https://robotics-transformer2.github.io

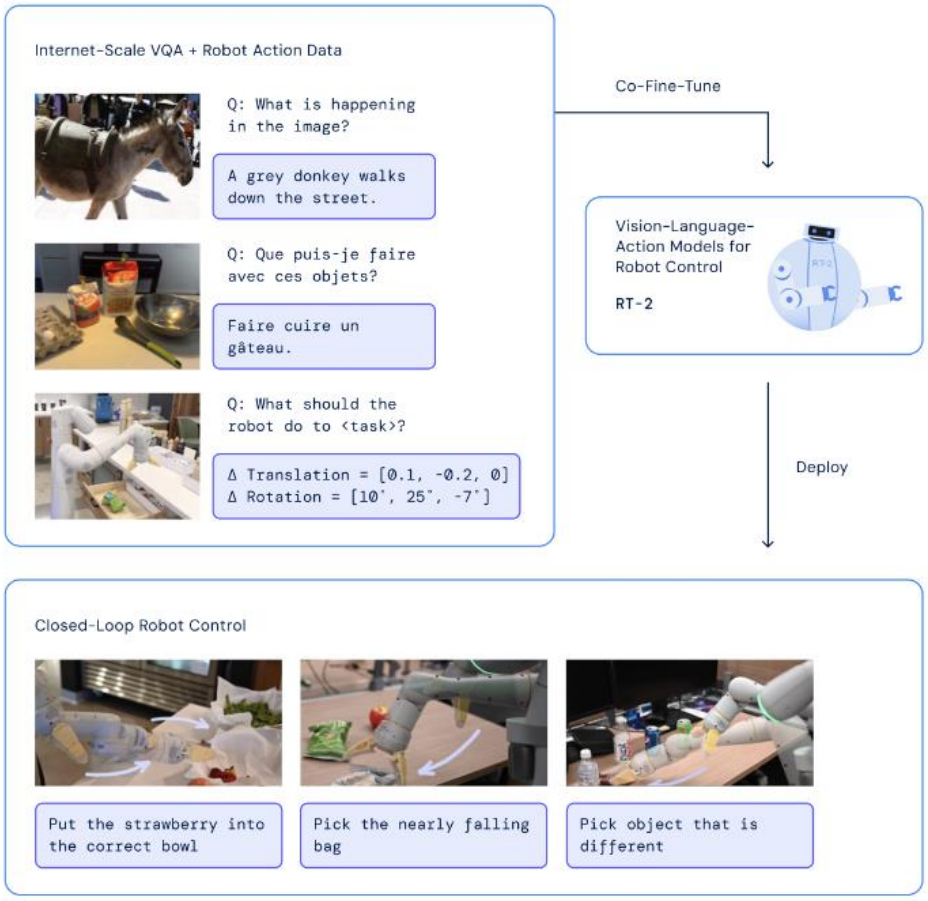
RT-2: Vision-Language-Action Models

Transfer Web Knowledge to Robotic Control

Abstract

We study how vision-language models trained on Internet-scale data can be incorporated directly into end-to-end robotic control to boost generalization and enable emergent semantic reasoning. Our goal is to enable a single end-to-end trained model to both learn to map robot observations to actions and enjoy the benefits of large-scale pretraining on language and vision-language data from the web. To this end, we propose to co-fine-tune state-of-the-art vision-language models on both robotic trajectory data and Internet-scale vision-language tasks, such as visual question answering. In contrast to other approaches, we propose a simple, general recipe to achieve this goal: in order to fit both natural language responses and robotic actions into the same format, we express the actions as text tokens and incorporate them directly into the training set of the model in the same way as natural language tokens. We refer to such category of models as vision-language-action models (VLA) and instantiate an example of such a model, which we call RT-2. Our extensive evaluation (6k evaluation trials) shows that our approach leads to performant robotic policies and enables RT-2 to obtain a range of emergent capabilities from Internet-scale training. This includes significantly improved generalization to novel objects, the ability to interpret commands not present in the robot training data (such as placing an object onto a particular number or icon), and the ability to perform rudimentary reasoning in response to user commands (such as picking up the smallest or largest object, or the one closest to another object). We further show that incorporating chain of thought reasoning allows RT-2 to perform multi-stage semantic reasoning, for example figuring out which object to pick up for use as an improvised hammer (a rock), or which type of drink is best suited for someone who is too sleepy (an energy drink).

