RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

Presenter: Ming Liu

Oct 24, 2023

RT-2: New Model Translates Vision and Language into Action

Robotic Transformer 2 (RT-2) is a novel **vision-language-action (VLA)** model that learns from both web and robotics data, and translates this knowledge into generalized instructions for robotic control.

Express lower-level **robotic actions as text tokens**

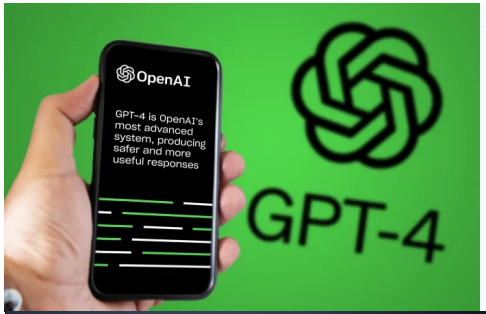
Ability to interpret commands **not present** in robot training; Perform **rudimentary reasoning** in response to user commands

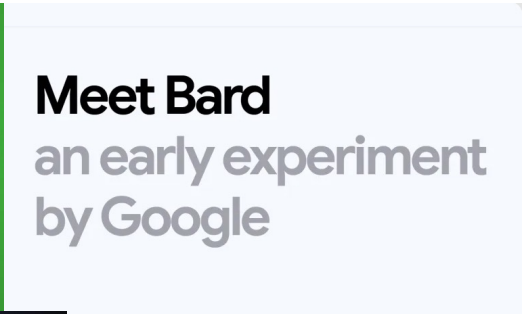
Motivation and Main Problem

High-capacity vision-language models (VLMs) are trained on **web-scale datasets**, making these systems remarkably good at recognizing visual or language patterns and operating across different languages.

But for robots to achieve a similar level of competency, we would need to collect **robot data**, first-hand, across every object, environment, task, and situation.

RT-X: the largest open-source robot dataset ever compiled, across 33 institutes, 22 robot hardware, 527 skills, and 1M episodes.

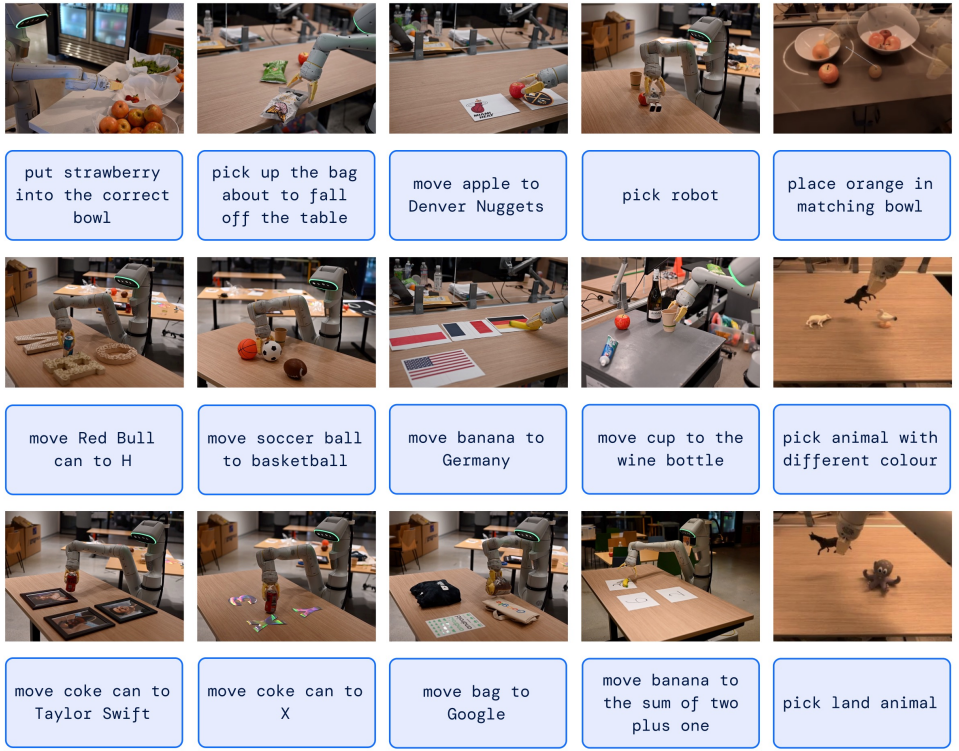




Motivation and Main Problem

Robots require grounded **low-level actions**, such as Cartesian end-effector commands

