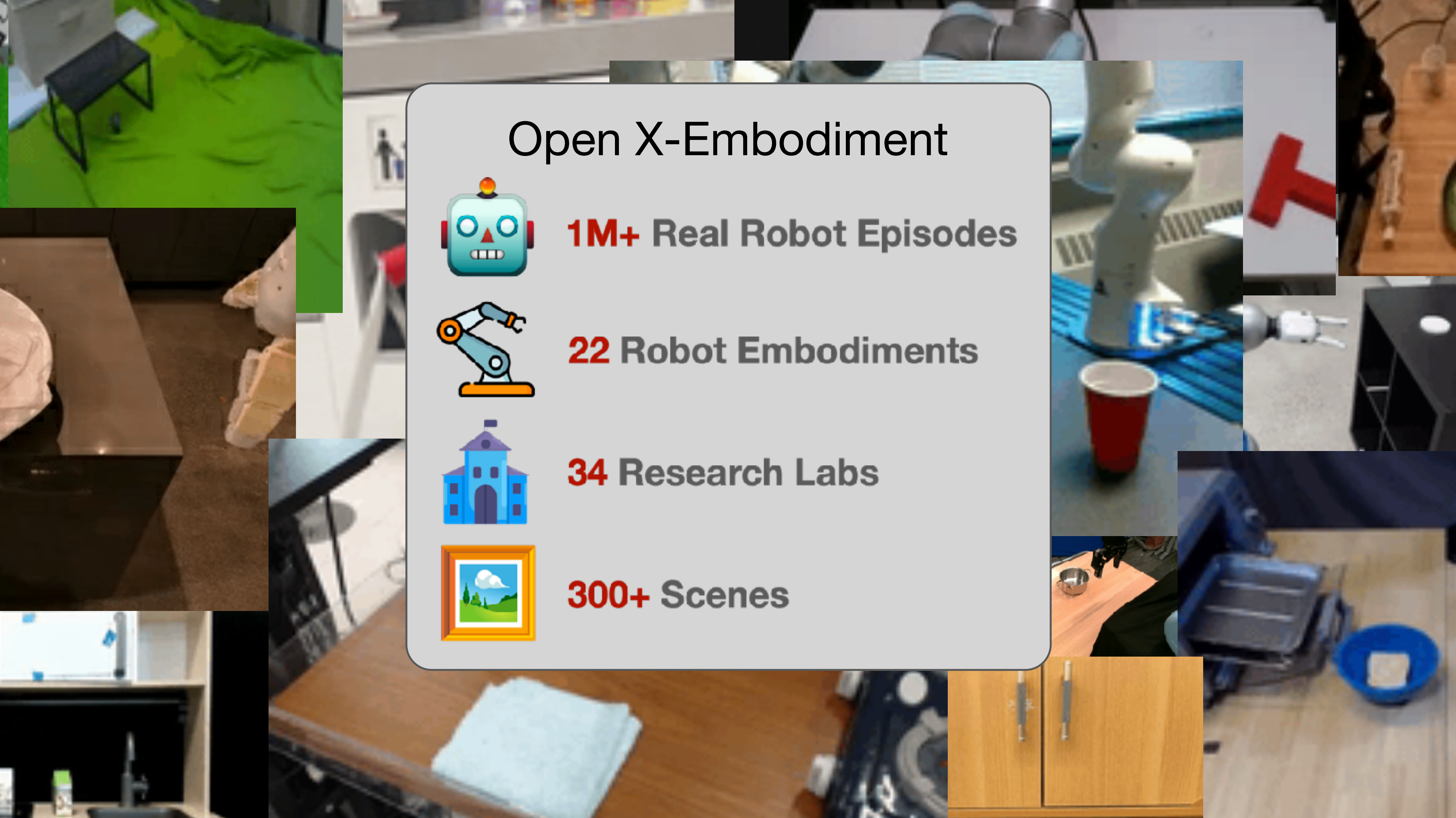


Sample Question

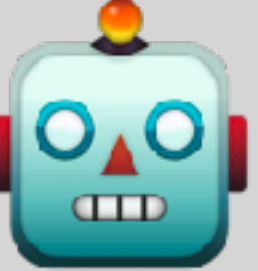





Selected
camera view

Question	Choices
Which language instruction best describes the robot's trajectory shown in the image? <small>ID: Mor_Oct_23_13_15_02_2023_01 Episode: Open the tap_190 trajectory_understanding</small>	<div><input type="radio"/> A. Drop the book into the platform</div> <div><input type="radio"/> B. Move the pan to the drawer</div> <div><input checked="" type="radio"/> C. Open the tap (Correct)</div> <div><input type="radio"/> D. Grab the cup with the gripper</div>