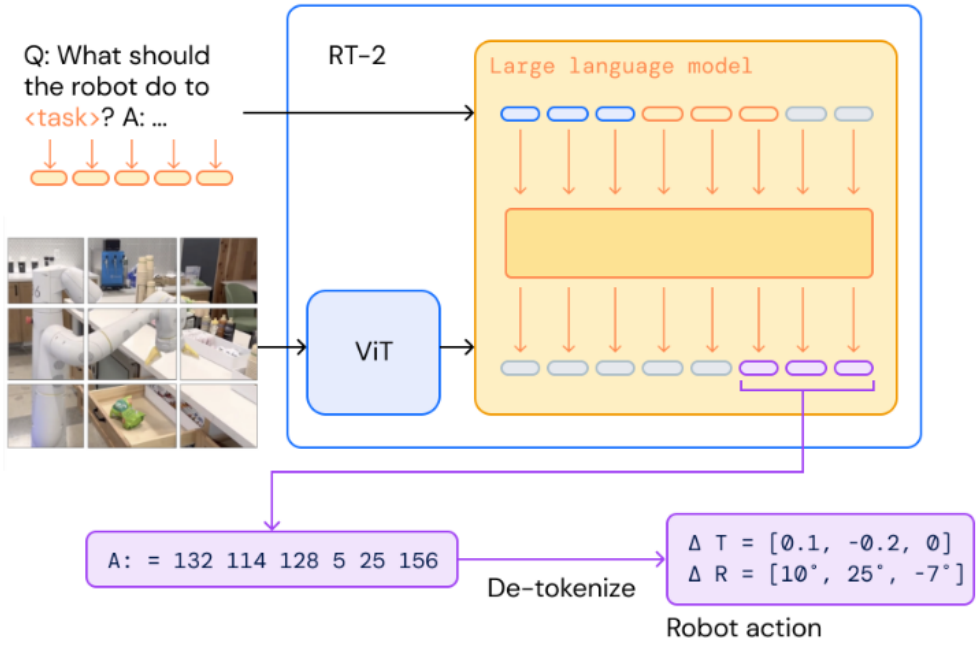
Approach Overview



To make RT-2 easily compatible with large, pre-trained vision-language models, our recipe is simple: we represent robot actions as another language, which can be cast into text tokens and trained together with Internet-scale vision-language datasets. In particular, we co-fine-tune (a combination of fine-tuning and co-training where we keep some of the old vision & text data around) an existing vision-language model with robot data. The robot data includes the current image, language command and the robot action at the particular time step. We represent the robot actions as text strings as shown below. An example of such a string could be a sequence of robot action token numbers: "1 128 91 241 5 101 127 217".

Since actions are represented as text strings, one can think of them as another language that allows us to operate the robot. This simple representation makes it straightforward to fine-tune any existing vision-language model and turn it into a vision-language-action model

During inference, the text tokens are de-tokenized into robot actions, enabling closed loop control. This allows us to leverage the backbone and pretraining of vision-language models in learning robotic policies, transferring some of their generalization, semantic understanding, and reasoning to robotic control.



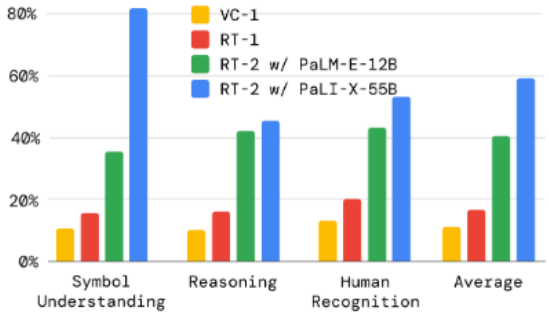
Results

We start the evaluation of RT-2 with testing the emergent properties of the model. Since we can't fully anticipate the extend of RT-2's generalization, we present a number of previously unseen objects to the robot and evaluate its performance on tasks that require semantic understanding that goes far beyond the robot data that the model was fine-tuned on. You can see qualitative examples of successful tasks that we found surprising below:

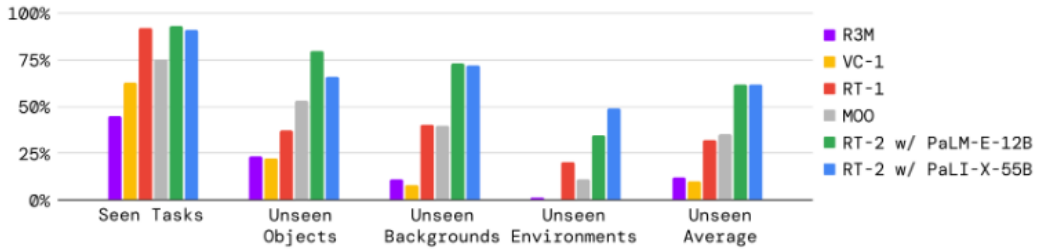
To quantify the emergent properties of RT-2, we categorize them into: symbol understanding, reasoning and human recognition and evaluate two variants of RT-2:

- RT-2 trained on top of PaLM-E (12B parameters),
- RT-2 trained on top of PaLI-X (55B parameters)

against its predecessor - RT-1 and another visual pre-training method - VC-1. The results below demonstrate a significant improvement of RT-2 compared to the baselines (3x).