How **large pre-trained vision-language models** trained on **Internet-scale data** can be incorporated directly into **end-to-end low-level** robotic control to boost **generalization and enable emergent semantic reasoning**?



Examples of emergent robotic skills that are not present in the robotics data and require knowledge transfer from web pre-training

Related Work

- Vision-language models (VLMs)

  (1) representation-learning models, e.g. CLIP (Radford et al., 2021): learn common embeddings for both modalities

  (2) **visual language models of the form {vision, text} → {text}** which learn to take vision and language as input and provide free-form text

  - Generally trained on many different tasks, such as image captioning, **vision-question answering** (VQA), and general language tasks on multiple datasets at the same time.

Q: What is happening in the image?

A grey donkey walks down the street.

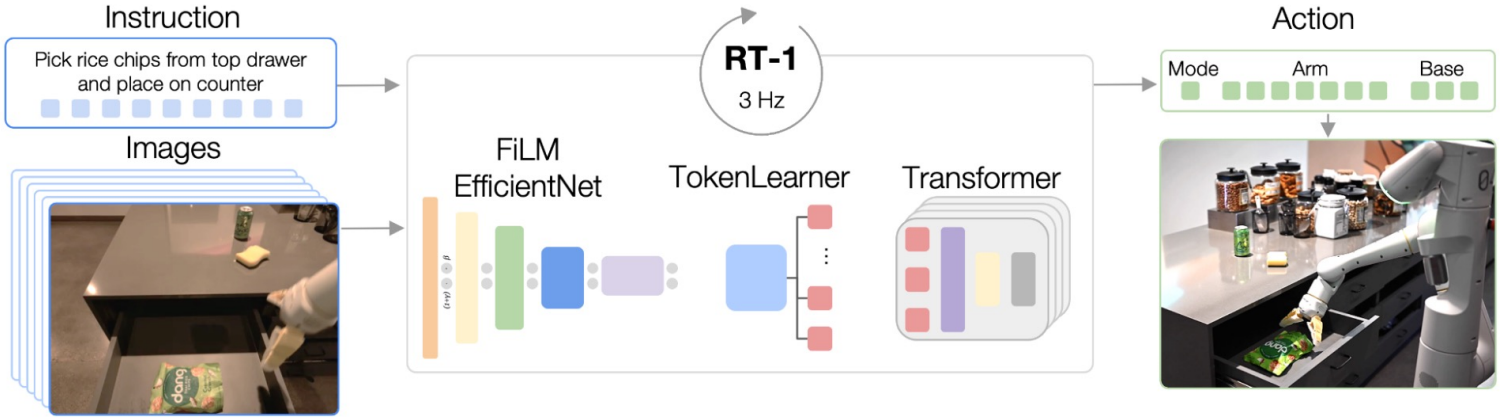
Related Work

- **Generalization** in robot learning
  - Prior methods have demonstrated how robots can generalize to novel object instances, to tasks involving novel combinations of objects and skills, to new goals or language instructions, to tasks with novel semantic object categories, to unseen environments.
  - RT-2: **A single model** that can generalize to **unseen conditions** along all of these axes.
  - Leverage pre-trained models that have been exposed to data that is much broader than the data seen by the robot.

- Pre-training for robotic manipulation.
  - Prior approaches use VLMs for visual state representations, for identifying, for high-level planning, or for providing supervision or success detection
  - RT-2: Not rely on a restricted 2D action space and **Not require a calibrated camera**
  - Leverage VLMs that generate language, and the unified output space of our formulation enables model **weights** to be entirely **shared** across language and action tasks, without introducing action-only model layer components.

