RT2 (Robotics Transformer 2) from DeepMind

AI Bites
14.3K subscribers

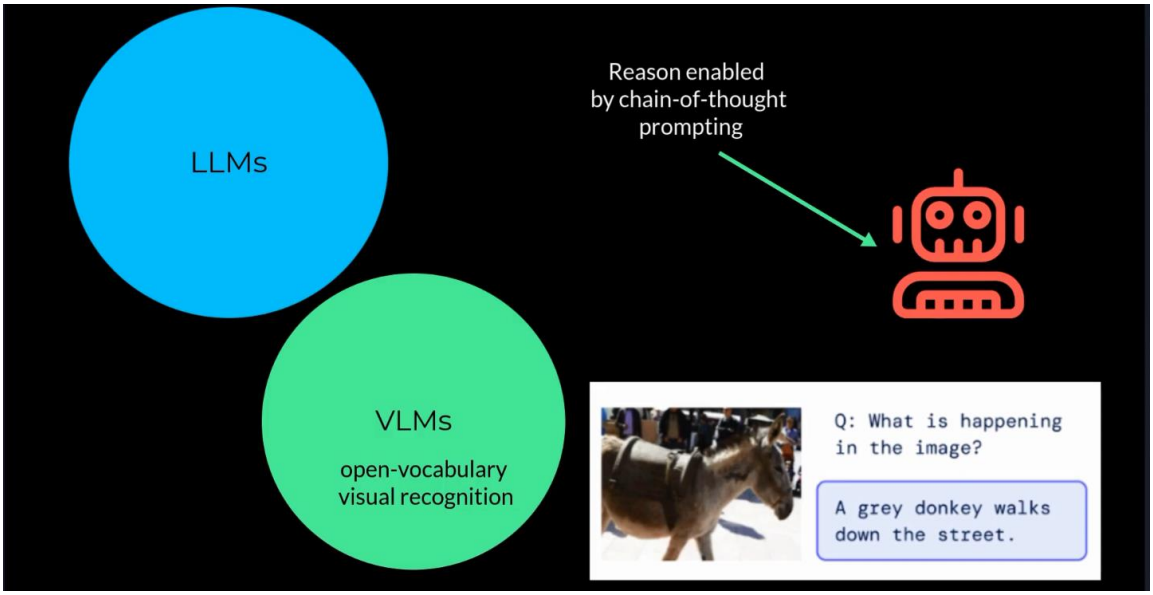
Join | Subscribe | 94 | Share | Ask | Save

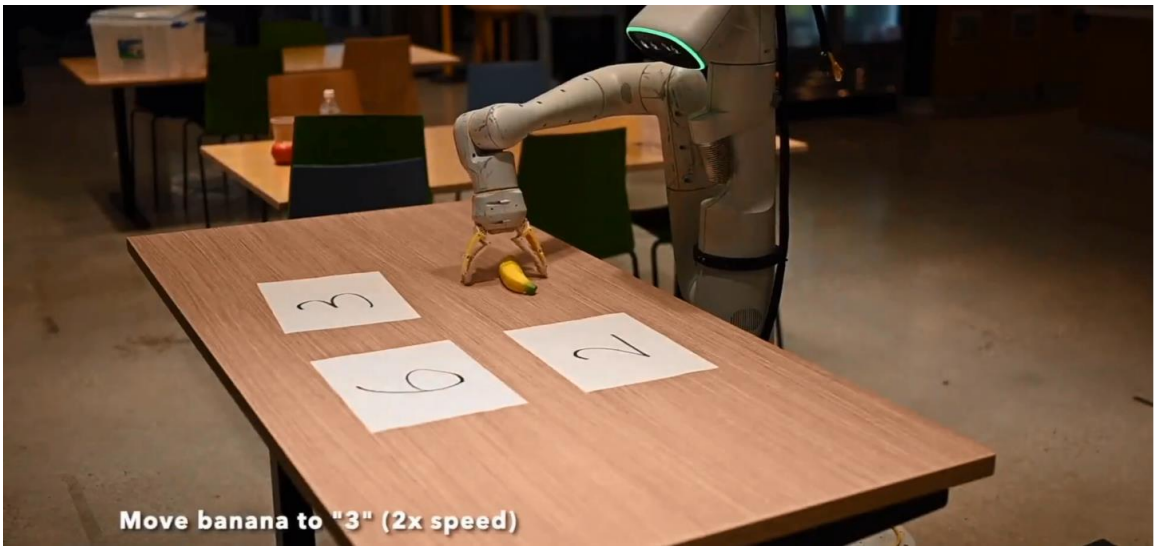
4,311 views Aug 10, 2023 #deeplearning #machinelearning #aibites