Open X-Embodiment

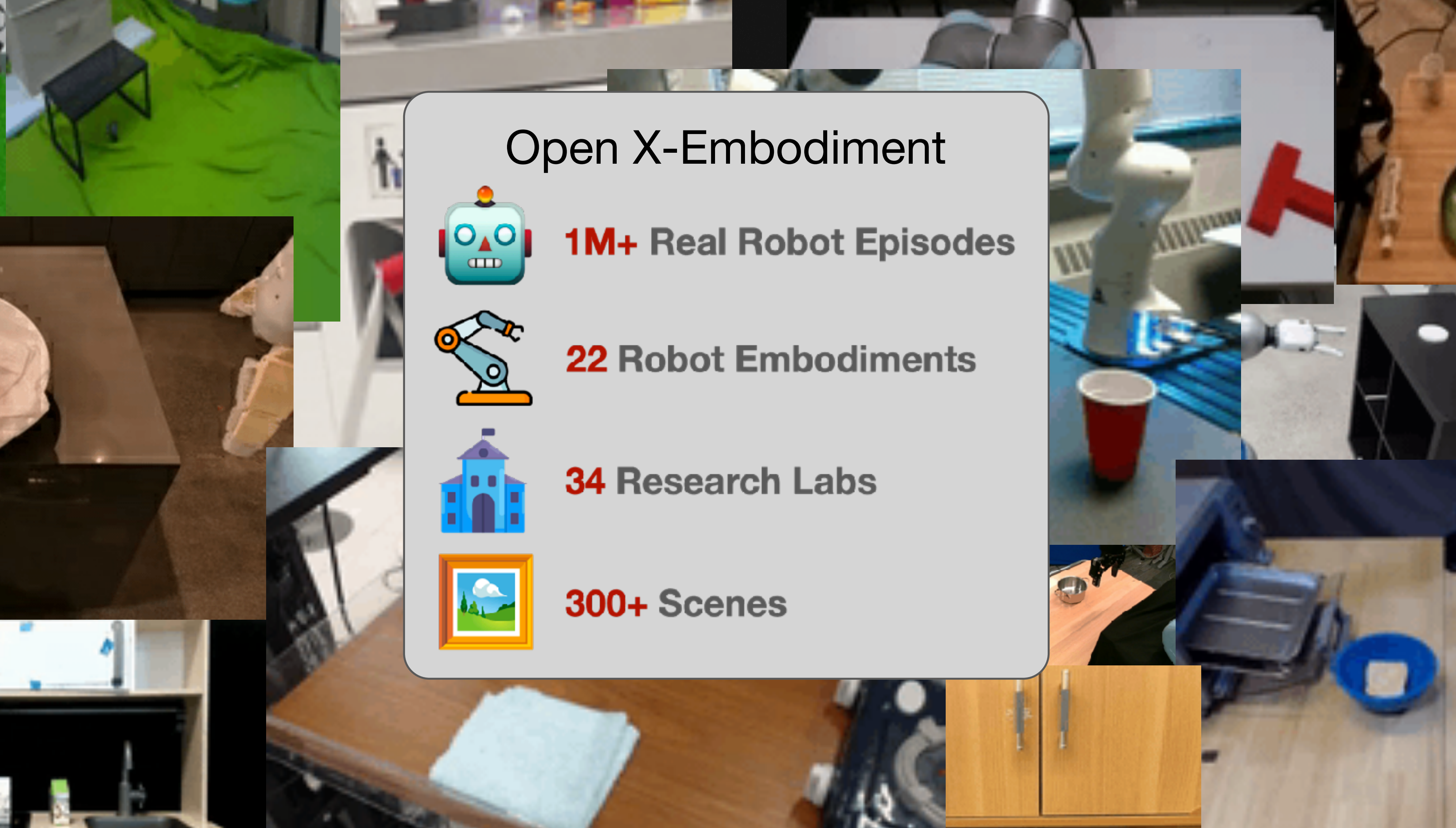
-  **1M+** Real Robot Episodes
-  **22** Robot Embodiments
-  **34** Research Labs
-  **300+** Scenes

Open X-Embodiment: Robotic Learning Datasets and RT-X Models. Open X-Embodiment Collaboration, et al. ICRA 2024

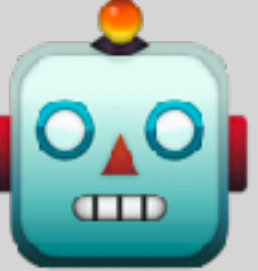



Open X-Embodiment

- 1M+** Real Robot Episodes
- 22** Robot Embodiments
- 34** Research Labs
- 300+** Scenes

