Join at slido.com
#3175898

You are developing mission-critical software for an autonomous drone. The software includes many learning-assisted functions implemented as periodic real-time tasks that generally read streaming data from the physical world (e.g., images, sensor readings, etc), analyze these readings, then respond accordingly. Proper drone operation depends on meeting strict deadlines on data processing. The underlying real-time OS supports multiple scheduling policies for such tasks. Two primary candidates are rate monotonic and EDF. Which policy would you choose?

Today's Debate:

Part I: Group #2 Argues for "EDF"

Madhav Khirwar
Kai-Siang Wang
Jiayi Xiao

Part II: Group #4 Argues for "Rate Monotonic"

Tomoyoshi Kimura
Yuheng Pan
Hongjue Zhao

Part III: Moderated Open Discussion: All are welcome to join in contributing any further pro/counter arguments

Closed Loop Control and Related Foundation Models (RT-2, RT-X, etc)

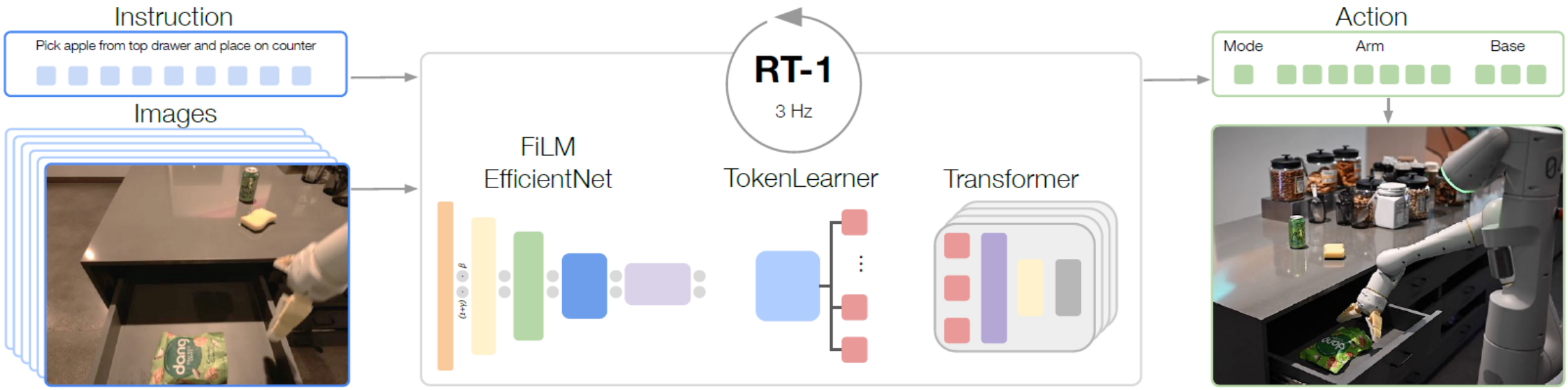
ROBOTICS FOUNDATION MODELS

Objectives

- Can we train a single, capable, large multi-task backbone model on data consisting of a wide variety of robotic tasks?
- Does such a model enjoy the benefits observed in other domains, exhibiting zero-shot generalization to new tasks, environments, and objects?

[1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan et al. "Rt-1: Robotics transformer for real-world control at scale." *arXiv preprint arXiv:2212.06817* (2022).

Robotic Transformer (RT-1)



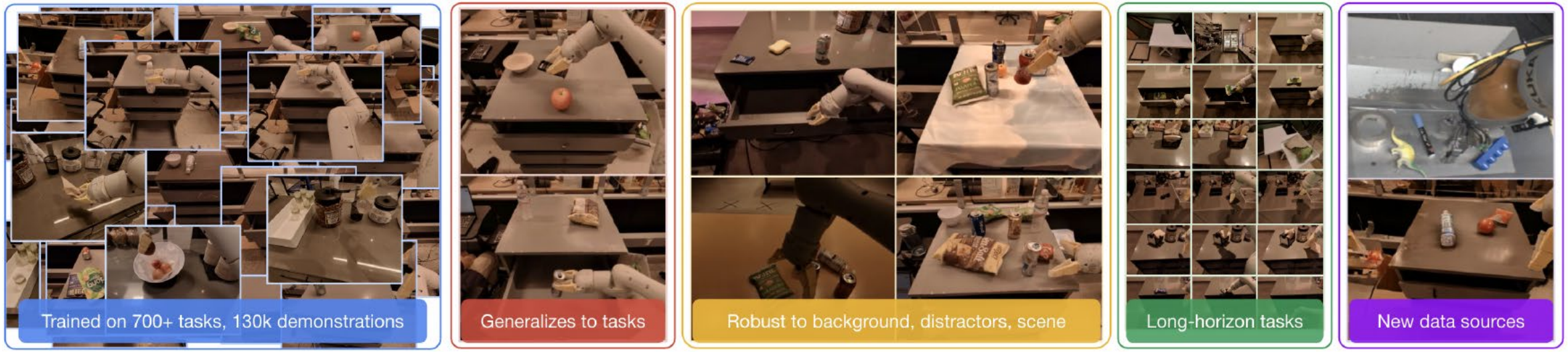
RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).

Training

- At timestep $t = 0$, the policy π is presented with a language instruction i and an initial image observation x_0.
- The policy produces an action distribution $\pi(\cdot \mid i, x_0)$ from which an action a_0 is sampled and applied to the robot.
- This process continues, producing actions a_t by sampling from a learned distribution $\pi(\cdot \mid i, \{x_j\}, 0 < j < T)$ and applying those actions to the robot. The interaction ends when a termination condition is achieved.
- The full interaction from the starting step $t = 0$ to terminating step T is referred to as an episode.
- At the end of an episode, the agent will be given a binary reward $r \in \{0, 1\}$ indicating whether the robot performed the instruction i correctly.
- The goal is to learn a policy π that maximizes the average reward, in expectation over a distribution of instructions, starting states x_0, and transition dynamics.

Training Data and Evaluation

- Training data collected over the course of 17 months with a fleet of 13 robots, containing ~130k episodes and over 700 tasks



RT-1's large-scale, real-world training (130k demonstrations) and evaluation (3000 real-world trials) show impressive generalization, robustness, and ability to learn from diverse data.

Training Tasks and Skills

Training tasks are grouped by the verb used in them. Verbs define “skills”. The table shows the different skills used and the number of tasks per 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

The Testbed

(a)

(b)

(c)

The Arm

(d)

(e)

(f)

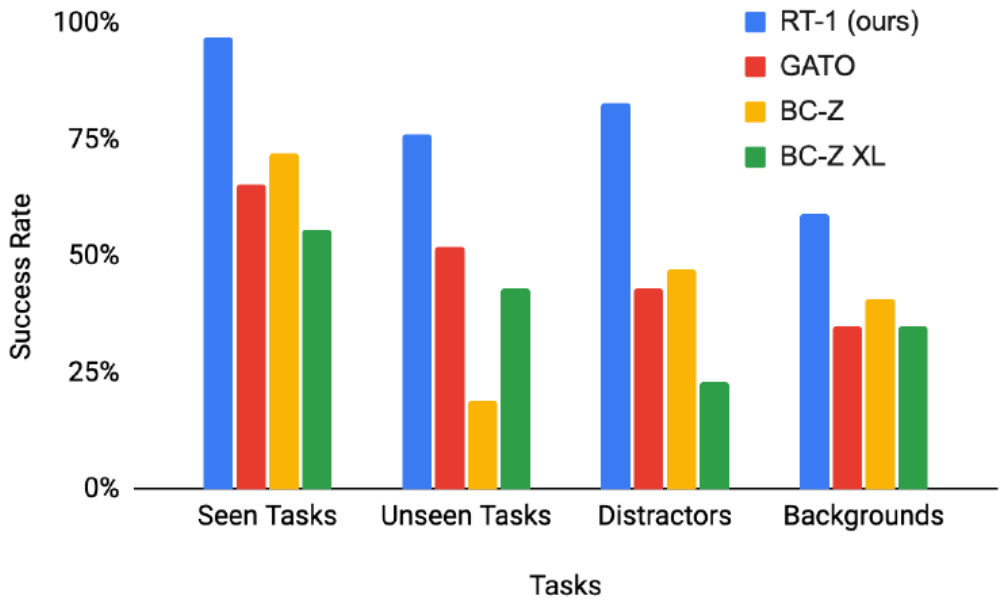
The Objects

Figure : (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.