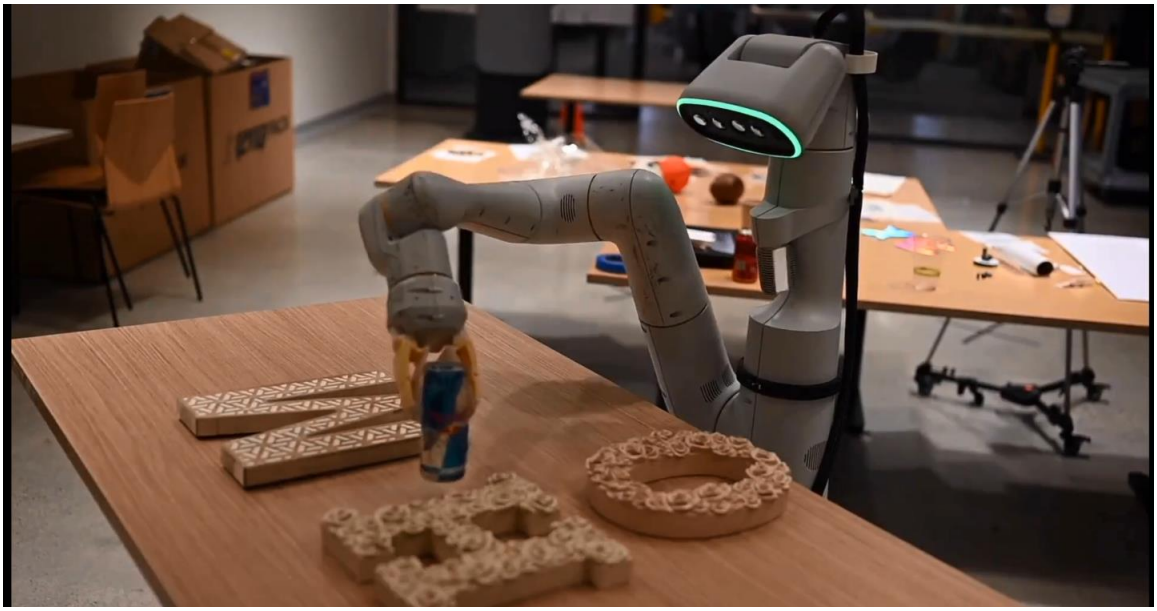
The biggest positive of LLMs today is that they have a good ability to reason. If robots can be equiped with this very ability to reason, then we will be one stop closer to intelligent robots.