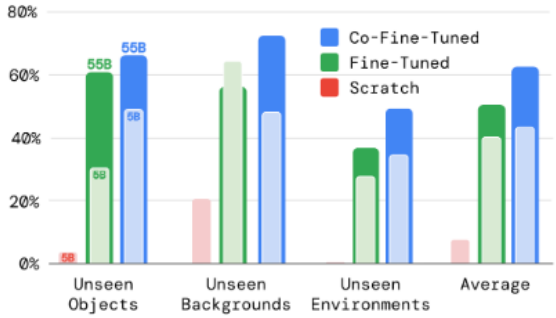
We evaluate the two variants of RT-2, together with more baselines in a blind A/B study and present the results across multiple generalization axes below. The resulting generalization improvement of RT-2 is approximately 2x.



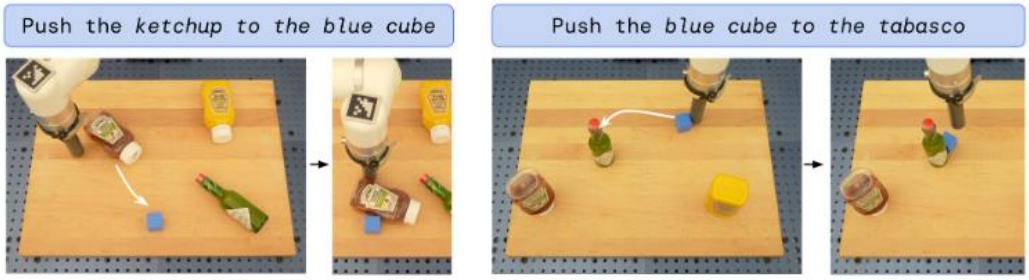
To better understand how different design choices of RT-2 impact the generalization results we ablate the two most significant design decisions:

- 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.

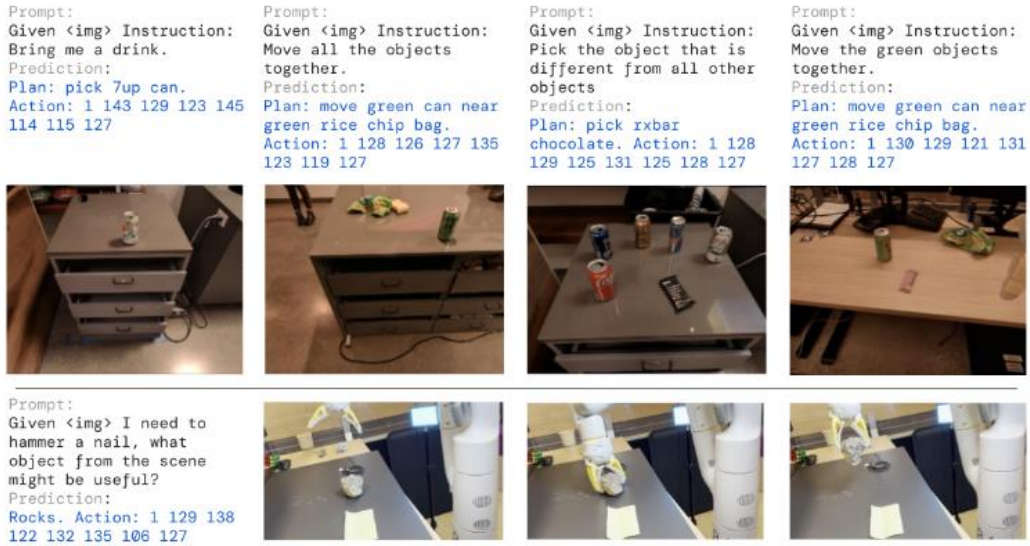
The results below indicate the importance of the pre-trained weights of the vision-langauge model as well as the trend of the model generalization improving with the model size.



We also evaluate RT-2 on an open-source language-table benchmark where we 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), we evaluate the resulting model in the real world. We demonstrate RT-2's generalization capabilities with the objects never seen in language table datasets before such as ketchup bottle, banana and others:



Lastly, 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, we demonstrate that RT-2 can perform chain-of-thought reasoning for control. In the examples below RT-2 first outputs a few reasoning steps in natural language which are then followed by the string: `Action:` and the resulting action tokens.



