Open X-Embodiment: Robotic Learning Datasets and RT-X Models

Open X-Embodiment: Robotic Learning Datasets and RT-X Models (robotics-transformer-x.github.io)

Backgrounds

- Large, high-capacity models trained on diverse datasets have shown remarkable successes on efficiently tackling downstream applications.

- In the domain of robotics, there is no pretrained models with general pretrained backbones serving as a starting point for downstream applications like those in NLP and CV.

- To achieve a general-purpose model in robotics, we need a large and diverse robotics dataset covering different environments, objects and tasks. **However, the existing largest robotics dataset lacks in size and diversity compared with those in CV and NLP.**

- Combination of data from different robots and environments provides a better coverage of variations in environments and robots. **Training over the cross-embodiment robotics data could be the solution to a general-purpose robotics model.**

Contributions

1. Demonstration for **the advantages of cross-embodiment training under a unified input and output scheme**, which enables efficient policy transfer.

2. Open source of the Open X-Embodiment (OXE) Repository, which includes a dataset with 22 different robotic embodiments from 21 different institutions.

Data Format Consolidation

- **Model inputs**: a history of recent images and language instructions.
- **Model outputs**: a normalized 7-dimensional action vector controlling the end-effector(x, y, z, roll, pitch, yaw, and gripper opening or the rates of these quantities).
- Images from different datasets are resized to the same resolution. For those with multiple views, choose the canonical one as inputs.
- Actions from different datasets are converted into **the normalized 7-DoF end-effector action** for discretization. The coordinate frames of actions across datasets are not aligned, and the action values could represent either absolute or relative positions or velocities as the original control scheme of the dataset.