Robotic Transformer 2 (RT-2)

Robotic Transformer 2 (RT-2) builds upon Robotic Transformer 1 (RT-1), a model trained on multi-task demonstrations, which can learn combinations of tasks and objects seen in the robotic data. More specifically, their work used RT-1 robot demonstration data that was collected with 13 robots over 17 months in an office kitchen environment.



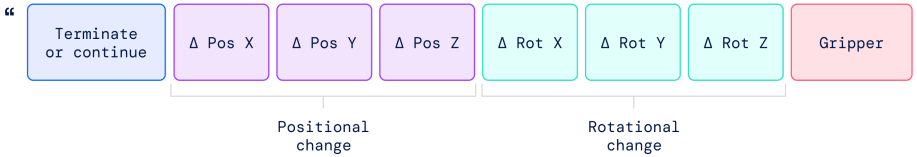
RT-1's architecture: The model takes a text instruction and set of images as inputs, encodes them as tokens via a pre-trained FiLM EfficientNet model and compresses them via TokenLearner. These are then fed into the Transformer, which outputs action tokens.

Vision-Language-Action Models

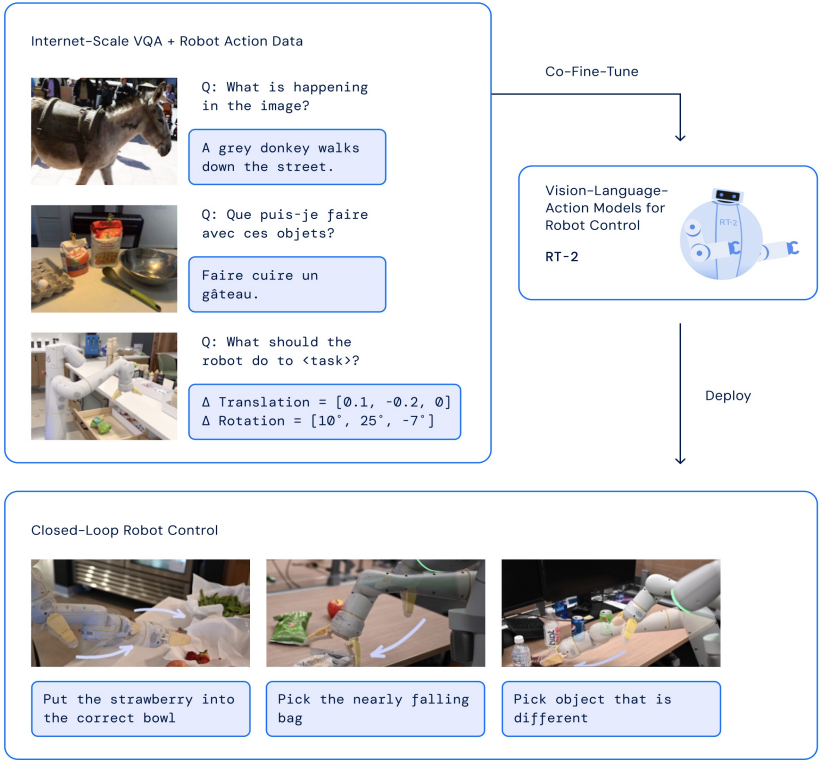
Adapting VLMs for robotic control:

Pathways Language and Image model (PaLI-X) (55B parameters) & Pathways Language model Embodied (PaLM-E) (12B parameters), to act as the backbones of RT-2.

RT-2 architecture and training: co-fine-tune a pre-trained VLM model on robotics and web data. The resulting model takes in robot camera images and directly predicts actions for a robot to perform.



Representation of an action string example: a sequence of robot action token numbers, e.g.“1 128 91 241 5 101 127 217”. The continuous dimensions (except for the discrete termination command) are discretized into 256 bins uniformly.