This shows the promise of fully integrated VLA models that can transfer not only some of the semantic concepts across different modalities (e.g. generalize robot actions to new semantic categories) but also some of the properties of the underlying models (e.g. chain-of-thought reasoning).

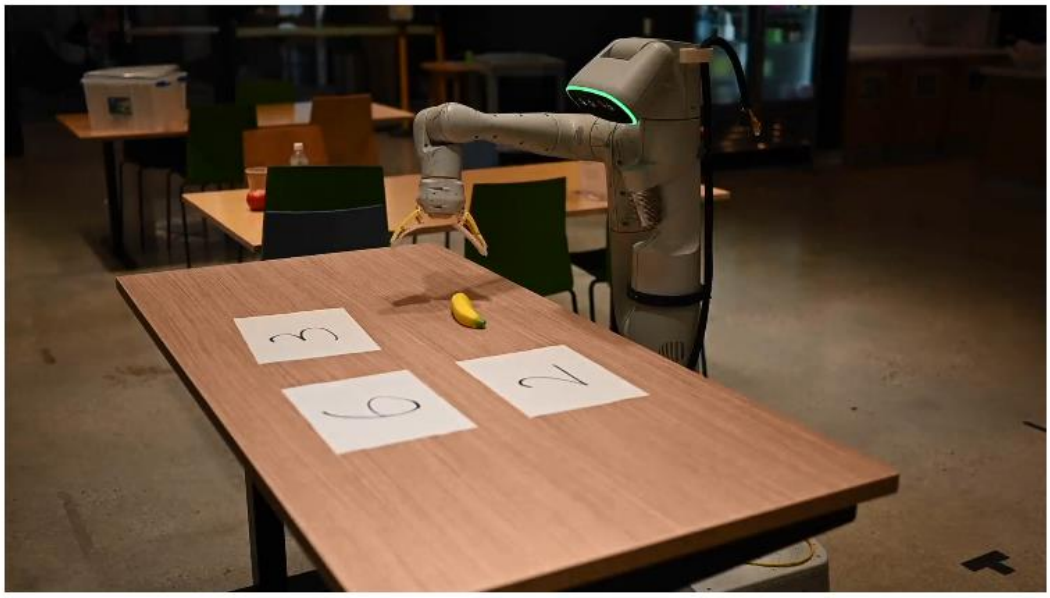
Demo

Prompt text in gray.

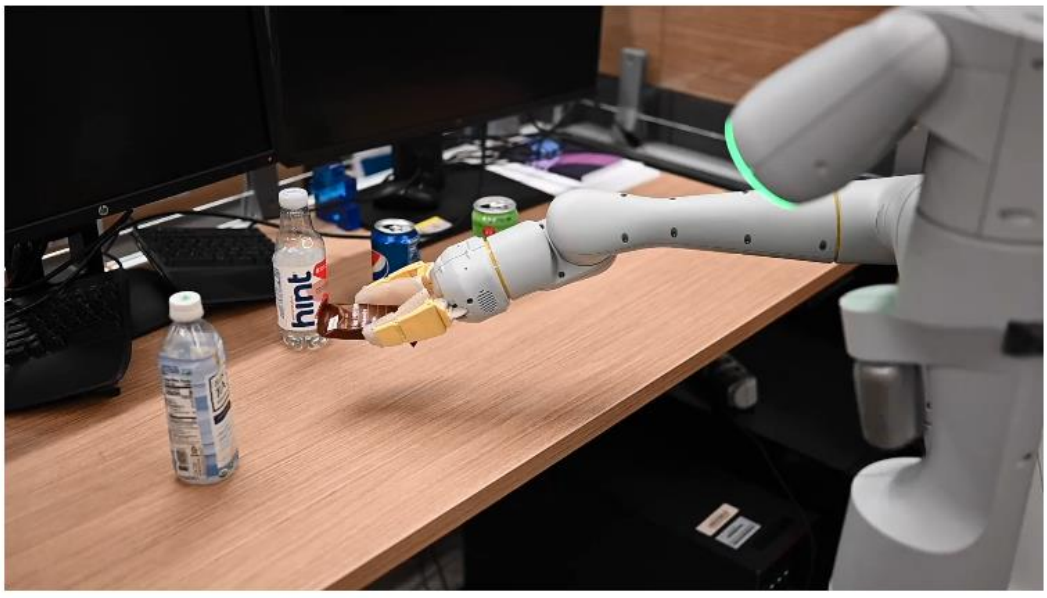
```
RT2 response (de-tokenized) shown within this block.
```