The Open X-Embodiment Repository

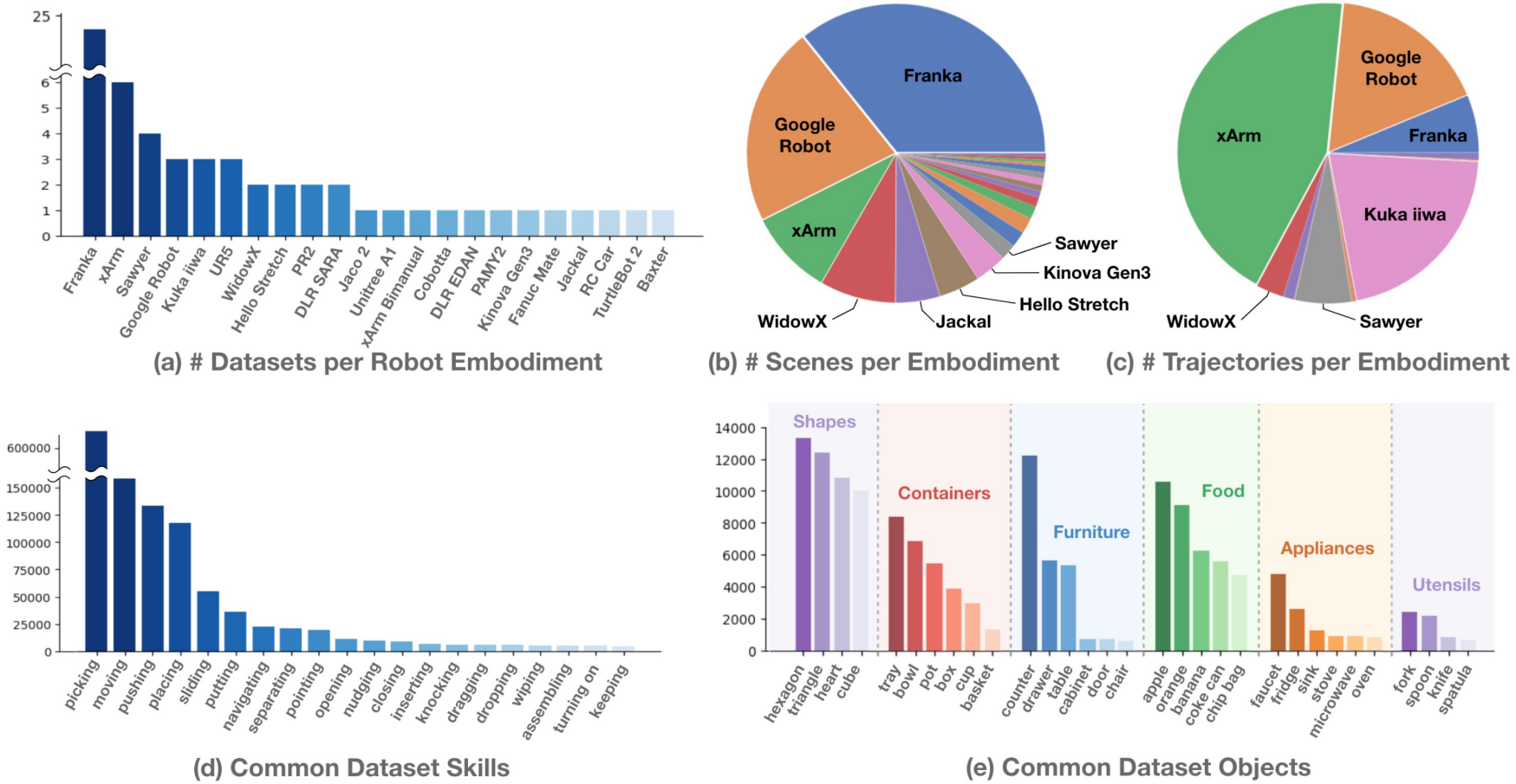


Fig. 2: The Open X-Embodiment Dataset. **(a)**: the dataset consists of 60 individual datasets across 22 embodiments. **(b)**: the Franka robot has the largest diversity in visually distinct scenes due to the large number of Franka datasets, **(c)**: xArm and Google Robot contribute the most number of trajectories due to a few large datasets, **(d, e)**: the dataset contains a great diversity of skills and common objects.

RT-X Design

The training details and architectures of RT-1-X and RT-2-X are similar to those of the original RT-1 and RT-2, while the major differences lie in the cross-embodiment dataset.

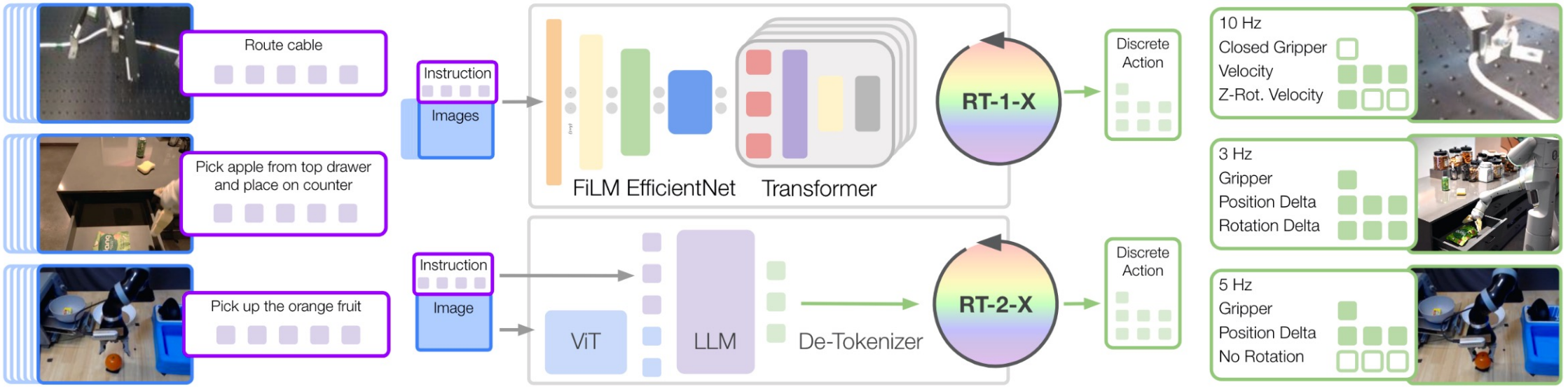


Fig. 3: RT-1-X and RT-2-X both take images and a text instruction as input and output discretized end-effector actions. RT-1-X is an architecture designed for robotics, with a FiLM [116] conditioned EfficientNet [117] and a Transformer [118]. RT-2-X builds on a VLM backbone by representing actions as another language, and training action text tokens together with vision-language data.