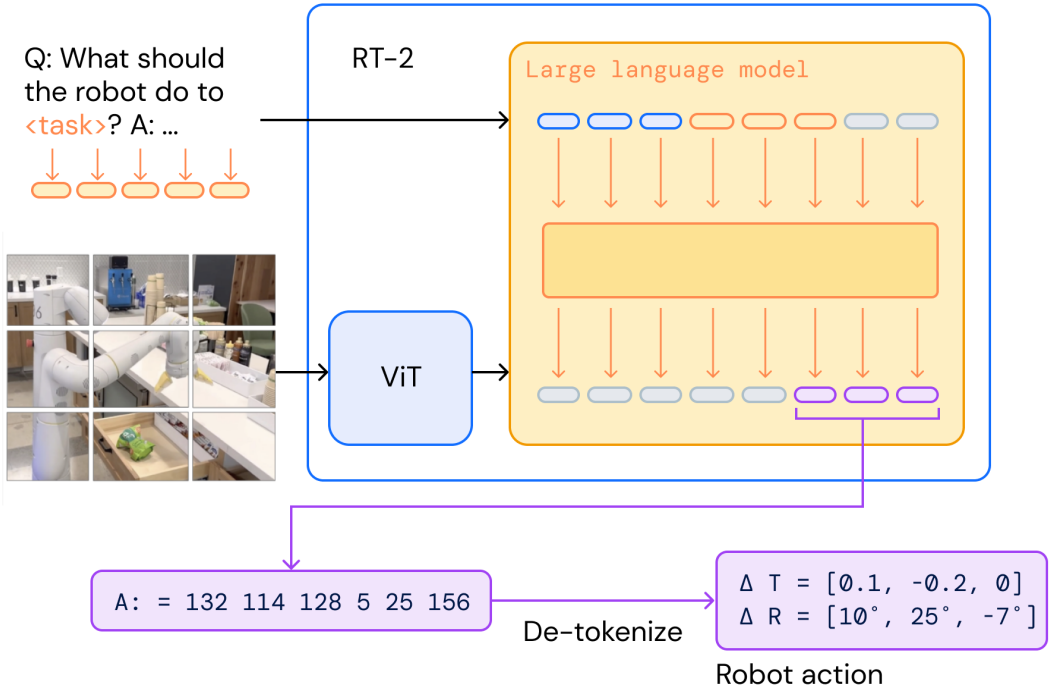


Vision-Language-Action Models

During inference, the text tokens are **de-tokenized into robot actions**, enabling closed loop control.

Leverage the backbone and pre-training of vision-language models in learning robotic policies, transferring some of their generalization, semantic understanding, and reasoning to robotic control.

To ensure that RT-2 outputs valid action tokens during decoding, constrain its output vocabulary via only sampling valid action tokens when the model is prompted with a robot-action task