Videos

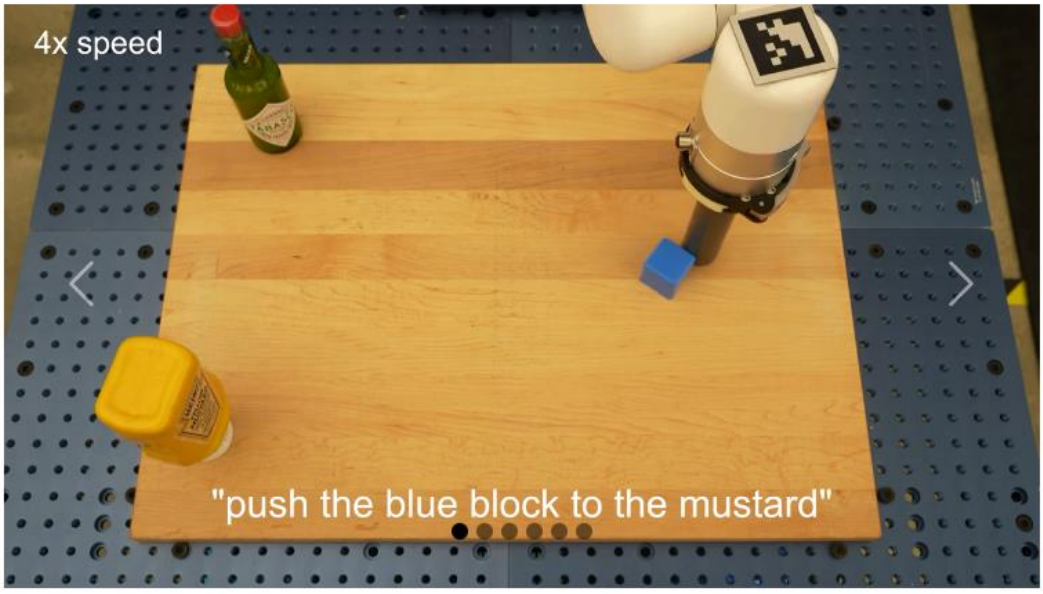
Below, we show a few videos showing examples of RT-2 execution. We show that RT-2 is able to generalize to new objects, new environments, and new tasks. RT-2 is able to generalize to a variety of real-world situations that require reasoning, symbol understanding, and human recognition.



RT-2 can exhibit signs of chain-of-thought reasoning similarly to vision-language models. We qualitatively observe that RT-2 with chain-of-thought reasoning is able to answer more sophisticated commands due to the fact that it is given a place to plan its actions in natural language first. This is a promising direction that provides some initial evidence that using LLMs or VLMs as planners can be combined with low-level policies in a single VLA model.



Finally, we show that RT-2 can work on another embodiment, Language Table environment. We show that RT-2 can handle real-world out-of-distribution behaviors in the Language Table environment.



Citation

```
@inproceedings{rt22023arxiv,
    title={RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic
Control},
    author={Anthony Brohan and Noah Brown and Justice Carbajal and Yevgen Chebotar and
Xi Chen and Krzysztof Choromanski and Tianli Ding and Danny Driess and Avinava Dubey
and Chelsea Finn and Pete Florence and Chuyuan Fu and Montse Gonzalez Arenas and
Keerthana Gopalakrishnan and Kehang Han and Karol Hausman and Alex Herzog and Jasmine
Hsu and Brian Ichter and Alex Irpan and Nikhil Joshi and Ryan Julian and Dmitry
Kalashnikov and Yuheng Kuang and Isabel Leal  and Lisa Lee and Tsang-Wei Edward Lee and
Sergey Levine and Yao Lu and Henryk Michalewski and Igor Mordatch and Karl Pertsch and
Kanishka Rao and Krista Reymann and Michael Ryoo and Grecia Salazar and Pannag Sanketi
and Pierre Sermanet and Jaspiar Singh and Anikait Singh and Radu Soricut and Huong Tran
and Vincent Vanhoucke and Quan Vuong and Ayzaan Wahid and Stefan Welker and Paul
Wohlhart and  Jialin Wu and Fei Xia and Ted Xiao and Peng Xu and Sichun Xu and Tianhe
Yu and Brianna Zitkovich},
    booktitle={arXiv preprint arXiv:2307.15818},
    year={2023}
}
```

Acknowledgements

We would like to thank John Guilyard for the amazing animations used for this website and beyond. The authors would like to acknowledge Fred Alcober, Jodi Lynn Andres, Carolina Parada, Joseph Dabis, Rochelle Dela Cruz, Jessica Gomez, Gavin Gonzalez, Tomas Jackson, Jie Tan, Scott Lehrer, Dee M, Utsav Malla, Sarah Nguyen, Jane Park, Emily Perez, Elio Prado, Jornell Quiambao, Clayton Tan, Jodexty Therlonge, Eleanor Tomlinson, Wenxuan Zhou, Boyuan Chen, and the greater Google DeepMind team for their feedback and contributions.