In-Distribution Evaluation

Evaluation results on domains with small-scale datasets

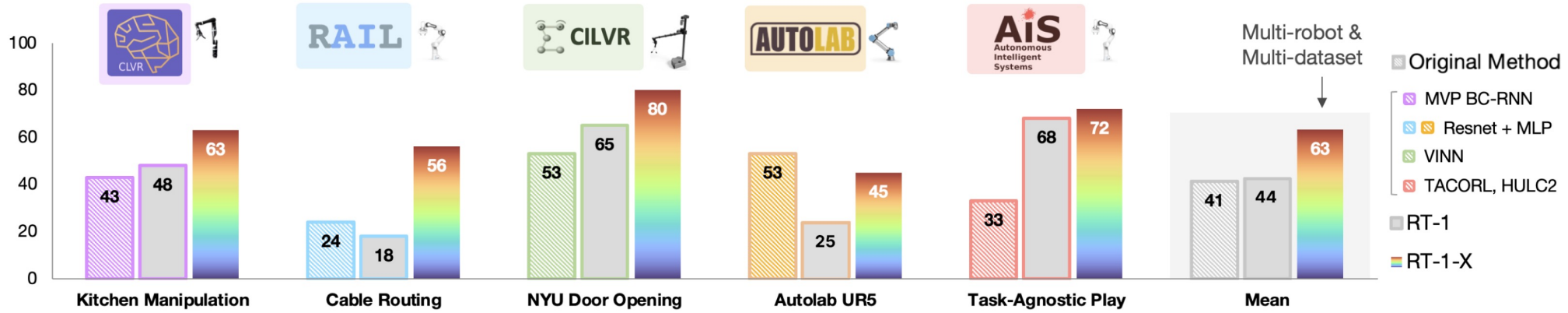


Fig. 4: RT-1-X mean success rate is 50% higher than that of either the Original Method or RT-1. RT-1 and RT-1-X have the same network architecture. Therefore the performance increase can be attributed to co-training on the robotics data mixture. The lab logos indicate the physical location of real robot evaluation, and the robot pictures indicate the embodiment used for the evaluation.

- Original method — the baseline model from the corresponding paper of the dataset
- RT-1 — an RT-1 model trained on the specific dataset
- RT-1-X — an RT-1 model trained on the cross-embodiment dataset

In-Distribution Evaluation

Evaluation results on domains with large-scale datasets

Evaluation Setting	Bridge	Bridge	RT-1 paper 6 skills
Evaluation Location	IRIS (Stanford)	RAIL Lab (UCB)	Google Robotic Lab
Robot Embodiment	WidowX	WidowX	Google Robot
Original Method	LCBC [95]	LCBC [95]	-
Original Method	13%	13%	-
RT-1	40%	**30%**	**92%**
RT-1-X	27%	27%	73%
RT-2-X (55B)	**50%**	**30%**	**91%**

TABLE I: Parameter count scaling experiment to assess the impact of capacity on absorbing large-scale diverse embodiment data. For these large-scale datasets (Bridge and RT-1 paper data), RT-1-X underfits and performs worse than the Original Method and RT-1. RT-2-X model with significantly many more parameters can obtain strong performance in these two evaluation scenarios.

Cross-embodiment training improves performance in the domains with large-scale datasets, only when utilizing a proper high-capacity architecture.

Out-of-Distribution Evaluation

Emergent skills evaluation: evaluate the model on one embodiment**(the Google Robot dataset, Google Robot)** while the skills come from another dataset**(the Bridge dataset)** with a different embodiment**(Widow X)**.

Row	Model	Size	History Length	Dataset	Co-Trained w/ Web	Initial Checkpoint	Emergent Skills Evaluation	RT-2 Generalization Evaluation
(1)	RT-2	55B	none	Google Robot action	Yes	Web-pretrained	27.3%	**62%**
(2)	RT-2-X	55B	none	Robotics data	Yes	Web-pretrained	**75.8%**	**61%**
(3)	RT-2-X	55B	none	Robotics data except Bridge	Yes	Web-pretrained	42.8%	54%
(4)	RT-2-X	5B	2	Robotics data	Yes	Web-pretrained	44.4%	52%
(5)	RT-2-X	5B	none	Robotics data	Yes	Web-pretrained	14.5%	30%
(6)	RT-2-X	5B	2	Robotics data	No	From scratch	0%	1%
(7)	RT-2-X	5B	2	Robotics data	No	Web-pretrained	48.7%	47%

TABLE II: Ablations to show the impact of design decisions on generalization (to unseen objects, backgrounds, and environments) and emergent skills (skills from other datasets on the Google Robot), showing the importance of Web-pretraining, model size, and history.

- (1) & (2): cross-embodiment training improves the range of tasks even in domains with large amount of data.
- (2) & (3): source of new skills (the Bridge dataset) is crucial in such transfer behavior.
- (2) & (5): architecture capacity matters.
- (4) & (5): the history could improve generalization performance.
- ……