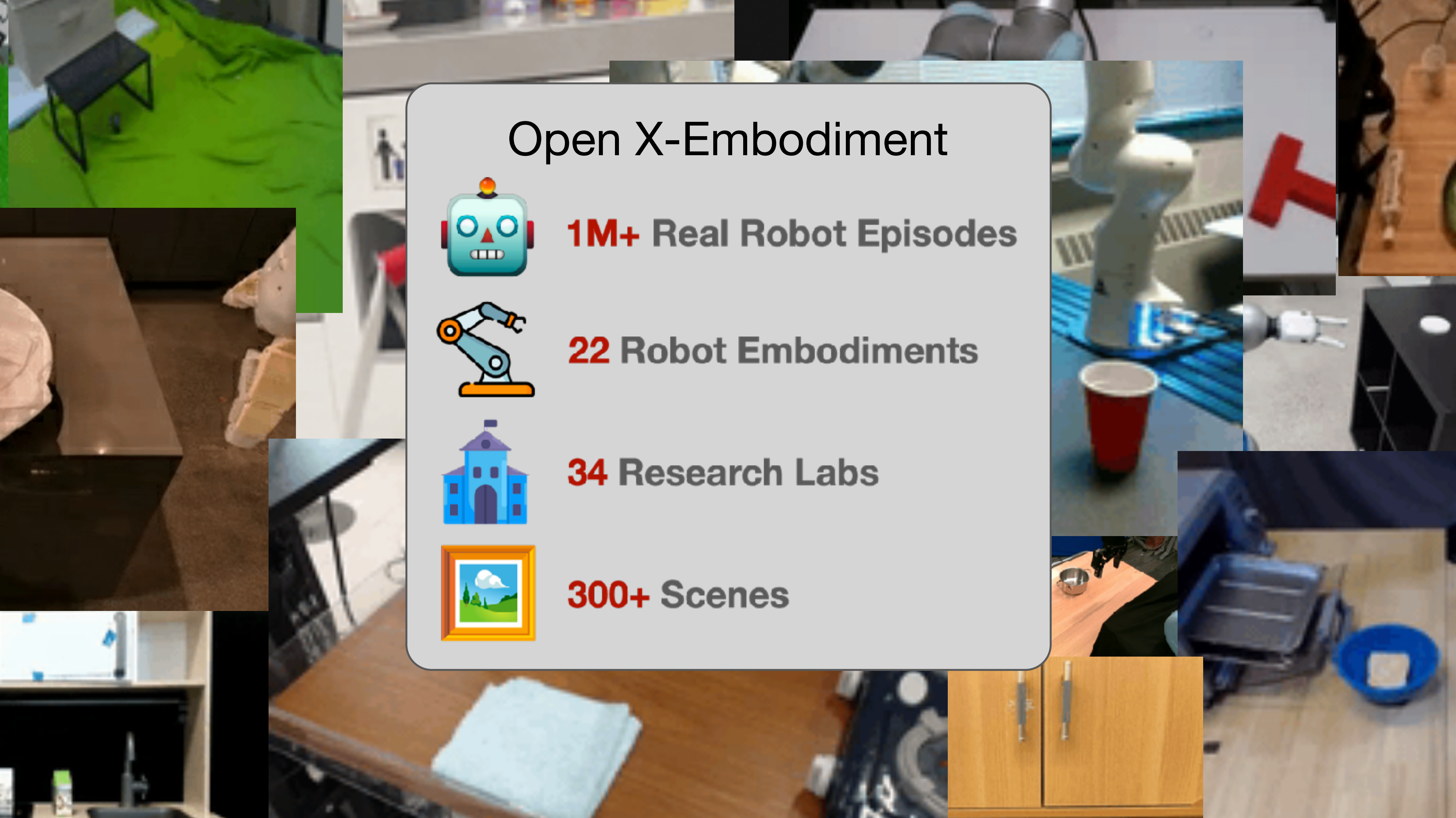
Open X-Embodiment: Robotic Learning Datasets and RT-X Models. Open X-Embodiment Collaboration, et al. ICRA 2024



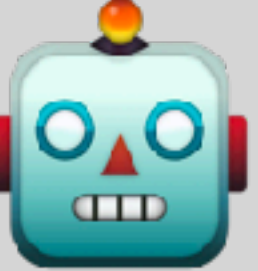



Open X-Embodiment

-  **1M+** Real Robot Episodes
-  **22** Robot Embodiments
-  **34** Research Labs
-  **300+** Scenes

Open X-Embodiment: Robotic Learning Datasets and RT-X Models. Open X-Embodiment Collaboration, et al. ICRA 2024