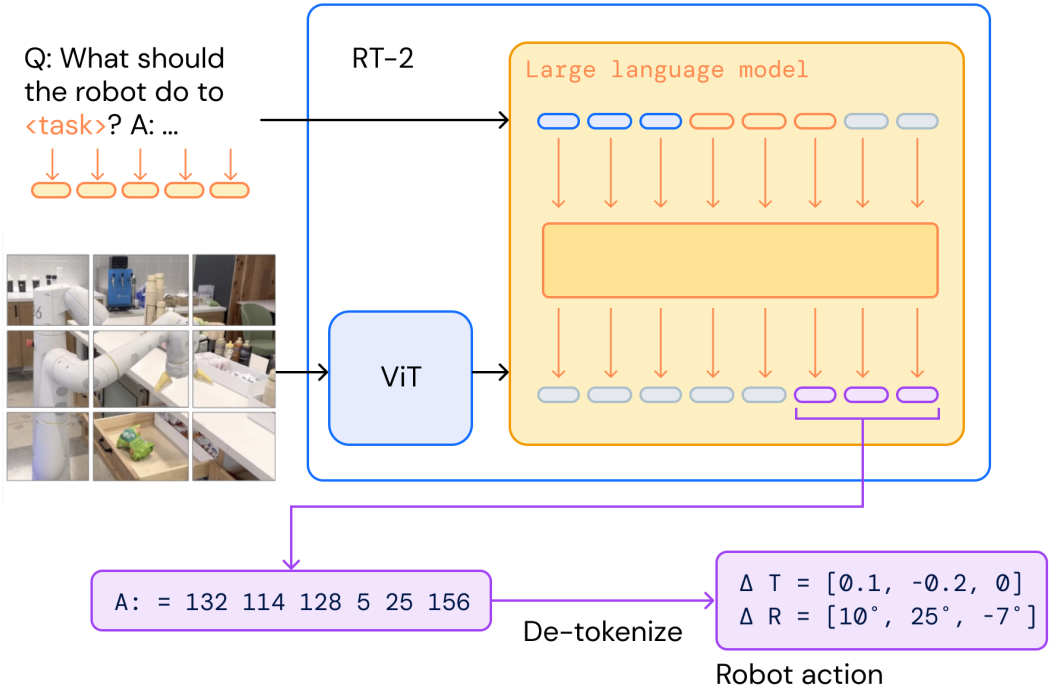


Real-Time Inference

Develop a protocol to run RT-2 models on robots by deploying them in a **multi-TPU cloud service** and querying this service over the network.

Achieve a suitable frequency of control and also serve multiple robots using the **same cloud service**.

The largest model, the **55B parameter RT-2-PaLI-X-55B model**, can run at a frequency of **1-3 Hz**. The smaller version of that model, consisting of 5B parameters, can run at a frequency of around **5 Hz**.



Experimental Setup

The experiments focus on real-world generalization and emergent capabilities of RT-2 and aim to answer the following questions:

1. How does RT-2 perform on seen tasks and more importantly, generalize over **new objects, backgrounds, and environments**?
2. Can we observe and measure any **emergent capabilities** of RT-2?
3. How does the generalization vary with parameter count and other design decisions?
4. Can RT-2 exhibit signs of **chain-of-thought reasoning** similarly to vision-language models?

Train two specific instantiations of RT-2 that leverage pre-trained VLMs:

1. **RT-2-PaLI-X** is built from 5B and 55B PaLI-X (Chen et al., 2023a)
2. **RT-2-PaLM-E** is built from 12B PaLM-E (Driess et al., 2023)

Evaluate the approach and several baselines with about **6,000 evaluation trajectories** in a variety of conditions

Experimental Setup

Training Data:

- Leverage the original **web scale data** from Chen et al. (2023a) and Driess et al. (2023), which consists of visual question answering, captioning, and unstructured interwoven image and text examples;
- **Robot demonstration data** from Brohan et al. (2022), which was collected with 13 robots over 17 months in an office kitchen environment.
- Each robot demonstration trajectory is annotated with a natural language instruction that describes the task performed, consisting of a verb describing the skill (e.g., “pick”, ”open”, “place into”) and one or more nouns describing the objects manipulated (e.g., “7up can”, “drawer”, “napkin”)
- For all RT-2 training runs they adopt the hyperparameters from the original **PaLI-X** (Chen et al., 2023a) and **PaLM-E** (Driess et al., 2023) papers, including learning rate schedules and regularizations.

Baselines:

- **RT-1** (35M parameter transformer-based model)
- To compare against state-of-the-art pre-trained representations, they use **VC-1** (Majumdar et al., 2023a) and **R3M** (Nair et al., 2022b), with policies implemented by training an RT-1 backbone to take their representations as input.
- To compare against other architectures for using VLMs, they use **MOO** (Stoneet al., 2023), which uses a VLM to create an additional image channel for a semantic map, which is then fed into an RT-1 backbone.