Evaluation

RT-1 outperforms competition on both tasks seen during training as well as unseen tasks, even in the presence of distractors and different backgrounds

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**

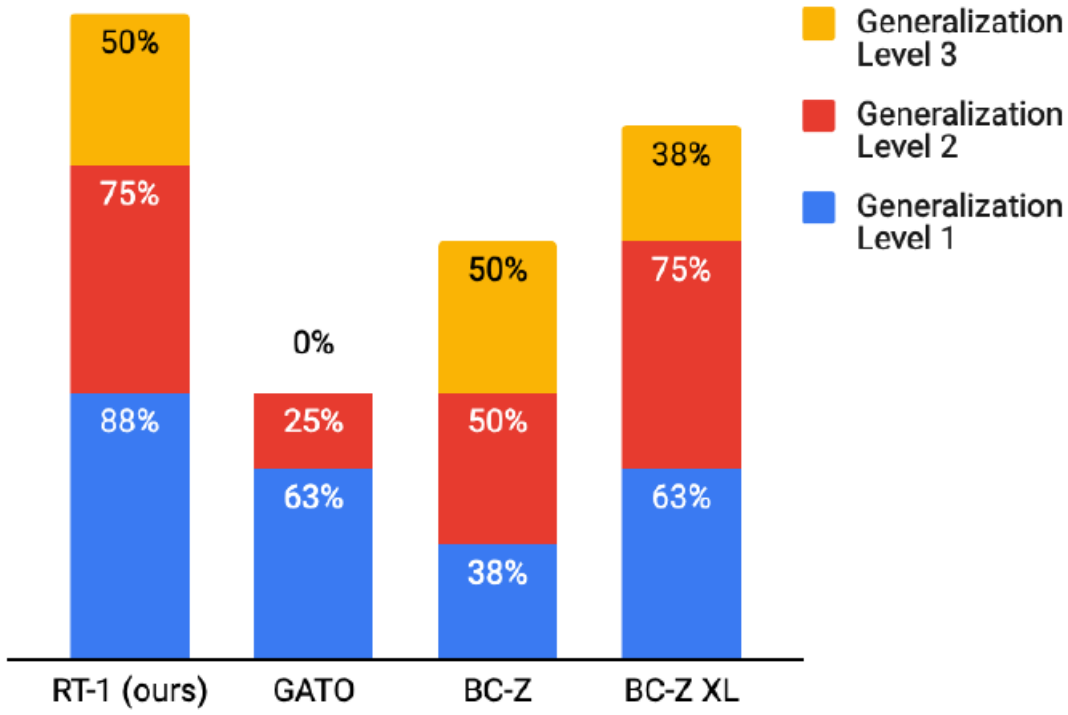


Evaluation

RT-1 outperforms competition on tasks requiring generalization:

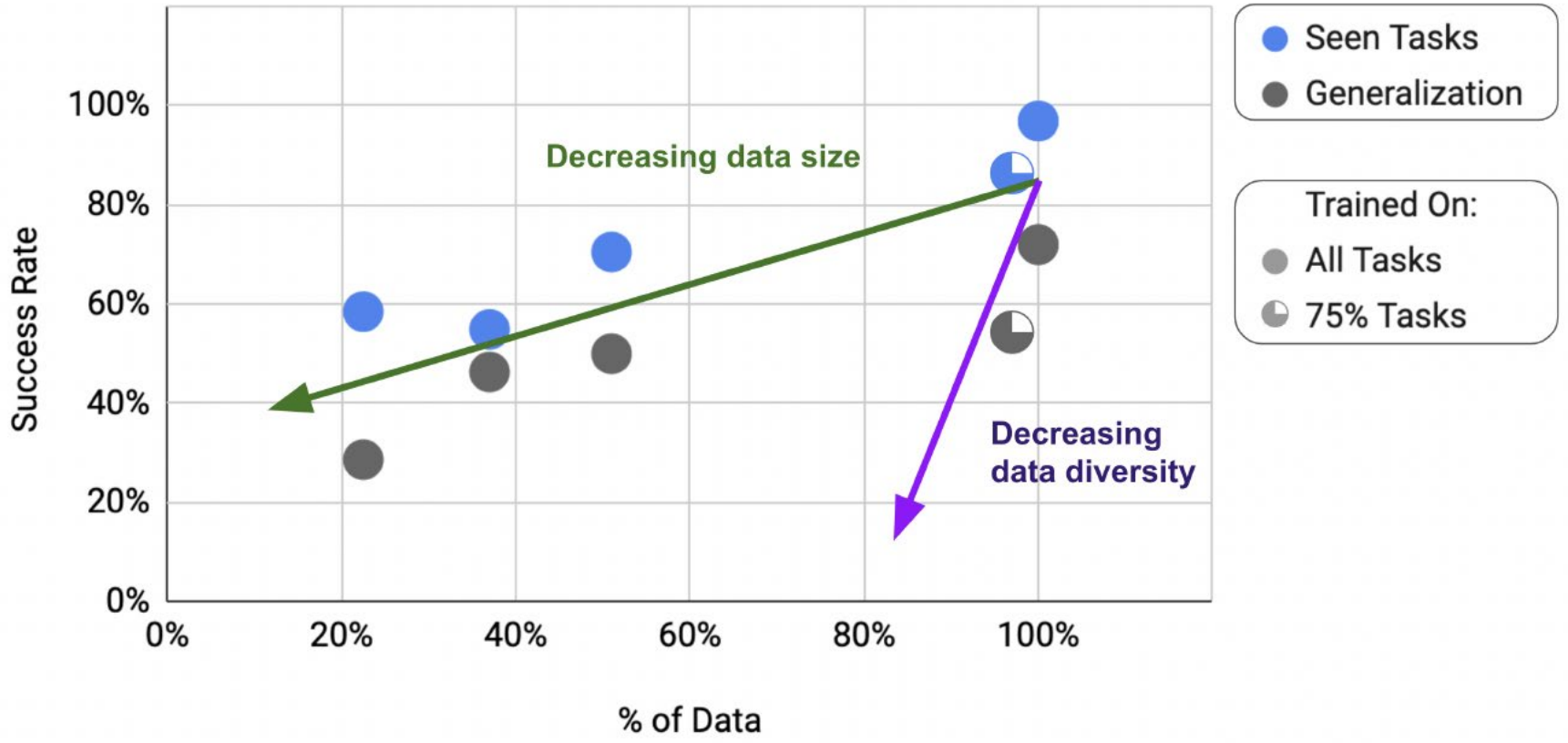
- L1: generalization to the new counter-top layout and lighting conditions
- L2: additionally, generalization to unseen distractor objects
- L3 for additionally generalization to new task settings, new task objects or in unseen locations

		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**



Evaluation

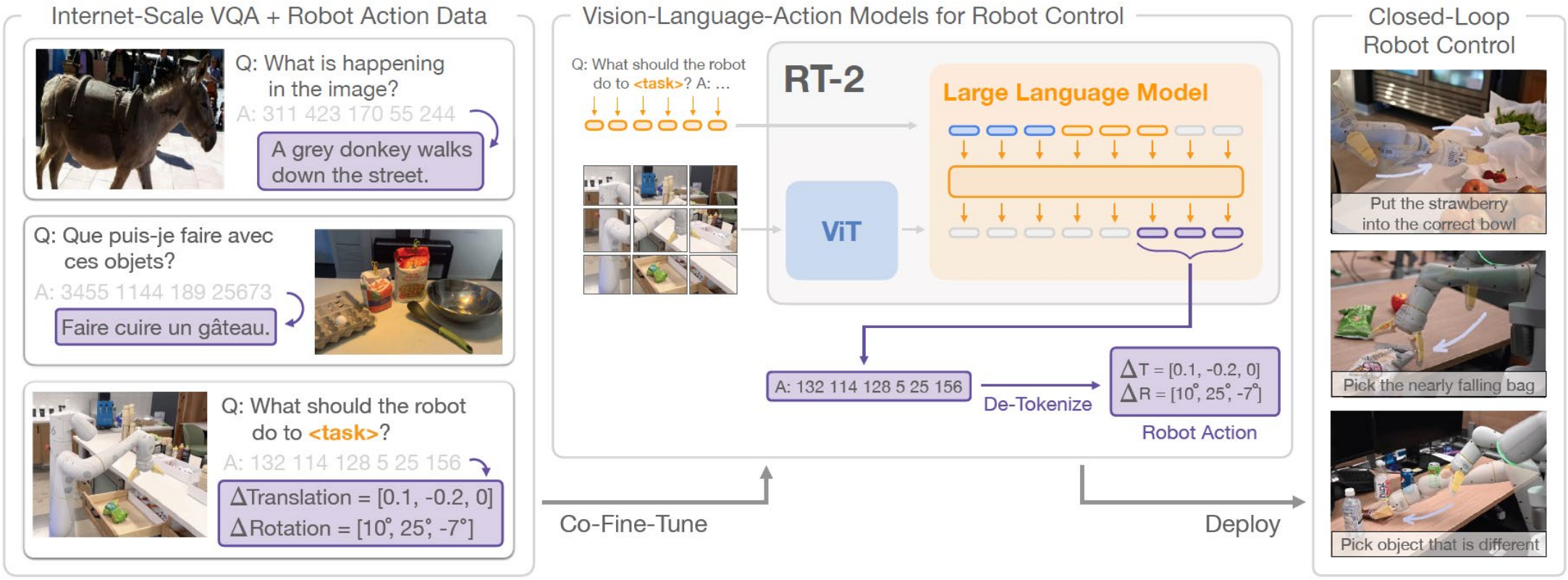
Impact of training data size



[2] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." In *Conference on Robot Learning*, pp. 2165-2183. PMLR, 2023.