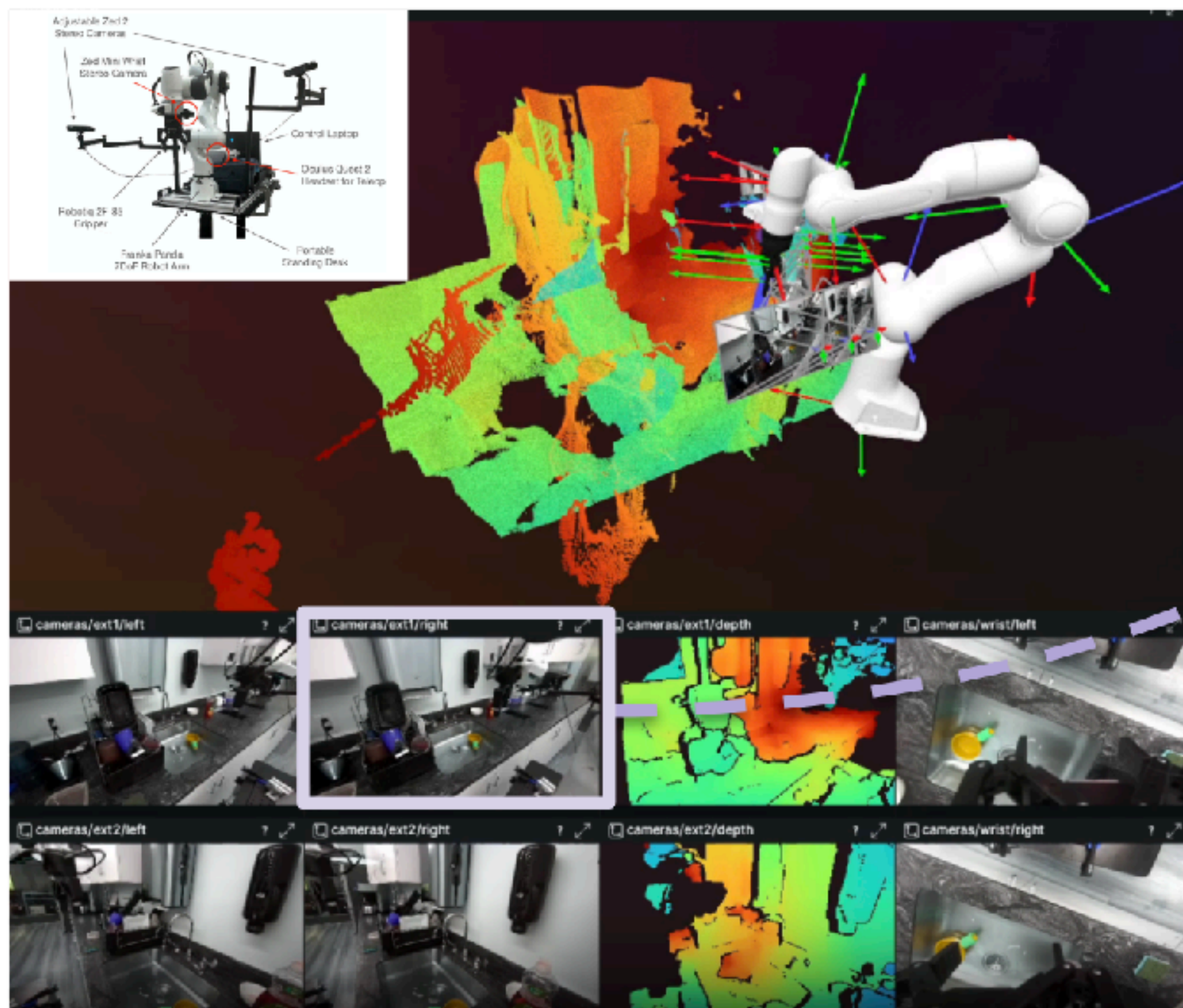
Open X-Embodiment

-  **1M+** Real Robot Episodes
-  **22** Robot Embodiments
-  **34** Research Labs
-  **300+** Scenes

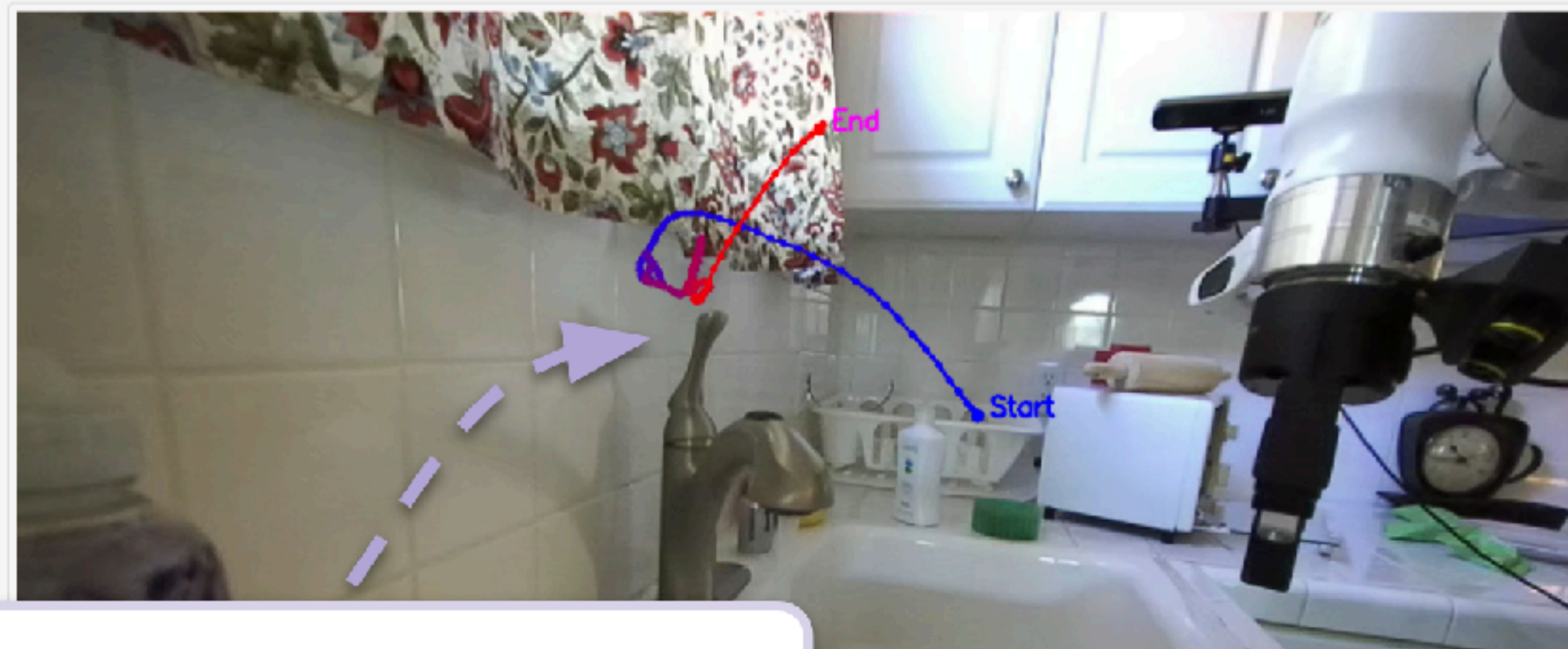
Open X-Embodiment: Robotic Learning Datasets and RT-X Models. Open X-Embodiment Collaboration, et al. ICRA 2024

[illegible]

A detailed example VQA generation



Images



Spatial Query Projection

Question

Which language instruction best describes the robot's trajectory shown in the image?