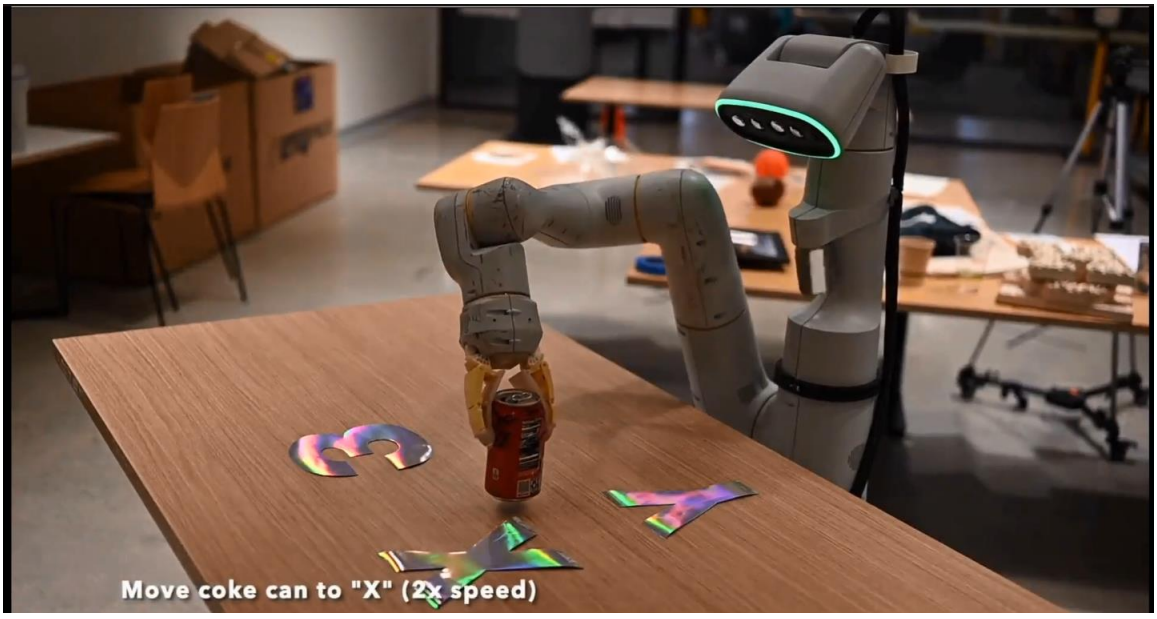
This video is a deep dive into the Robotics Transformer 2 (RT-2) paper and the model.

RT2 Blog: https://www.deepmind.com/blog/rt-2-ne...
RT2 Demo: https://robotics-transformer2.github....
RT2 Paper Download: https://robotics-transformer2.github....
PALM Model: https://ai.googleblog.com/2022/04/pat...
PALM-E Model: https://ai.googleblog.com/2023/03/pal...
PALI Model: https://ai.googleblog.com/2022/09/pal...
PALI-X Model: https://arxiv.org/pdf/2305.18565.pdf




Move banana to "3" (2x speed)




Move coke can to "X" (2x speed)

RT-2: Vision-Language-Action Models

Transfer Web Knowledge to Robotic Control

Anthony Brohan Noah Brown Justice Carbajal Yevgen Chebotar Xi Chen Krzysztof Choromanski Tianli Ding
Danny Driess Avinava Dubey Chelsea Finn Pete Florence Chuyuan Fu Montse Gonzalez Arenas Keerthana Gopalakrishnan
Kehang Han Karol Hausman Alex Herzog Jasmine Hsu Brian Ichter Alex Irpan Nikhil Joshi Ryan Julian
Dmitry Kalashnikov Yuheng Kuang Isabel Leal Lisa Lee Tsang-Wei Edward Lee Sergey Levine Yao Lu Henryk Michalewski
Igor Mordatch Karl Pertsch Kanishka Rao Krista Reymann Michael Ryoo Grecia Salazar Pannag Sanketi Pierre Sermanet
Jaspiar Singh Anikait Singh Radu Soricut Huong Tran Vincent Vanhoucke Quan Vuong Ayzaan Wahid Stefan Welker
Paul Wohlhart Jialin Wu Fei Xia Ted Xiao Peng Xu Sichun Xu Tianhe Yu Brianna Zitkovich

Authors listed in alphabetical order (see paper appendix for contribution statement).

Google DeepMind

Pre-Trained Vision-Language Models

Vision-Language Model

A: Move the object to the table

Q: What should the robot do?

Vision-Language Action Model

PaLI-X

PaLM-E.

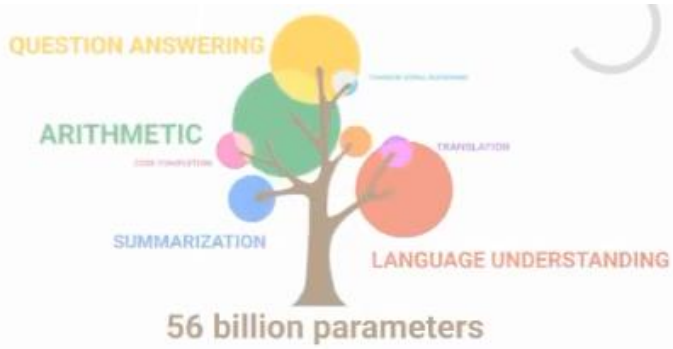
from more diverse sources. Yet much work remains in understanding the capabilities that emerge with few-shot learning as we push the limits of model scale.