RT-2: Improving Generalization

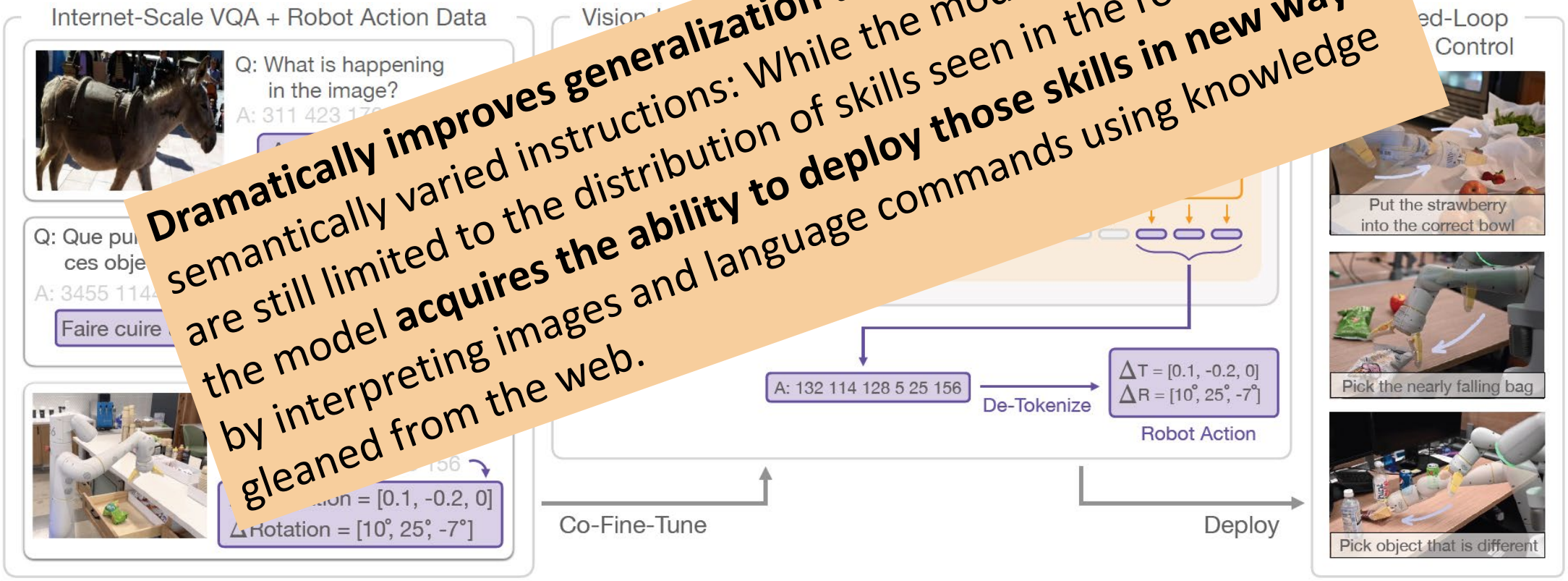
Idea: directly train vision-language models designed for open-vocabulary visual question answering and visual dialogue to output low-level robot actions, along with solving other Internet-scale vision-language tasks.



RT-2: Improving Generalization

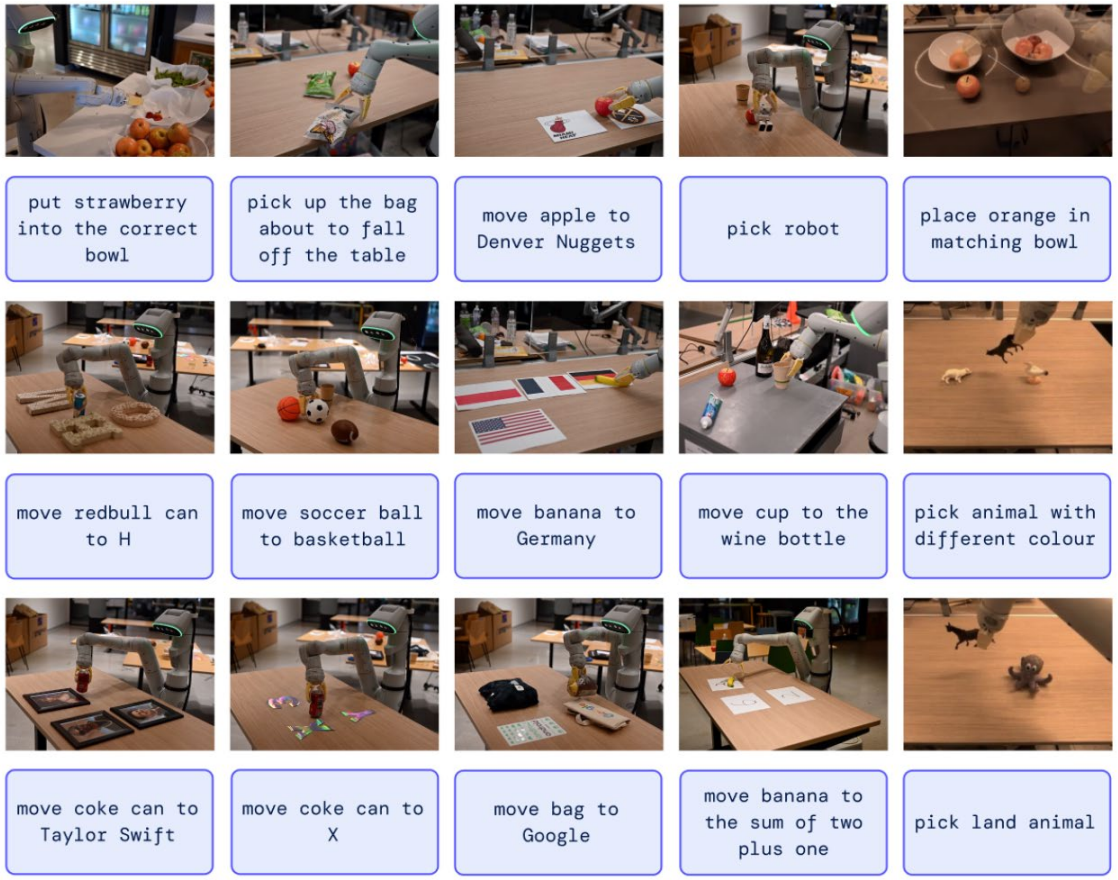
Idea: directly train vision-language models designed for open-vo… g and visual dialogue to output low-level robot actions, along with s… tasks.

Dramatically improves generalization to novel objects and semantically varied instructions: While the model's physical skills are still limited to the distribution of skills seen in the robot data, the model **acquires the ability to deploy those skills in new ways** by interpreting images and language commands using knowledge gleaned from the web.



More General Task Examples

RT-2 (based on a 55B parameter vision-language model) is able to generalize to a variety of real-world situations that require reasoning, symbol understanding, and human recognition.