ID: Mon_Oct_23_13_15_02_2023_q1 Episode: Open the tap_190

[trajectory_understanding](#)

Choices

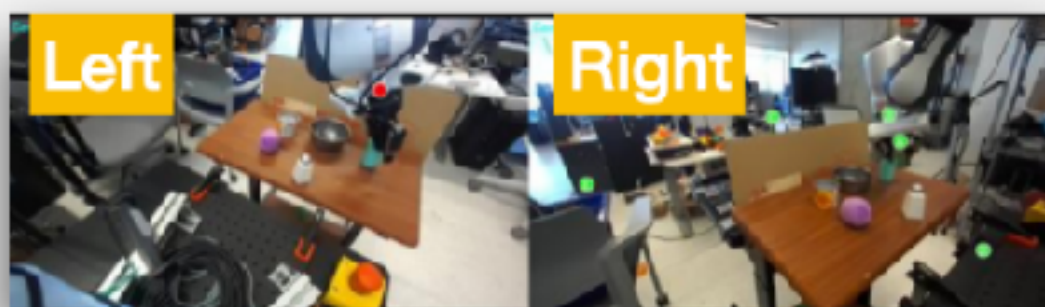
- A. Drop the book into the platform
- B. Move the pen to the drawer
- C. Open the tap (Correct)
- D. Grab the cup with the gripper

QA generation:

Scene-Interaction Understanding

Embodied Question Template

Keyframe Selection



Multiple View

Q: In the left image (ext1 camera), a red dot is marked. Which point is the closest point in the right image (ext2 camera) corresponding to the same 3D location?

Key Modalities: 2 views, stereo (depth) img



Scene Understanding

Q: In the image from ext2, which colored point is CLOSEST to the camera?

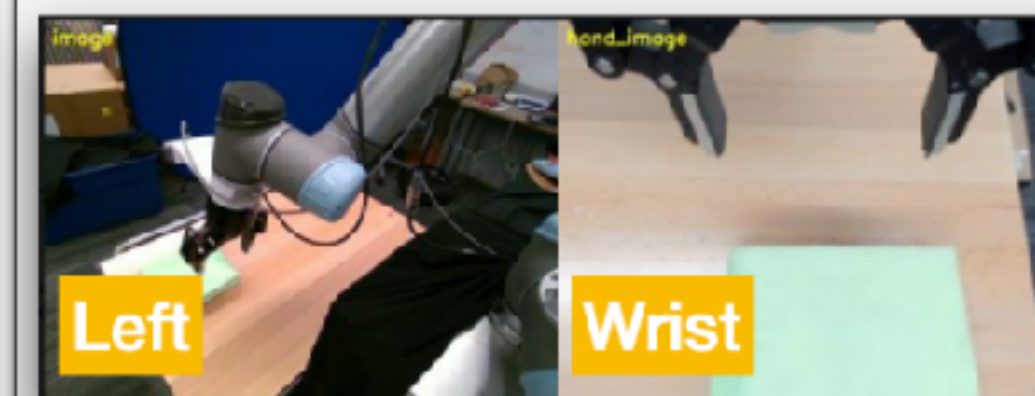
Key Modalities: stereo (depth) images



Task State - Success

Q: The robot is to turn on the toaster. Has the robot successfully completed the task?

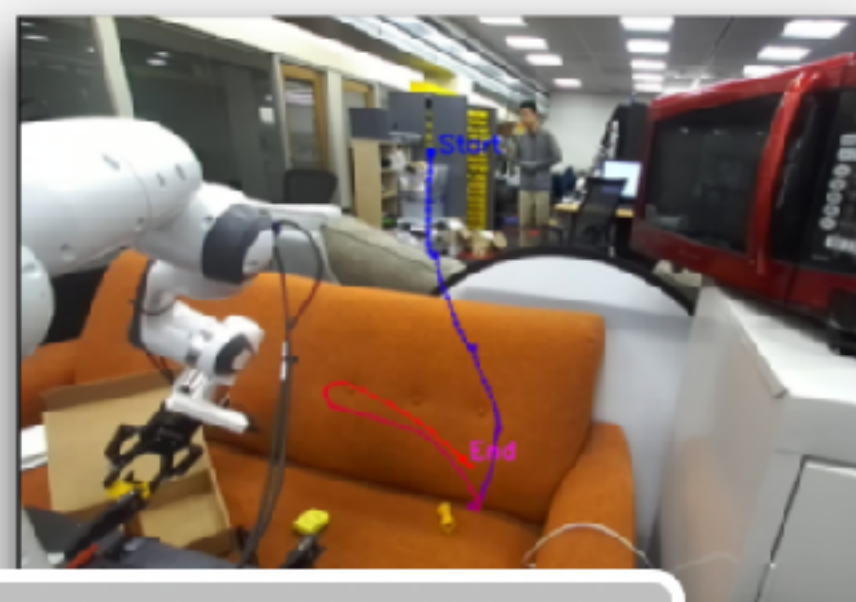
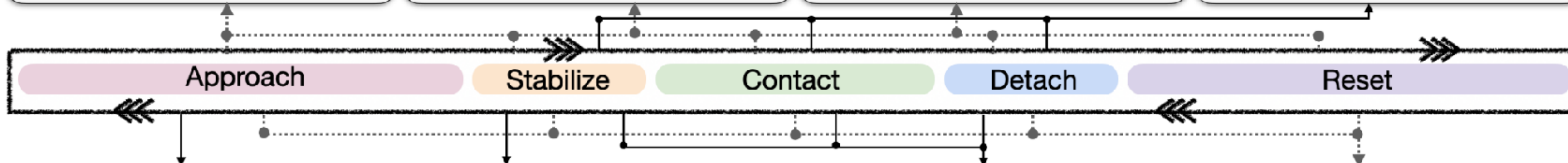
Key Modalities: 2 Side View, Wrist View



Robot State - Gripper

Q: Is the robot gripper open?

Key Modalities: gripper state, Side+Wrist View



Trajectory Understanding

Q: what language instruction best describes the robot's trajectory shown in the image?

Key Modalities: End effector pose, language



Spatial Relationship

Q: The robot is tasked to move the spoon, which arrow shows the most possible direction to move next?