Last year Google Research announced our vision for Pathways, a single model that could generalize across domains and tasks while being highly efficient. An important milestone toward realizing this vision was to develop the new Pathways system to orchestrate distributed computation for accelerators. In "PaLM: Scaling Language Modeling with Pathways", we introduce the Pathways Language Model (PaLM), a 540-billion parameter, dense decoder-only Transformer model trained with the Pathways system, which enabled us to efficiently train a single model across multiple TPU v4 Pods. We evaluated PaLM on hundreds of language understanding and generation tasks, and found that it achieves state-of-the-art few-shot performance across most tasks, by significant margins in many cases.

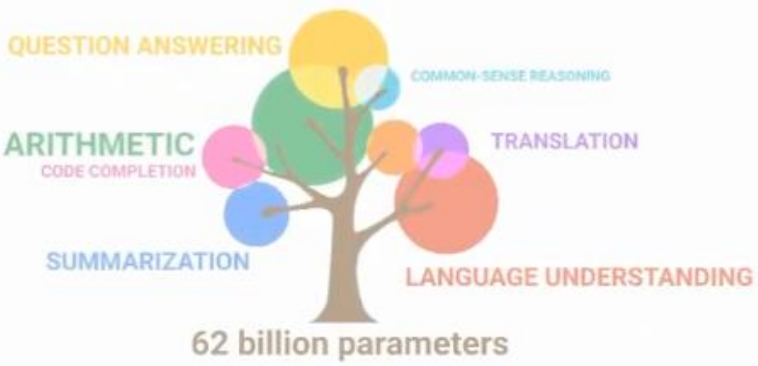
As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.

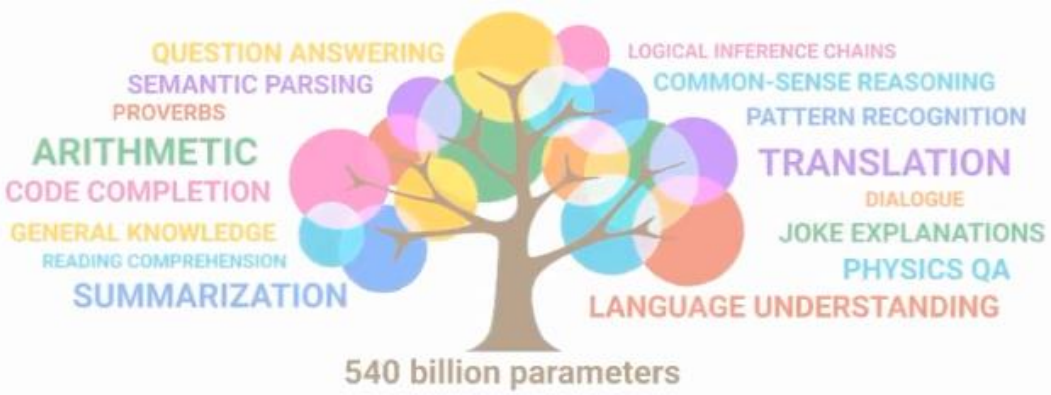
Training a 540-Billion Parameter Language Model with Pathways

PaLM demonstrates the first large-scale use of the Pathways system to scale training to 6144 chips, the largest TPU-



As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.





As the scale of the model increases, the performance improves across tasks while also unlocking new capabilities.