The website template was borrowed from Jon Barron.

--

July 28, 2023 Research

RT-2: New model translates vision and language into action

Yevgen Chebotar, Tianhe Yu

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 generalised instructions for robotic control

High-capacity vision-language models (VLMs) are trained on web-scale datasets, making these systems remarkably good at recognising visual or language patterns and operating across different languages. But for robots to achieve a similar level of competency, they would need to collect robot data, first-hand, across every object, environment, task, and situation.

In our paper, we introduce Robotic Transformer 2 (RT-2), a novel vision-language-action (VLA) model that learns from both web and robotics data, and translates this knowledge into generalised instructions for robotic control, while retaining web-scale capabilities.

A visual-language model (VLM) pre-trained on web-scale data is learning from RT-1 robotics data to become RT-2, a visual-language-action (VLA) model that can control a robot.

This work 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, our work used RT-1 robot demonstration data that was collected with 13 robots over 17 months in an office kitchen environment.

RT-2 shows improved generalisation capabilities and semantic and visual understanding beyond the robotic data it was exposed to. This includes interpreting new commands and responding to user commands by performing rudimentary reasoning, such as reasoning about object categories or high-level descriptions.

We also show that incorporating chain-of-thought reasoning allows RT-2 to perform multi-stage semantic reasoning, like deciding which object could be used as an improvised hammer (a rock), or which type of drink is best for a tired person (an energy drink).

Adapting VLMs for robotic control

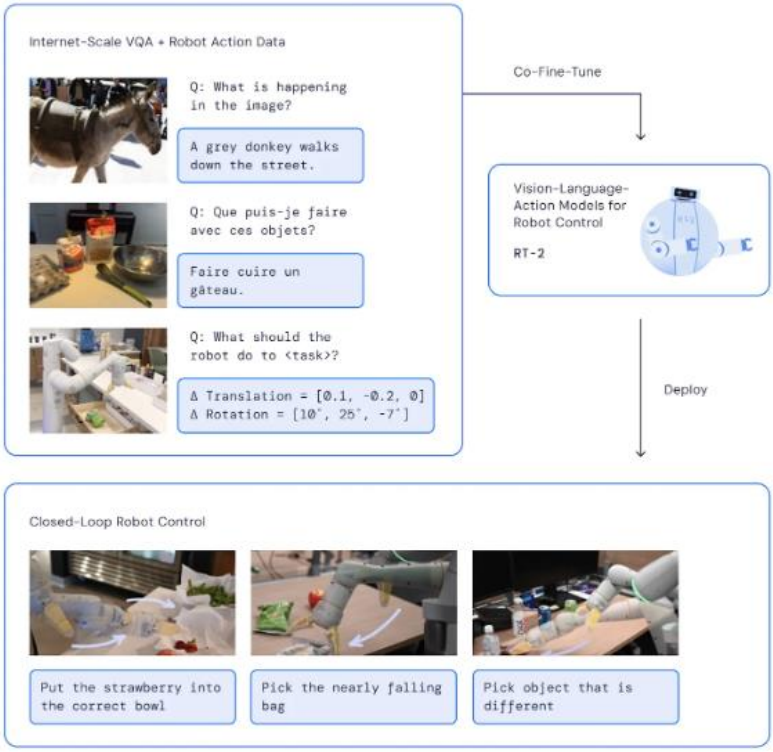
RT-2 builds upon VLMs that take one or more images as input, and produces a sequence of tokens that, conventionally, represent natural language text. Such VLMs have been successfully trained on web-scale data to perform tasks, like visual question answering, image captioning, or object recognition. In our work, we adapt Pathways Language and Image model (PaLI-X) and Pathways Language model Embodied (PaLM-E) to act as the backbones of RT-2.

To control a robot, it must be trained to output actions. We address this challenge by representing actions as tokens in the model's output – similar to language tokens – and describe actions as strings that can be processed by standard natural language tokenizers, shown here:

Representation of an action string used in RT-2 training. An example of such a string could be a sequence of robot action token numbers, e.g.“1 128 91 241 5 101 127 217”.

The string starts with a flag that indicates whether to continue or terminate the current episode, without executing the subsequent commands, and follows with the commands to change position and rotation of the end-effector, as well as the desired extension of the robot gripper.

We use the same discretised version of robot actions as in RT-1, and show that converting it to a string representation makes it possible to train VLM models on robotic data – as the input and output spaces of such models don't need to be changed.