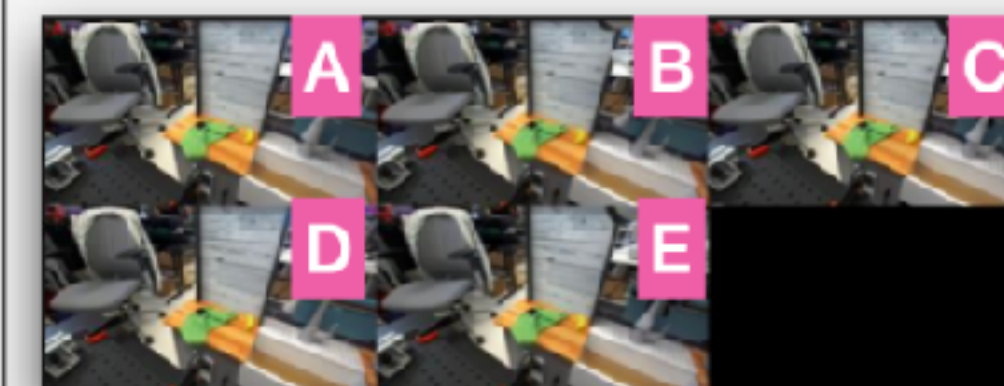
Key Modalities: End effector pose, language



Task State - Grasp

Q: Is the robot grasp Sponge stable?