BLOG ›

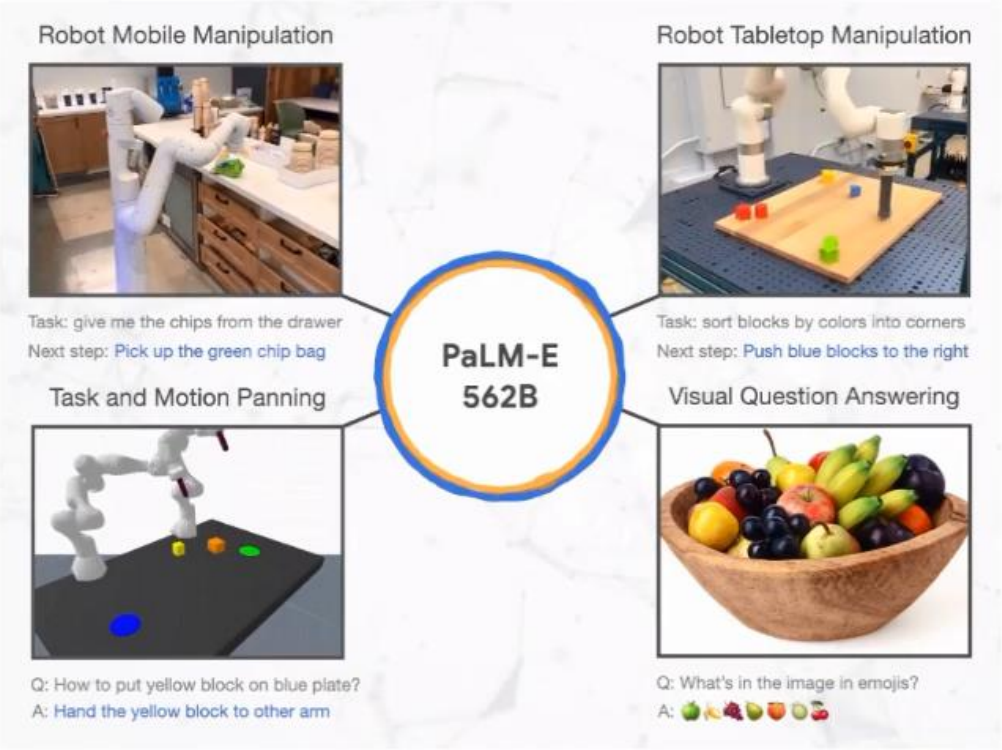
PaLM-E: An embodied multimodal language model

FRIDAY, MARCH 10, 2023

Posted by Danny Driess, Student Researcher, and Pete Florence, Research Scientist, Robotics at Google

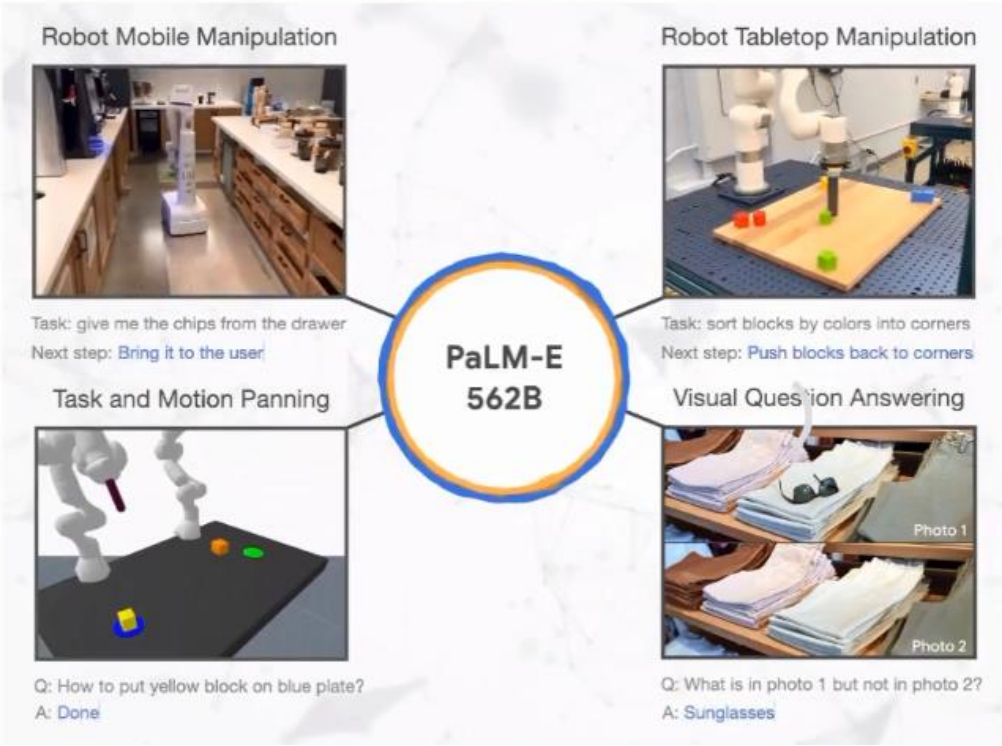
Recent years have seen tremendous advances across machine learning domains, from models that can explain jokes or answer visual questions in a variety of languages to those that can produce images based on text descriptions. Such innovations have been possible due to the increase in availability of large scale datasets along with novel advances that enable the training of models on these data. While scaling of robotics models has seen some success, it is outpaced by other domains due to a lack of datasets available on a scale comparable to large text corpora or image datasets.