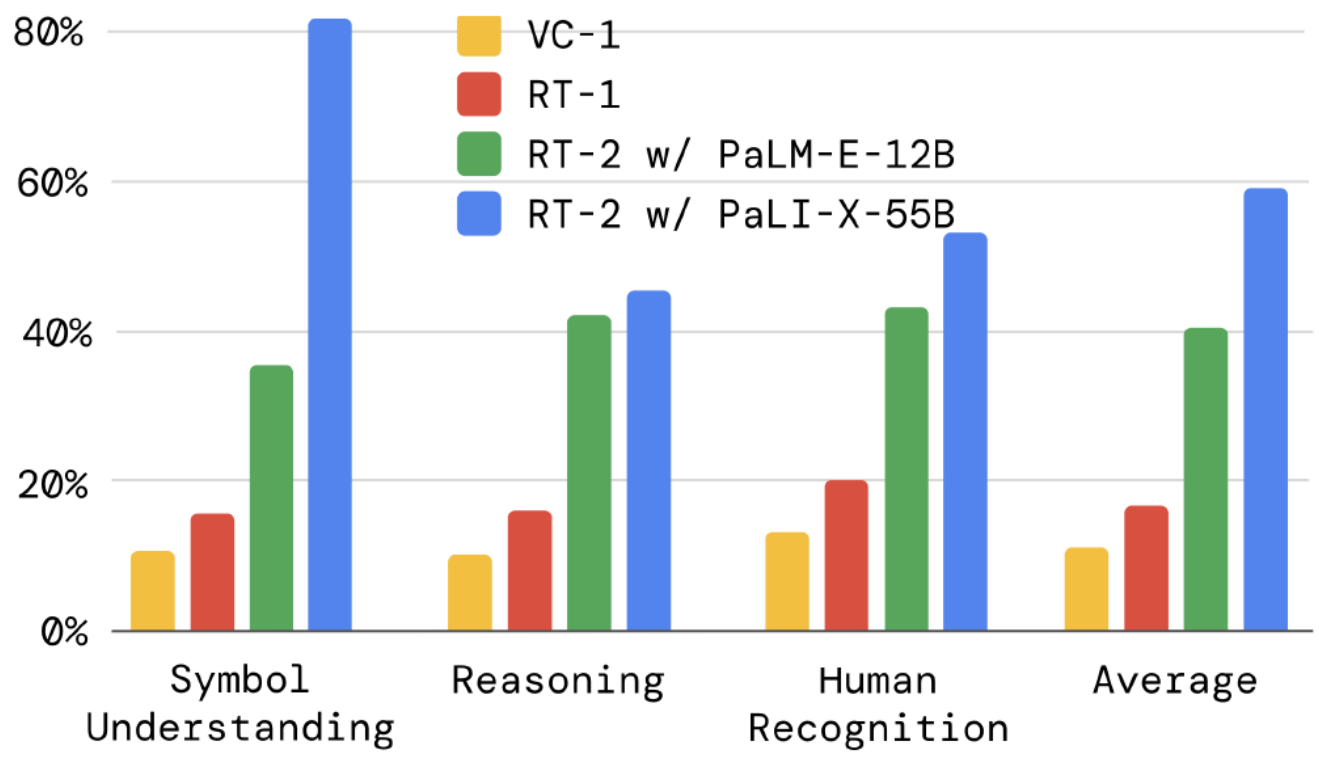


Emergent Symbolic Understanding, Reasoning, and Human Recognition

Symbol understanding: Tasks involving references to symbolic knowledge such as a country flag or a company logo.

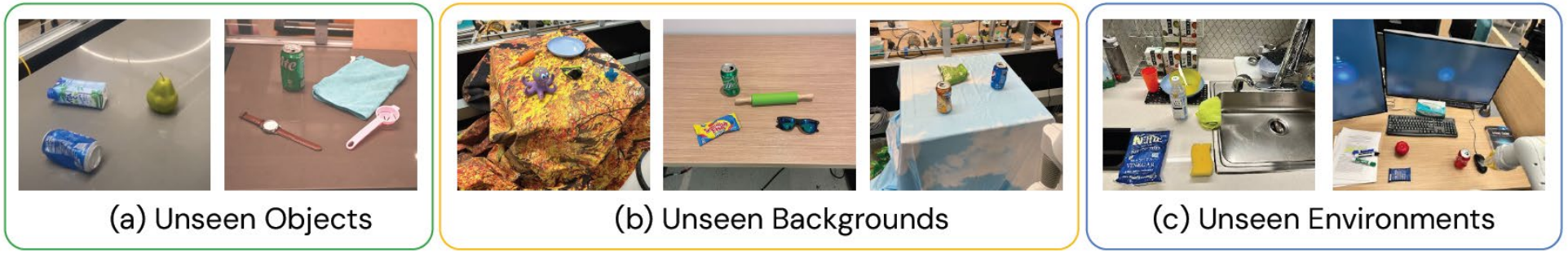
Reasoning: Tasks involving a reasoning ability such as putting items in the “correct container,” or pushing in an item that is “about to fall”.

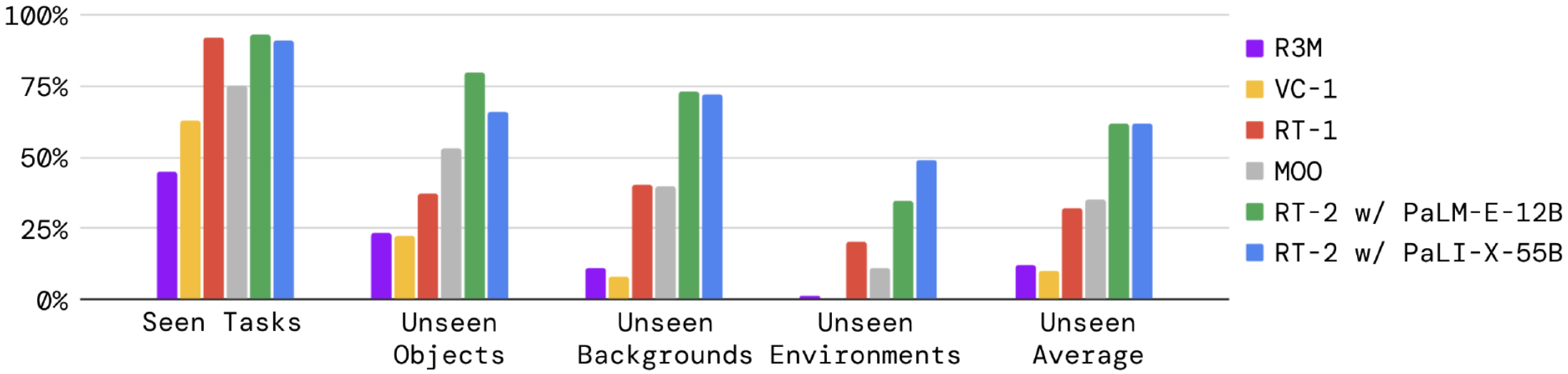
Human recognition: Tasks involving identifying a person by their visual features.



Unseen Objects, Backgrounds, and Environments

RT-2 shows a generalization ability to unseen objects backgrounds and environments.





[3] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6892-6903. IEEE, 2024.

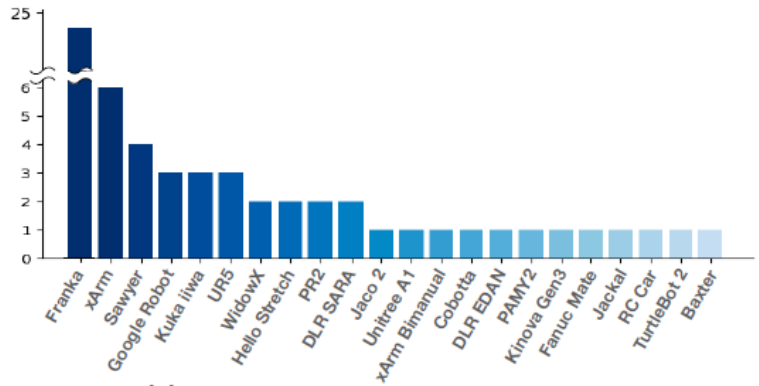
RT-X: Towards Generalist Models

Contributions:

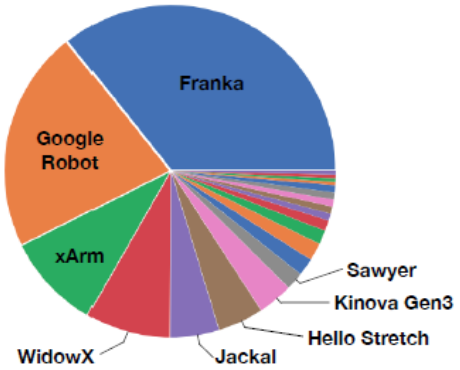
- A dataset from 22 different robots collected through a collaboration between 21 institutions, demonstrating 527 skills (160266 tasks).
- A high-capacity model trained on this data, which we call RT-X, exhibits positive transfer and improves the capabilities of multiple robots by leveraging experience from other platforms.