Limitations

1. No generalization experiments over a new embodiment.

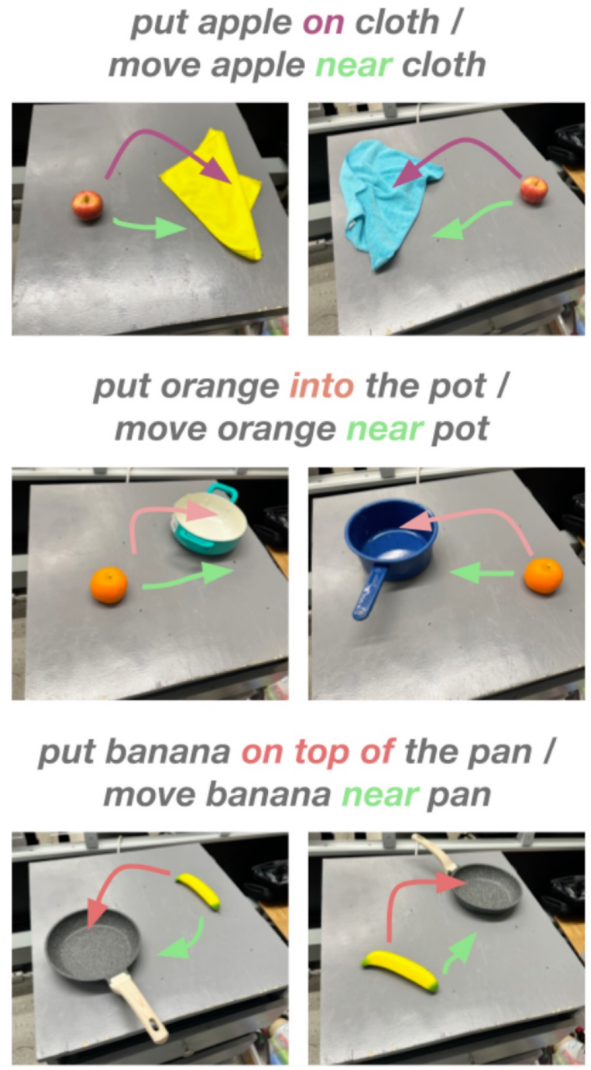
2. Embodiment control conditioned on single image and language instructions, without the rich data from different sensing and actuation modalities that could be possibly helpful.

3. No criterion for whether positive transfer happens.

Discussion

- **Assumption: necessary abilities** to finish a vision-language conditioned robotic manipulation task

    1. **Entity of interest extraction and vision-language alignment,** including the embodiment, the object to interact with, obstacles, and the background.
    2. **Recognition of spatial relationships among entities,** including the knowledge of the relative positions and orientations of the robot's embodiment(the end effector), objects, and obstacles.
    3. **Plan of the next possible move:** Based on the information above, the robot needs to be able to plan possible next actions effectively to accomplish the task.

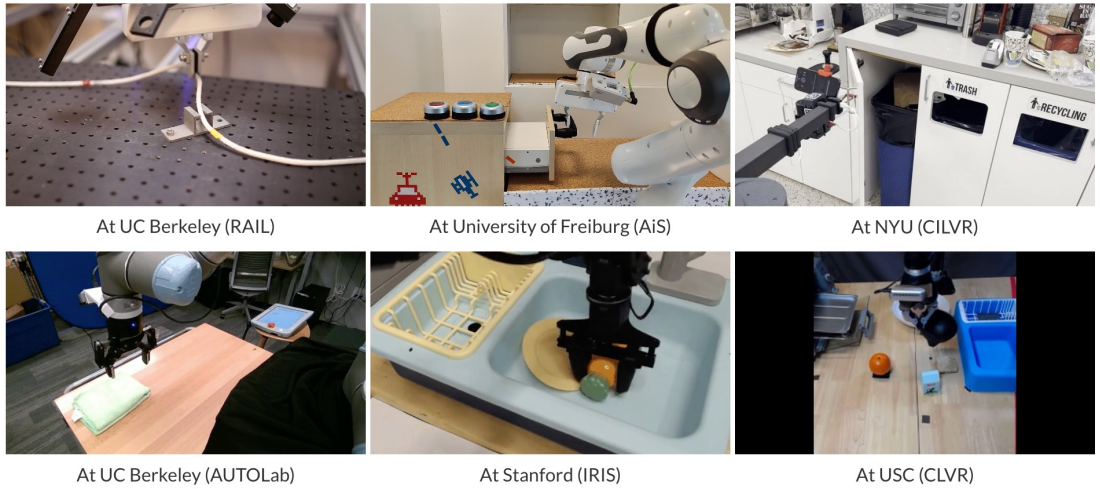


Discussion

- **How does the transfer happen?**

  Differences between the embodiments no longer matter, as the action spaces have been unified. The core task of the model is guiding the end effector to the right position with the correct orientation. However, the necessary abilities to complete the task are essentially **embodiment-agnostic**.

  All the involved embodiments can be seen as combinations of an end effector and a base. The unified 7-DoF end effector action space liberates the model from trivial low-level joint control, enabling it to transfer among embodiments and learn from other datasets.



At UC Berkeley (RAIL) | At University of Freiburg (AiS) | At NYU (CILVR)

At UC Berkeley (AUTOLab) | At Stanford (IRIS) | At USC (CLVR)

Discussion

- **Utilization of information from other modalities**
  - Depth, joint state, kinematic information from urdfs, image instruction, etc.

- **Few/Zero-shot transfer to new embodiment**
  - Differences in action space and camera poses
  - New challenging embodiment which do not follow the assumption.

- **Learn from diverse human video datasets**
  - How to learn from unlabeled human data?
  - Suppose the assumption is correct, how to design an effective self-supervised training scheme?