RT-2 Performance

Seen tasks category:

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

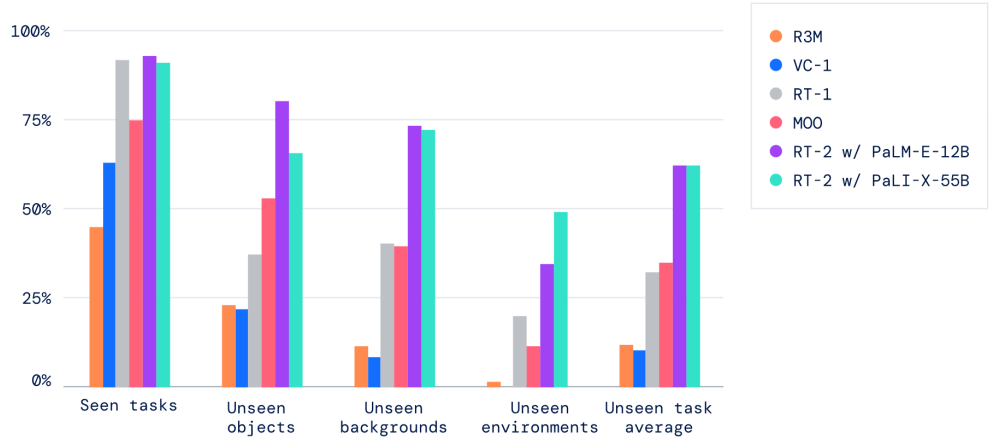
Unseen tasks category:

- Evaluations consists of over 280 tasks that focus primarily on pick and placing skills in many diverse scenarios.
- Example generalization evaluations, which are split into unseen categories (objects, backgrounds, and environments), and are additionally split into easy and hard cases.

Unseen objects | Unseen backgrounds | Unseen environments

RT-2 Performance

The performance on seen tasks is similar between the RT-2 models and RT-1, with other baselines attaining a lower success rate.
The difference between the RT-2 models and the baseline is most pronounced in the various generalization experiments, suggesting that the strength of vision-language-action models lies in transferring more generalizable visual and semantic concepts from their Internet-scale pre-training data.



Overall performance of two instantiations of RT-2 and baselines across seen training tasks as well as unseen evaluations measuring generalization to novel objects, novel backgrounds, and novel environments.

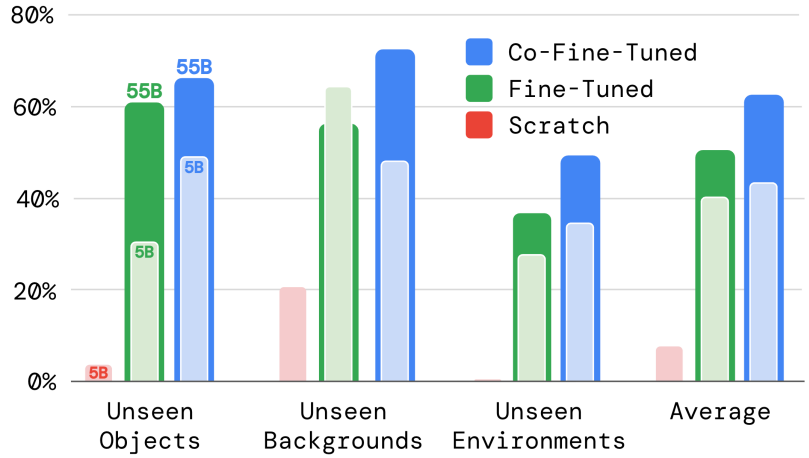
RT-2 Performance

The results below demonstrate a significant improvement of RT-2 compared to the baselines (3x).