A Dataset Collected Across Multiple Embodiments

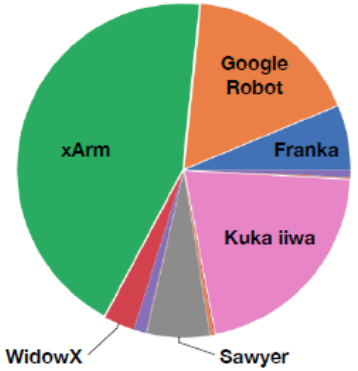
Multiple types of robots, environments (scenes), and trajectories, sharing some common skills and objects



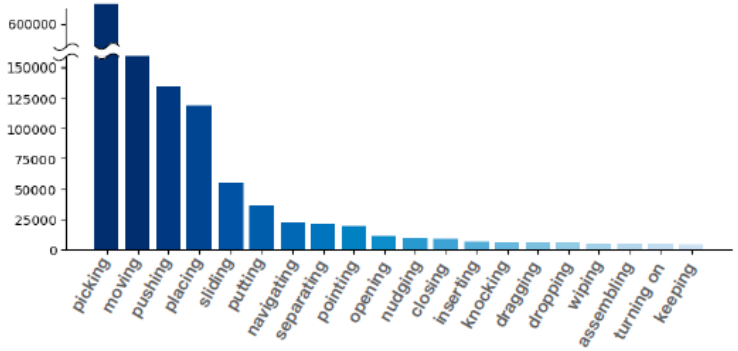
(a) # Datasets per Robot Embodiment



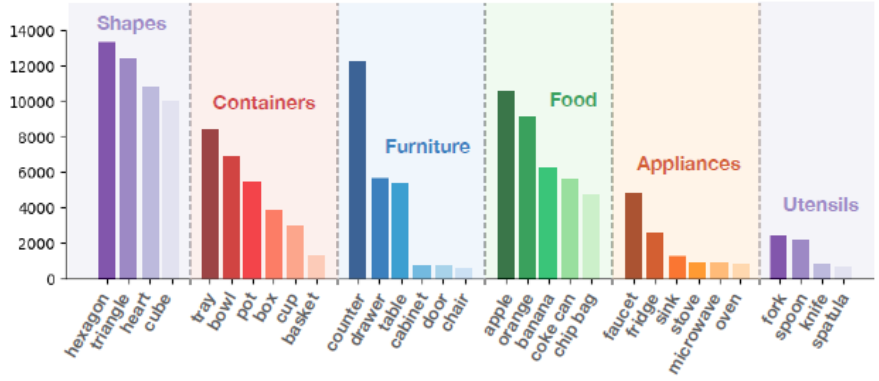
(b) # Scenes per Embodiment



(c) # Trajectories per Embodiment



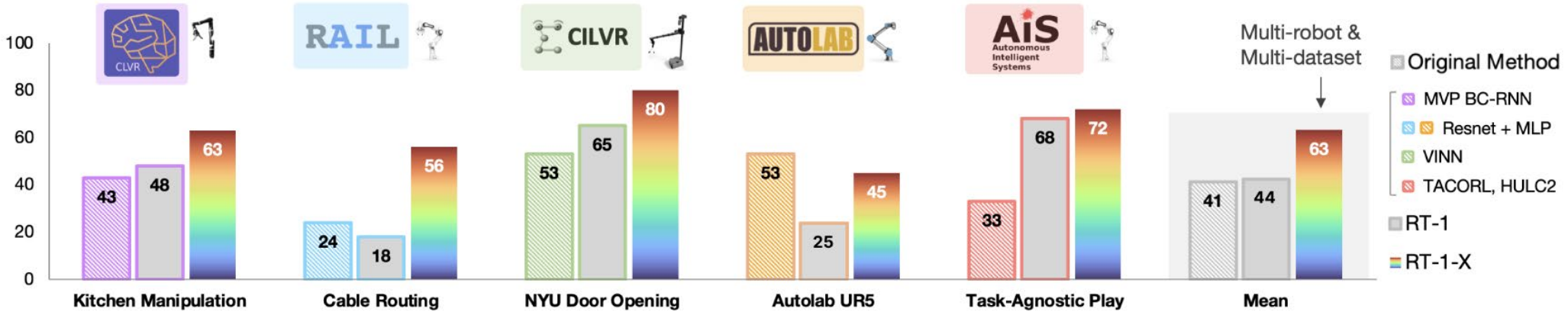
(d) Common Dataset Skills



(e) Common Dataset Objects

Training with X-Embodiment Data Improves Outcomes

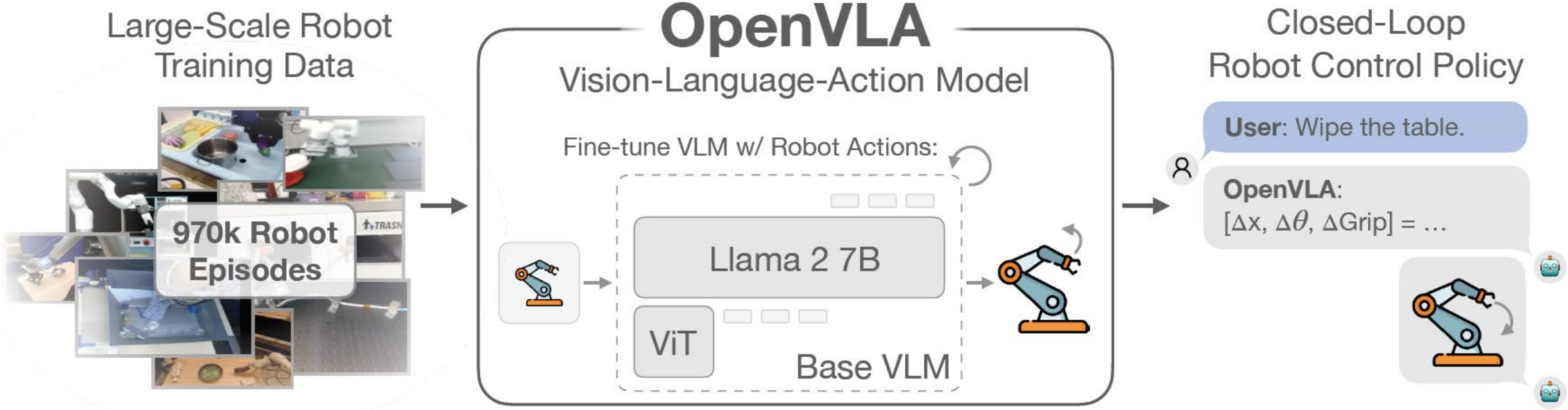
- RT-1-X mean success rate is 50% higher than that of either the Original Method or RT-1.
- RT-1 and RT-1-X have the same network architecture. Therefore, the performance increase can be attributed to co-training on the robotics data mixture.
- The lab logos indicate the physical location of real robot evaluation.
- The robot pictures indicate the embodiment used for the evaluation.



[4] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." In Conference on Robot Learning, pp. 2679-2713. PMLR, 2025.

OpenVLA

An open VLA to facilitate extensions to new robot embodiments and tasks



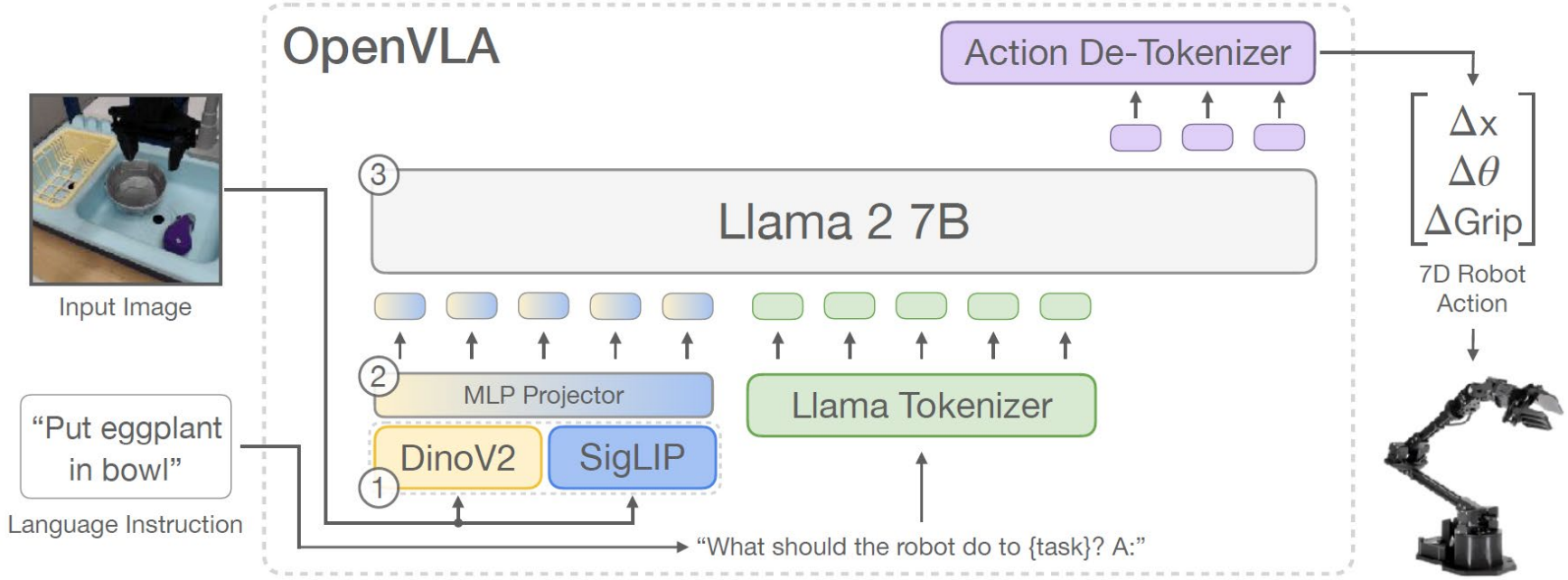
OpenVLA

Given an image observation and a language instruction, the model predicts 7-dimensional robot control actions. The architecture consists of three key components: (1) a vision encoder that concatenates Dino V2 and SigLIP features, (2) a projector that maps visual features to the language embedding space, and (3) the LLM backbone, a Llama 2 7B-parameter large language model.