Key Modalities: language, gripper state

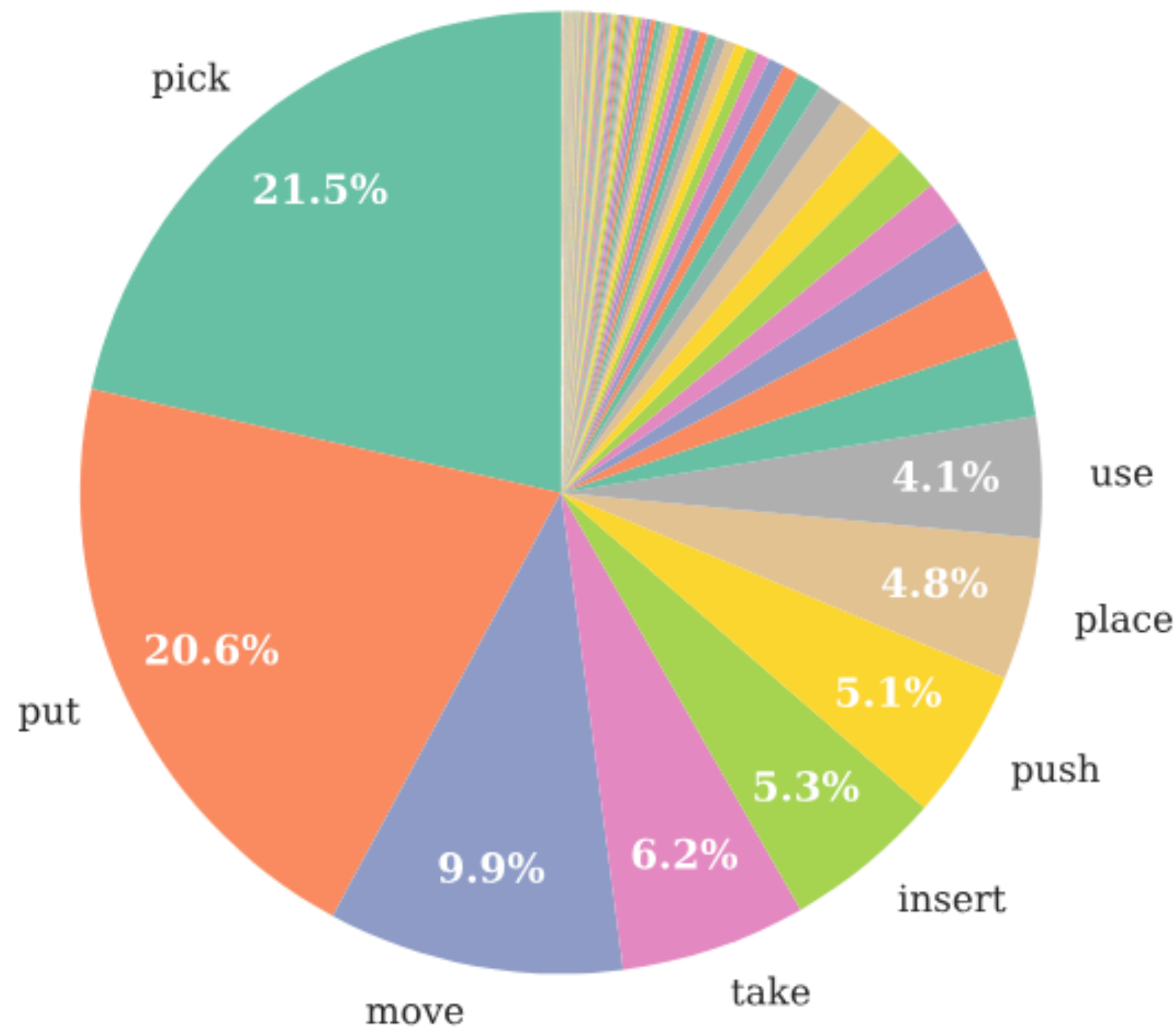
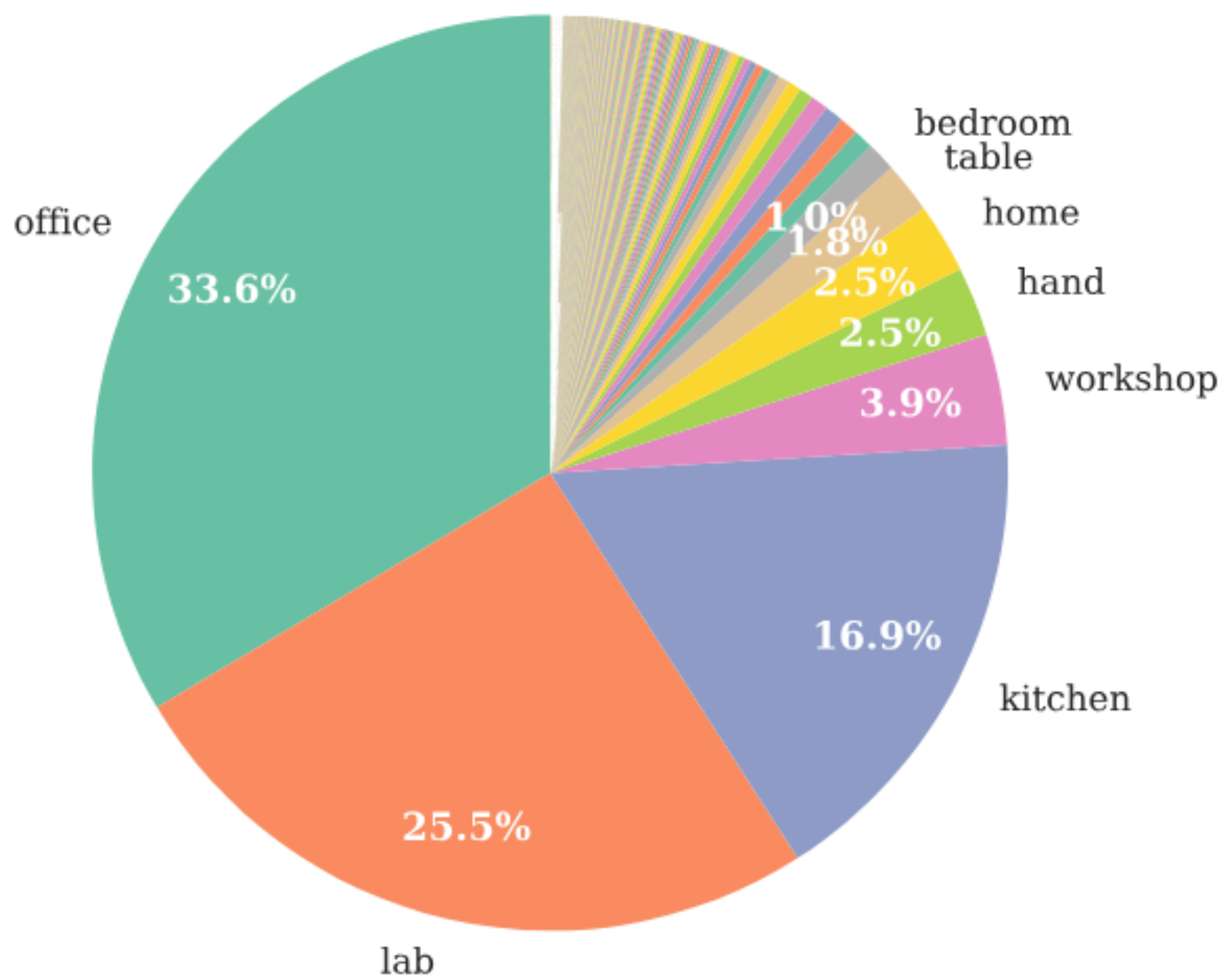