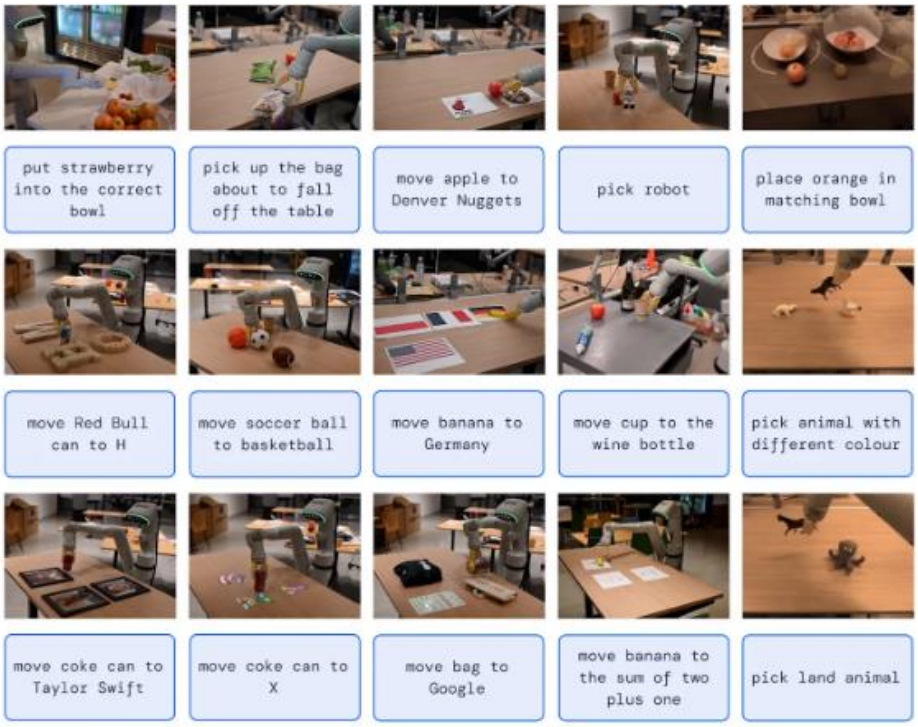


RT-2 architecture and training: We co-fine-tune a pre-trained VLM model on robotics and web data. The resulting model takes in robot camera images and directly predicts actions for a robot to perform.

Generalisation and emergent skills

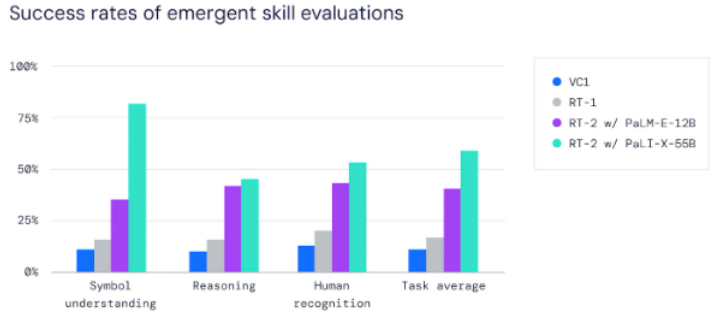
We performed a series of qualitative and quantitative experiments on our RT-2 models, on over 6,000 robotic trials. Exploring RT-2's emergent capabilities, we first searched for tasks that would require combining knowledge from web-scale data and the robot's experience, and then defined three categories of skills: symbol understanding, reasoning, and human recognition.

Each task required understanding visual-semantic concepts and the ability to perform robotic control to operate on these concepts. Commands such as "pick up the bag about to fall off the table" or "move banana to the sum of two plus one" – where the robot is asked to perform a manipulation task on objects or scenarios never seen in the robotic data – required knowledge translated from web-based data to operate.



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

Across all categories, we observed increased generalisation performance (more than 3x improvement) compared to previous baselines, such as previous RT-1 models and models like Visual Cortex (VC-1), which were pre-trained on large visual datasets.



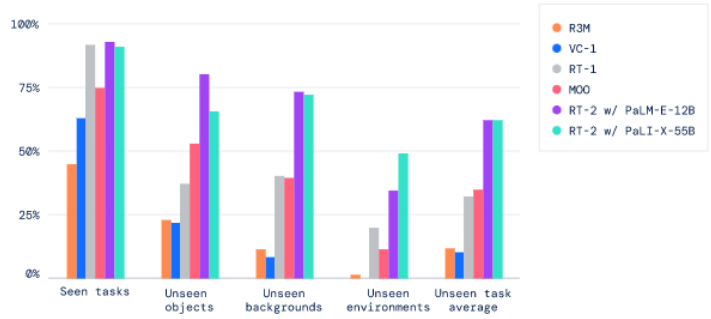
Success rates of emergent skill evaluations: our RT-2 models outperform both previous robotics transformer (RT-1) and visual pre-training (VC-1) baselines.