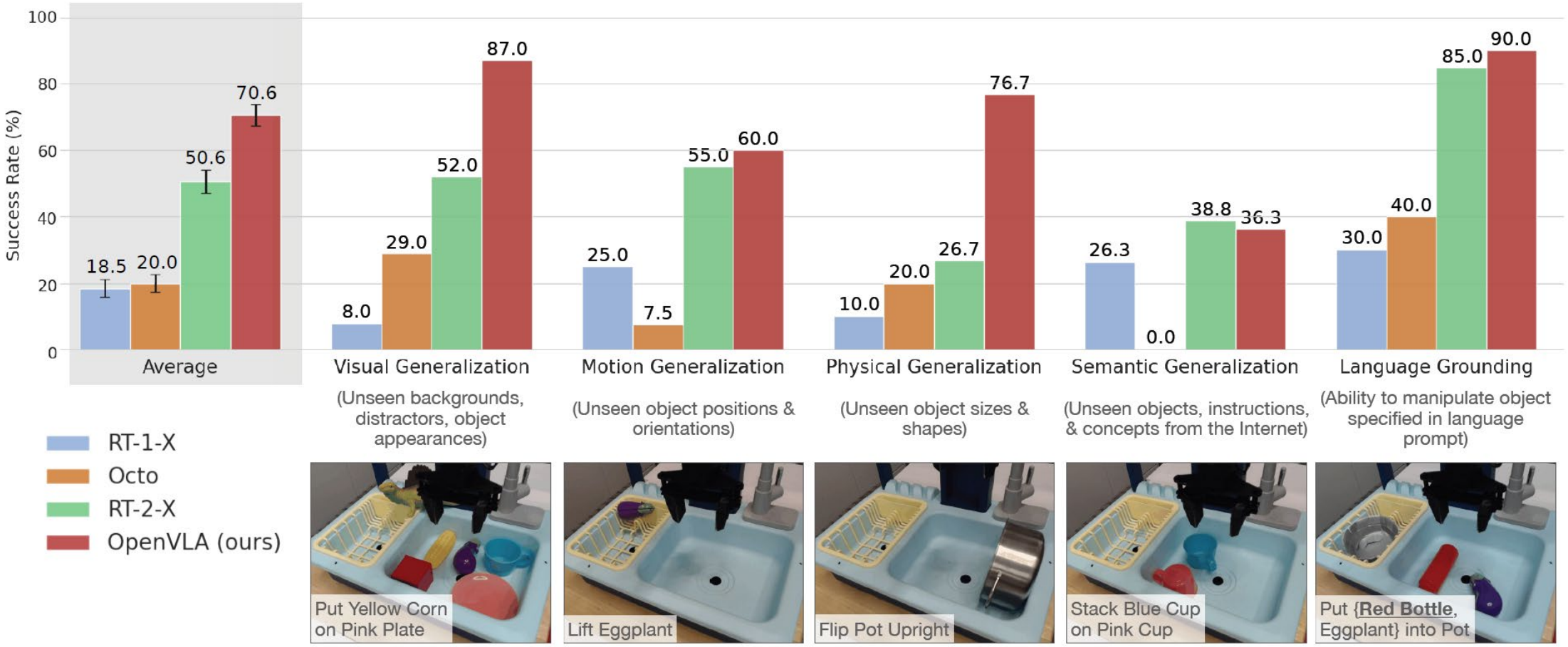
Contributions:

Larger training data.

More modular architecture



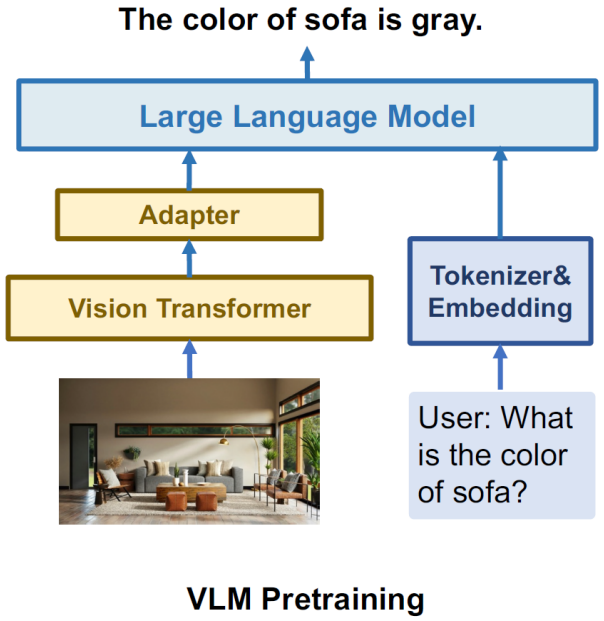
OpenVLA

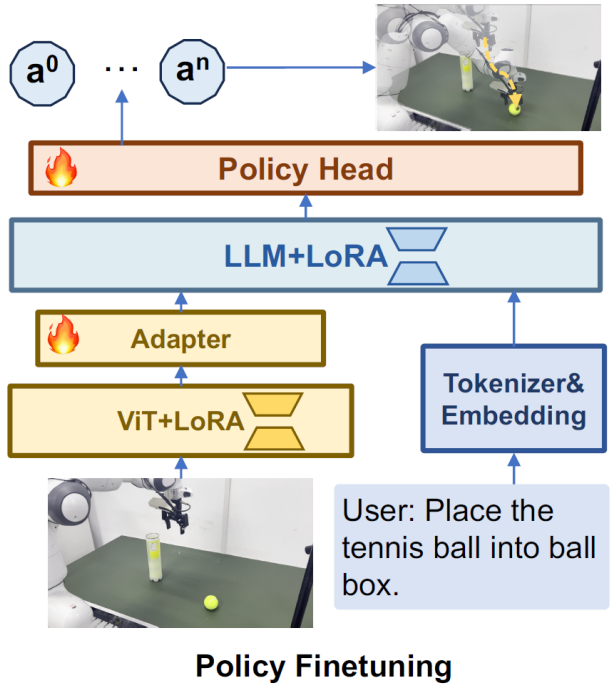

Success Rate (%)
100
80
60
40
20
0
18.5
20.0
50.6
70.6
Average
8.0
29.0
52.0
87.0
Visual Generalization
(Unseen backgrounds, distractors, object appearances)
25.0
7.5
55.0
60.0
Motion Generalization
(Unseen object positions & orientations)
10.0
20.0
26.7
76.7
Physical Generalization
(Unseen object sizes & shapes)
26.3
0.0
38.8
36.3
Semantic Generalization
(Unseen objects, instructions, & concepts from the Internet)
30.0
40.0
85.0
90.0
Language Grounding
(Ability to manipulate object specified in language prompt)
RT-1-X
Octo
RT-2-X
OpenVLA (ours)
Put Yellow Corn on Pink Plate
Lift Eggplant
Flip Pot Upright
Stack Blue Cup on Pink Cup
Put {Red Bottle, Eggplant} into Pot

[5] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu et al. "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation." *IEEE Robotics and Automation Letters* (2025).

TinyVLA

- Used a smaller vision and language model
- Used low-rank adaptors (LoRA) in fine-tuning to reduce the manipulated parameter space
- Used a new more efficient policy head based on a diffusion model to generate outputs