Task State - Goal

Q: The robot is tasked to move the tap, which configuration shows the goal state that the robot show achieve?

Key Modalities: language, gripper state

Robo2VLM-1 overview and statistic



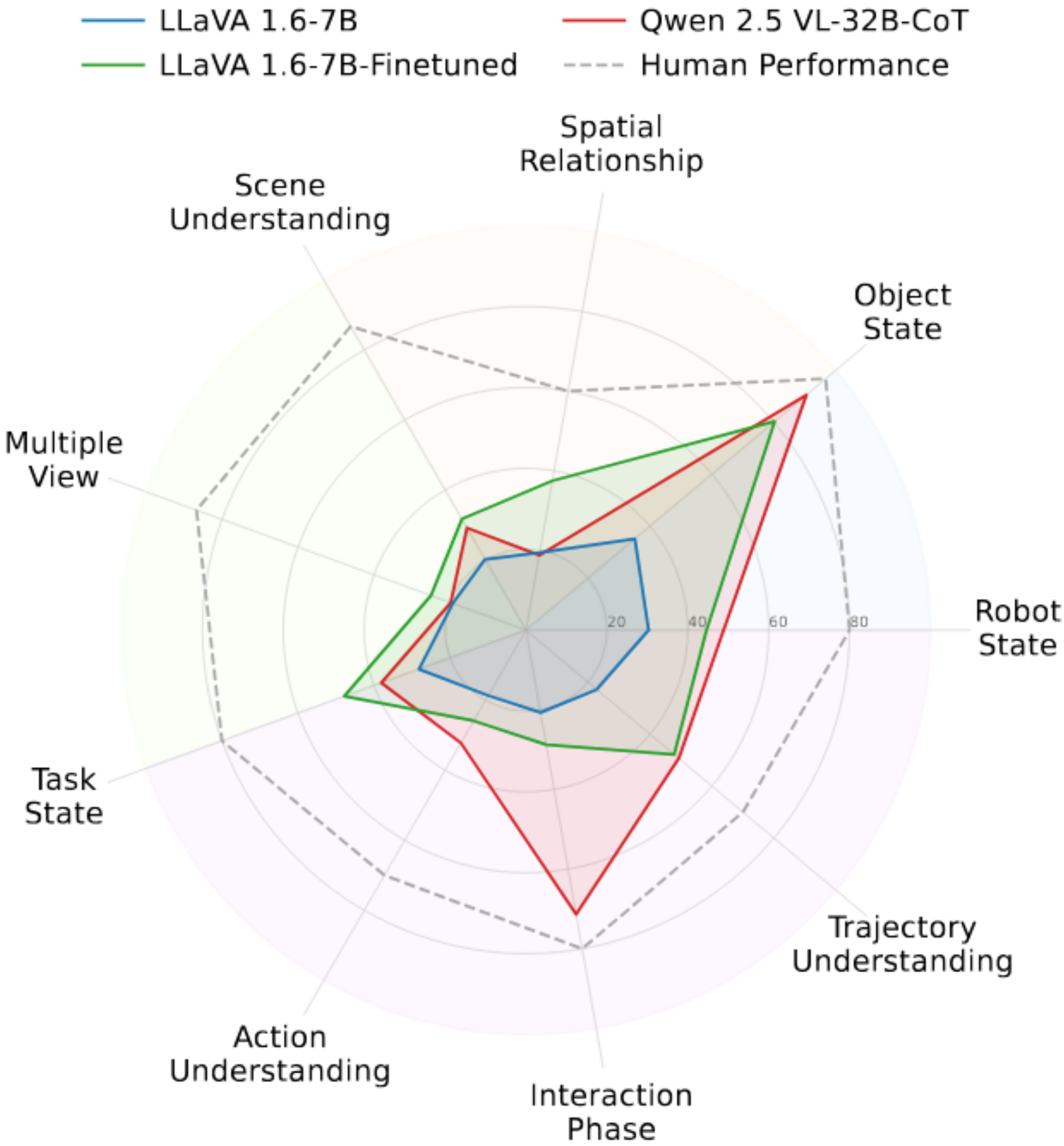
Overall	
Total samples:	60,000
Questions	
Avg. length:	108.69 ch
Median length:	113.00 ch
Min length:	28 ch
Max length:	378 ch
Choices	
Avg. # / Q:	4.65
Avg. length:	14.22 ch
Max length:	271 ch
Image Resolutions	
Avg. width:	520.66 px
Avg. height:	292.99 px
Most common:	640x360 (39.61%)
Unique res.:	19

Evaluation

Table 3: Performance Comparison of Multimodal Foundation Models on OpenX-VQA Benchmark Categories (%). Upper part: zero-shot. Lower part: with CoT prompting.

Model	Overall	Spatial Reasoning					Goal Reasoning			Interaction Reasoning			
		RS	OS	SR	SU	MV	TS-G	TS-S	TS-GL	AU	IP	TU	
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	
Zero-Shot													
LLaVA 1.5-7B	21.58	35.32	23.87	16.08	17.78	17.50	31.82	23.79	19.03	20.30	21.74	22.37	
LLaVA 1.6 Mistral-7B	24.09	30.31	35.13	19.42	20.24	19.29	34.20	30.77	19.52	18.67	20.70	22.83	
LLaVA 1.6-34B	24.94	26.66	29.75	21.47	23.18	17.86	29.19	29.40	17.90	19.49	36.98	30.59	
Llama 3.2-90B	28.60	31.94	55.87	18.51	26.61	16.43	28.23	35.27	8.06	18.13	51.56	49.77	
Qwen 2.5 VL-7B	30.63	41.68	55.63	21.55	24.38	17.32	33.01	42.57	7.82	25.71	46.61	39.73	
Qwen 2.5 VL-32B	37.68	49.39	71.37	21.85	28.53	17.50	34.21	55.08	12.90	30.45	63.80	49.32	
Qwen 2.5 VL-72B	37.76	38.84	85.00	22.31	28.23	15.71	28.47	51.89	10.08	33.96	71.09	54.79	
CoT Reasoning													
LLaVA 1.5-7B	21.61	28.28	21.00	17.37	20.90	18.93	25.36	24.19	21.53	21.24	20.31	20.09	
LLaVA 1.6 Mistral-7B	24.05	27.60	38.87	17.15	20.18	22.32	25.84	28.03	18.47	18.40	30.60	29.68	
LLaVA 1.6-34B	23.49	20.43	31.00	21.24	22.88	20.36	18.18	26.14	16.77	21.79	35.16	26.94	
Llama 3.2-90B	30.45	32.34	79.87	13.35	26.37	18.57	29.90	29.14	14.27	19.76	59.24	44.75	
Qwen 2.5 VL-7B	34.82	38.02	90.00	21.78	23.30	16.79	36.84	46.48	18.39	28.15	42.71	36.99	
Qwen 2.5 VL-32B	41.30	48.85	90.50	18.82	29.19	19.82	35.17	60.43	18.71	32.21	71.35	49.32	
Qwen 2.5 VL-72B	39.52	44.79	92.37	18.36	29.73	13.39	29.19	55.28	13.15	36.13	74.09	46.12	

Performance Comparison of Multimodal Foundation Models on Robo2VLM



Comparison with human annotation

VLM Finetuning Result



Robo2VLM: Visual Question Answering from Large-Scale In-the- Wild Robot Manipulation Datasets

Kaiyuan Chen*, Shuangyu Xie*, Zehan Ma, Pannag Sanketi, Ken Goldberg
<https://berkeleyautomation.github.io/robo2vlm/>