We also performed a series of quantitative evaluations, beginning with the original RT-1 tasks, for which we have examples in the robot data, and continued with varying degrees of previously unseen objects, backgrounds, and environments by the robot that required the robot to learn generalisation from VLM pre-training.

Examples of previously unseen environments by the robot, where RT-2 generalises to novel situations.

RT-2 retained the performance on the original tasks seen in robot data and improved performance on previously unseen scenarios by the robot, from RT-1's 32% to 62%, showing the considerable benefit of the large-scale pre-training.

Additionally, we observed significant improvements over baselines pre-trained on visual-only tasks, such as VC-1 and Reusable Representations for Robotic Manipulation (R3M), and algorithms that use VLMs for object identification, such as Manipulation of Open-World Objects (MOO).

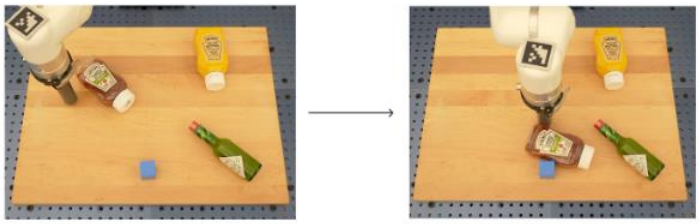


RT-2 achieves high performance on seen in-distribution tasks and outperforms multiple baselines on out-of-distribution unseen tasks.

Evaluating our model on the open-source Language Table suite of robotic tasks, we achieved a success rate of 90% in simulation, substantially improving over the previous baselines including BC-Z (72%), RT-1 (74%), and LAVA (77%).

Then we evaluated the same model in the real world (since it was trained on simulation and real data), and demonstrated its ability to generalise to novel objects, as shown below, where none of the objects except the blue cube were present in the training dataset.

Push the *ketchup to the blue cube*



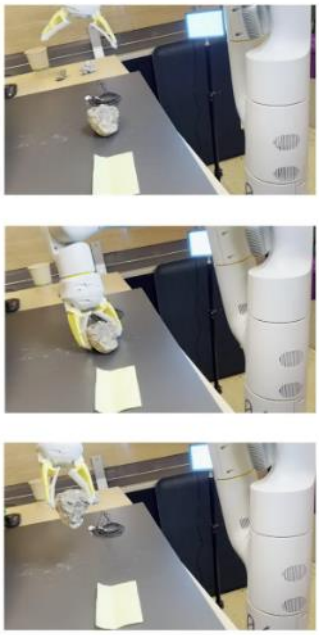
RT-2 performs well on real robot Language Table tasks. None of the objects except the blue cube were present in the training data.

Inspired by chain-of-thought prompting methods used in LLMs, we probed our models to combine robotic control with chain-of-thought reasoning to enable learning long-horizon planning and low-level skills within a single model.

In particular, we fine-tuned a variant of RT-2 for just a few hundred gradient steps to increase its ability to use language and actions jointly. Then we augmented the data to include an additional "Plan" step, first describing the purpose of the action that the robot is about to take in natural language, followed by "Action" and the action tokens. Here we show an example of such reasoning and the robot's resulting behaviour:

Instruction:
I need to hammer a nail, what object from the scene might be useful?

Prediction:
Rocks. Action: 1 129 138 122 132 132 106 127



Chain-of-thought reasoning enables learning a self-contained model that can both plan long-horizon skill sequences and predict robot actions.

With this process, RT-2 can perform more involved commands that require reasoning about intermediate steps needed to accomplish a user instruction. Thanks to its VLM backbone, RT-2 can also plan from both image and text commands, enabling visually grounded planning, whereas current plan-and-act approaches like SayCan cannot see the real world and rely entirely on language.

Advancing robotic control

RT-2 shows that vision-language models (VLMs) can be transformed into powerful vision-language-action (VLA) models, which can directly control a robot by combining VLM pre-training with robotic data.

With two instantiations of VLAs based on PaLM-E and PaLI-X, RT-2 results in highly-improved robotic policies, and, more importantly, leads to significantly better generalisation performance and emergent capabilities, inherited from web-scale vision-language pre-training.

RT-2 is not only a simple and effective modification over existing VLM models, but also shows the promise of building a general-purpose physical robot that can reason, problem solve, and interpret information for performing a diverse range of tasks in the real-world.