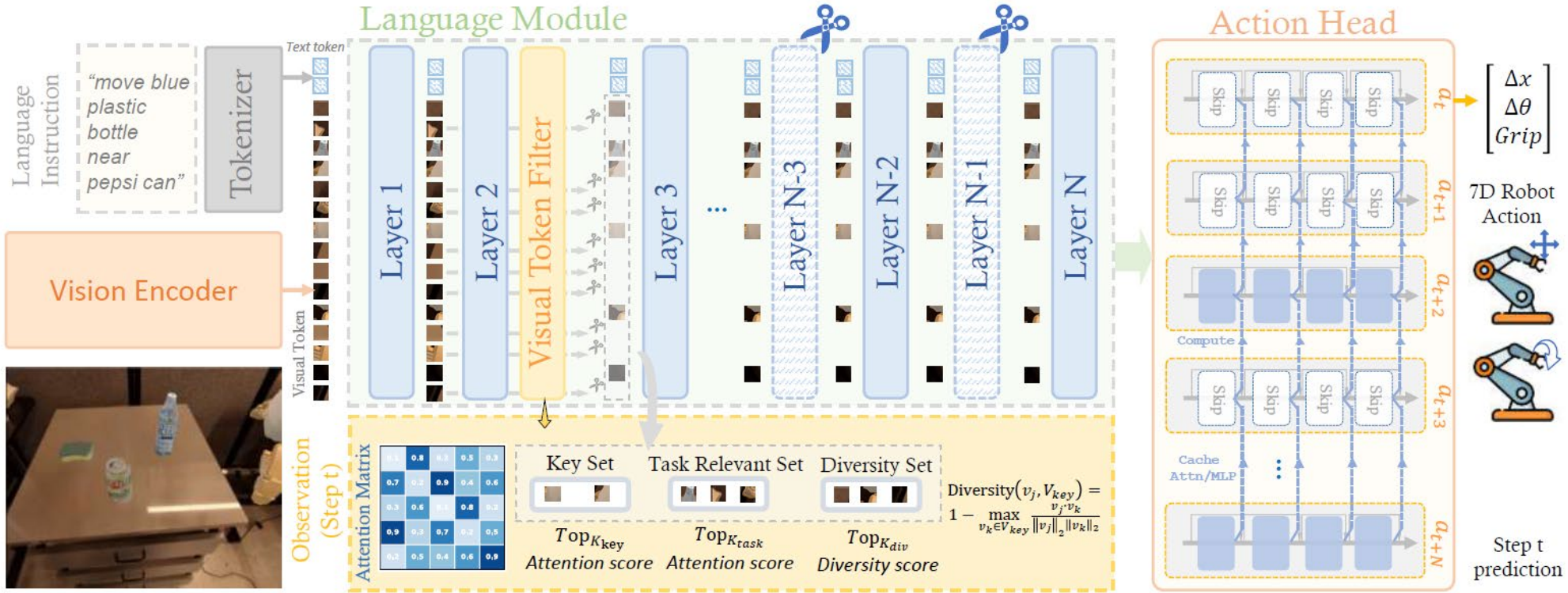


Model \ Tasks	Pre-trained Trajectory	Total Params	Trainable Params	RealWorld(5 tasks) PlaceTennis	FlipMug	StackCubes	CloseDrawer	OpenBox	Avg.
Diffusion Policy [3]	N/A	111M	111M	16.7±0.6	30±0.2	3.3±0.1	73.3±0.1	53.3±0.1	35.3
Multimodal Diffusion [38]	N/A	230M	230M	23.3±0.3	13.3±1.3	6.7±0.3	36.7±0.3	10.0±0	18.0
OpenVLA [10]	970K	7.2B	195M	83.3±1.1	51.7±3.1	40.0±0.1	85.0±1	81.7±0.6	68.3
TinyVLA-S	N/A	422M	101M	8.3±0.1	6.7±0.1	6.7±0.1	60.0±0.2	35.0±0.3	23.3
TinyVLA-B	N/A	740M	138M	76.7±0.6	76.7±0.1	71.7±0.1	81.7±0.1	80.0±0.2	77.4
TinyVLA-H	N/A	1.3B	143M	**90.0±0.2**	**98.3±0.1**	**98.3±0.1**	**96.7±0.3**	**86.7±0.1**	**94.0**

[6] Yantai Yang, Yuhao Wang, Zichen Wen, Luo Zhongwei, Chang Zou, Zhipeng Zhang, Chuan Wen, and Linfeng Zhang. "Efficientvla: Training-free acceleration and compression for vision-language-action models." arXiv preprint arXiv:2506.10100 (2025).

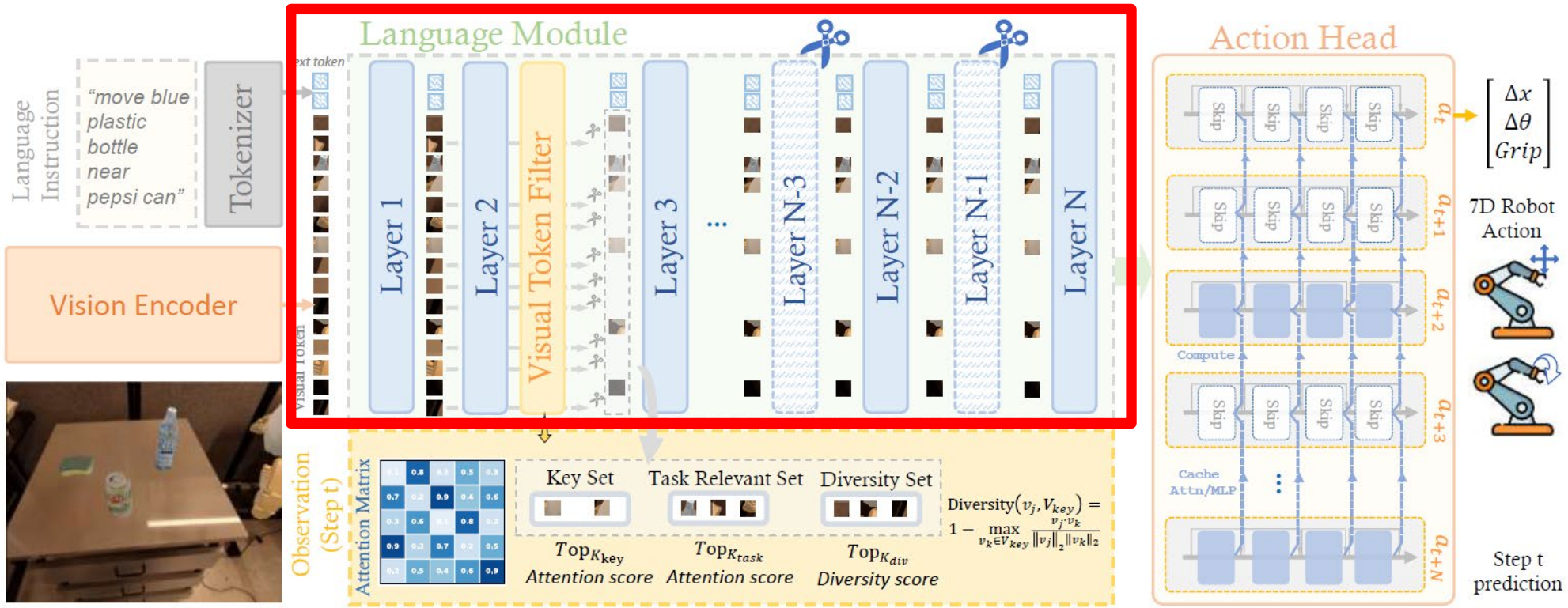
EfficientVLA

Additional efficiency gains due to: (1) pruning of redundant language module layers; (2) VLA task-aware visual token selection (balancing task relevance and informational diversity); and (3) temporal caching of intermediate features in the diffusion action head.



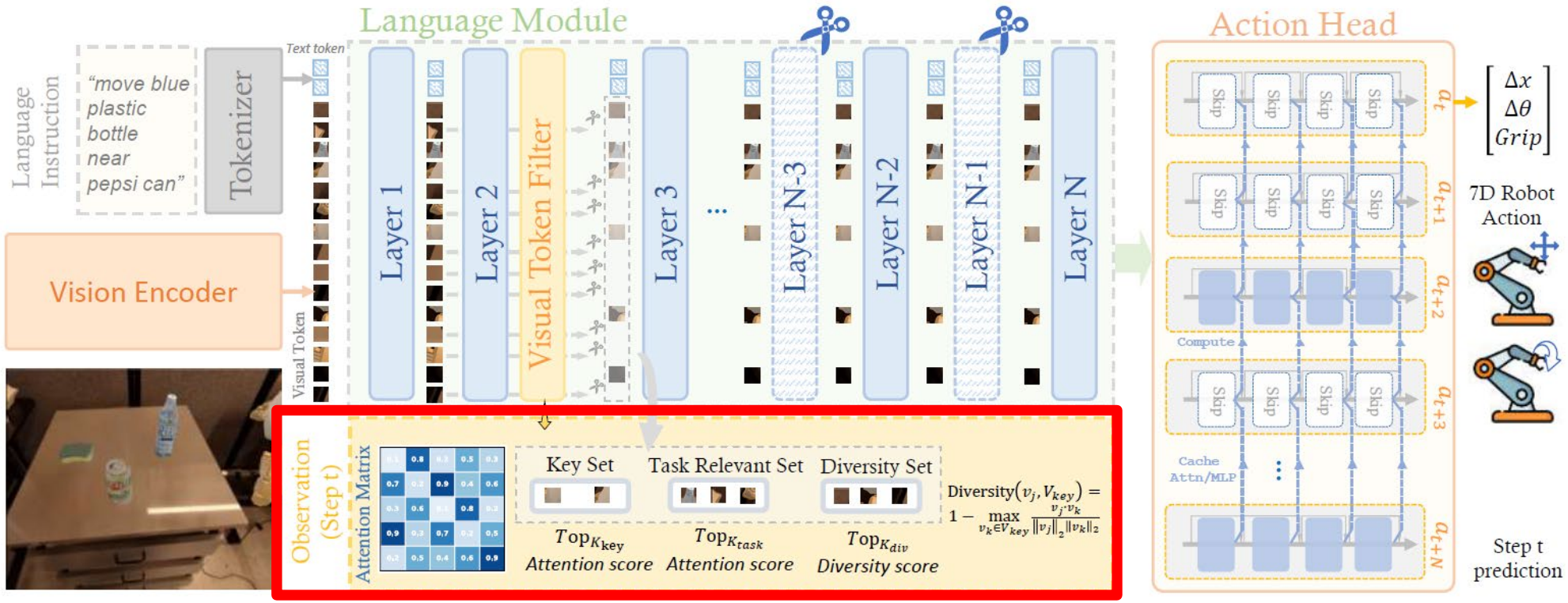
EfficientVLA

(1) Pruning: Remove layers with a high cosign similarity between their inputs and outputs