RT-2 Performance

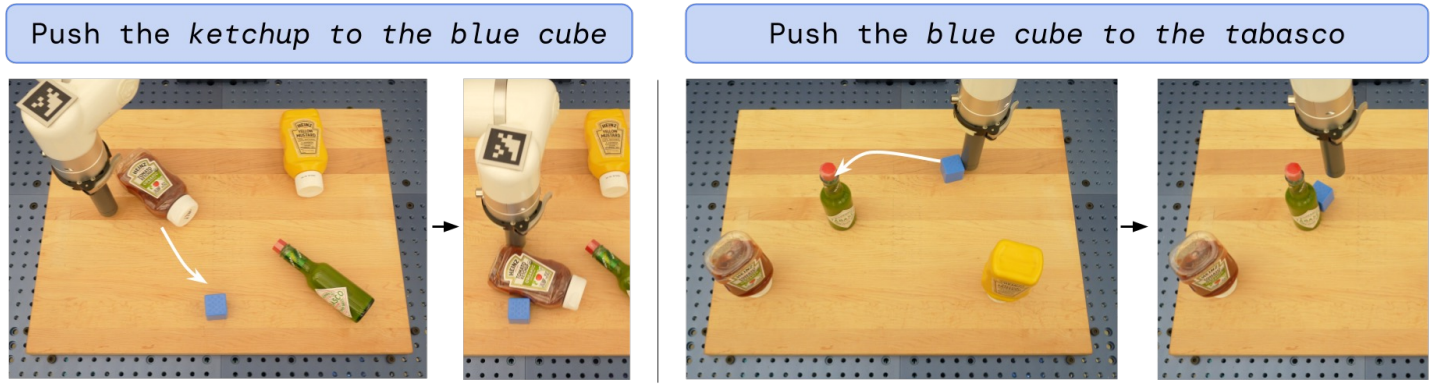
To better understand how different design choices of RT-2 impact the generalization results they ablate the two most significant design decisions:

- the model size: 5B vs 55B for the RT-2 PaLI-X variant
- training recipe: training the model from scratch vs fine-tuning vs co-fine-tuning.
- The results below indicate the importance of the pre-trained weights of the vision-language model as well as the trend of the model generalization improving with the model size.



RT-2 Performance

They also evaluate RT-2 on an open-source language-table benchmark where they train RT-2 on simulation and real language-table data.
In addition to achieving the state-of-the-art result on the simulation benchmark (90% vs 77% of the previous SoTA), they evaluate the resulting model in the real world, and demonstrate RT-2's generalization capabilities with the objects never seen in language table datasets before such as ketchup bottle, banana and others:



RT-2 Performance

Since the resulting RT-2 PaLM-E version of the model is a vision-language-action model that can act as an LLM, VLM and a robotic controller all in a single neural network, they demonstrate that RT-2 can perform **chain-of-thought reasoning** for control.

The example shows RT-2 first outputs a few reasoning steps in natural language which are then followed by the string: `Action:` and the resulting action tokens.

Limitations

- Although including web-scale pre-training via VLMs boosts generalization over semantic and visual concepts, the robot **does not acquire any ability to perform new motions** by virtue of including this additional experience.
- The model's physical skills are still **limited to the distribution of skills seen in the robot data**, but it learns to deploy those skills in new ways.
- An exciting direction for future work is to study how new skills could be acquired through new data collection paradigms such as video of humans.

- Although it could run large VLA models in real time, **the computation cost of these models is high**, and as these methods are applied to settings that demand high-frequency control, real-time inference may become a major bottleneck.
- An exciting direction for future research is to explore quantization and distillation techniques that might enable such models to run at higher rates or on lower-cost hardware.

- There are only a small number of generally available VLM models that can be used to create RT-2.

Conclusions

- Demonstrated the effectiveness of Vision-Language-Action (VLA) models
- Combined vision-language model (VLM) pre-training with robotic data
- Introduced RT-2-PaLM-E and RT-2-PaLI-X VLA models
- Fine-tuned models with robotic trajectory data to generate textual robot actions
- **Achieved exceptional robotic policy performance**
- **Enhanced generalization and knowledge transfer from web-scale pre-training**
- Foresee robotics benefiting from improved vision-language models
- Positioned robot learning at the forefront of advancements in various fields.

- Large Language Models (LLMs) have excelled as high-level semantic planners for sequential decision-making tasks. How to harness them to learn complex low-level manipulation tasks, such as dexterous pen spinning?