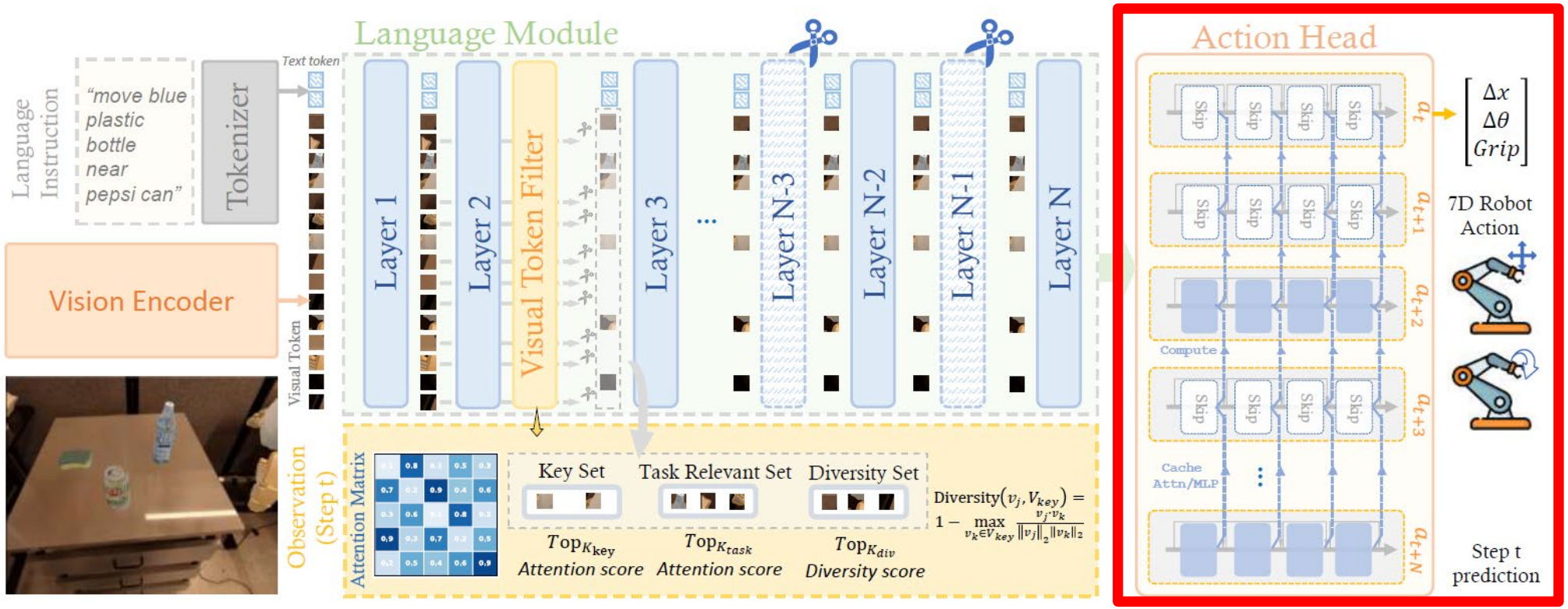
EfficientVLA

(2) VLA task-aware visual token selection: Prune irrelevant and redundant visual tokens.



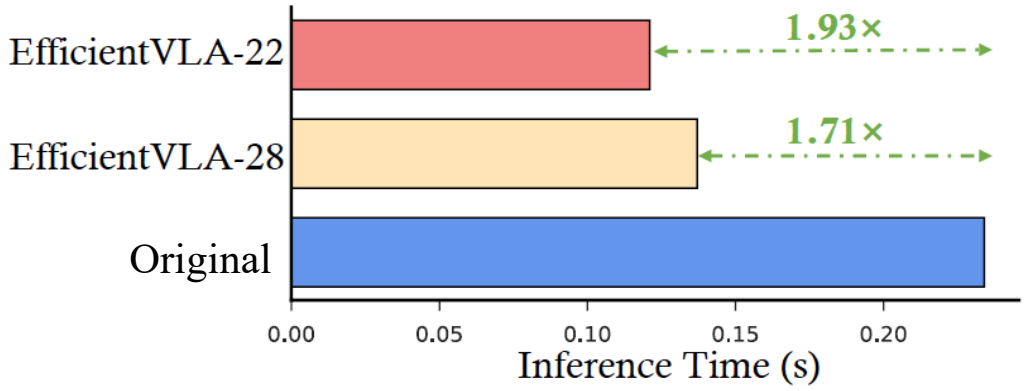
EfficientVLA

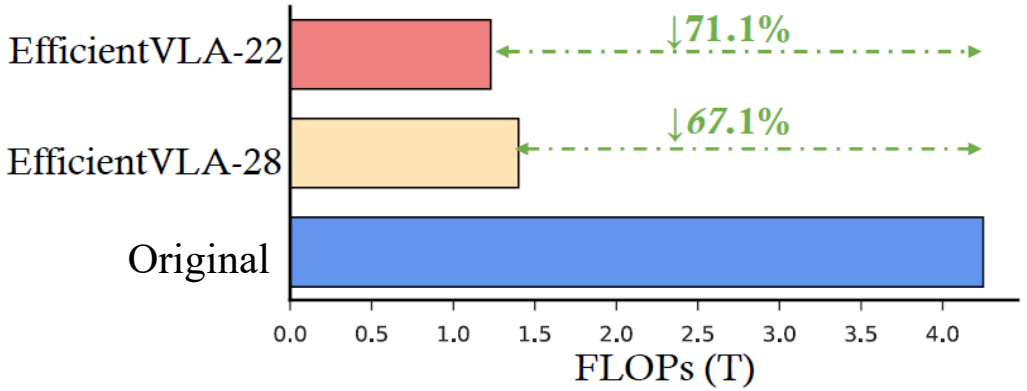
(3) Temporal Caching: Store intermediate features in the diffusion action head for N-1 out of N successive samples



Evaluation

1. Effect of model layer pruning: Comparing the original model to one reduced to 22 and 28 layers, respectively.





2. Effect of visual token reduction.

Token Ratio	56	72	96	112	256 100.0%
Accuracy	95.0%	95.3%	95.0%	96.0%	91.3%
Inference time (s)	0.1866	0.1870	0.1889	0.1956	0.2342
FLOPs (T)	1.76	1.96	2.25	2.45	4.19

3. Effect of cache interval.

Cache Interval	1	2	3	4	5
Accuracy	91.3%	94.0%	93.7%	90.3%	93.7%
Inference time (s)	0.2342	0.2031	0.1987	0.1953	0.1909
FLOPs (T)	4.190	4.161	4.155	4.150	4.144

Reminders and Announcements

Student-Led Project Presentations

- Starting next week (4/28)
- Each presentation is 15 minutes + 3-4 minutes Q&A.
- Content of the presentation:
  - **Introduction:** Summarize what you are working on, why it is important, how it is usually done (if applicable), and what your key innovation is. (This part is similar to the elevator pitch.)
  - **Design:** Detail the technical design. Motivate and explain design decisions. Do not just say: "Here is how we built it". Rather, explain why you decided to build it this way.
  - **Evaluation:** Show experimental results to demonstrate that the presented design achieved its stated objectives.
  - **Conclusion slide:** Add a slide on conclusions, lessons learning, and possible future directions for this work.

References

[1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan et al. "Rt-1: Robotics transformer for real-world control at scale." arXiv preprint arXiv:2212.06817 (2022).

[2] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." In Conference on Robot Learning, pp. 2165-2183. PMLR, 2023.

[3] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6892-6903. IEEE, 2024.

[4] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." In Conference on Robot Learning, pp. 2679-2713. PMLR, 2025.

[5] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu et al. "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation." IEEE Robotics and Automation Letters (2025).

[6] Yantai Yang, Yuhao Wang, Zichen Wen, Luo Zhongwei, Chang Zou, Zhipeng Zhang, Chuan Wen, and Linfeng Zhang. "Efficientvla: Training-free acceleration and compression for vision-language-action models." arXiv preprint arXiv:2506.10100 (2025).