Today we introduce PaLM-E, a new generalist robotics model that overcomes these issues by transferring knowledge from varied visual and language domains to a robotics system. We began with PaLM, a powerful large language model, and "embodied" it (the "*E*" in PaLM-E), by complementing it with sensor data from the robotic agent. This is the key difference from prior efforts to bring large language models to robotics — rather than relying on only textual input, with PaLM-E we train the language model to directly ingest raw streams of robot sensor data. The resulting model not only enables highly effective robot learning, but is also a state-of-the-art general-purpose visual-language model, while maintaining excellent language-only task capabilities.



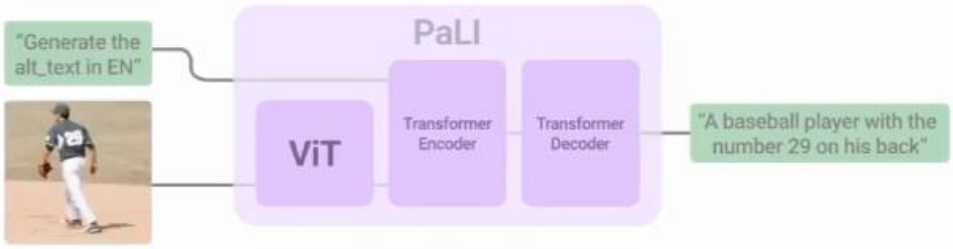
An *embodied* language model, and also a visual-language generalist

On the one hand, PaLM-E was primarily developed to be a model for robotics, and it solves a variety of tasks on *multiple* types of robots and for *multiple* modalities (images, robot states, and neural scene representations). At the



One goal of this project is to examine how language and vision models interact at scale and specifically the scalability of language-image models. We explore both per-modality scaling and the resulting cross-modal interactions of scaling. We train our largest model to 17 billion (17B) parameters, where the visual component is scaled up to 4B parameters and the language model to 13B.

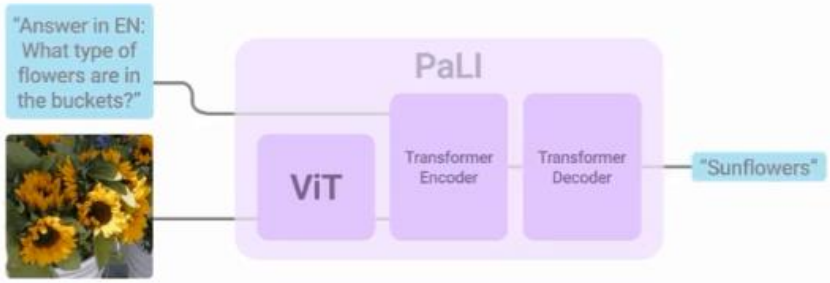
The PaLI model architecture is simple, reusable and scalable. It consists of a **Transformer** encoder that processes the input text, and an auto-regressive Transformer decoder that generates the output text. To process images, the input to the Transformer encoder also includes "visual words" that represent an image processed by a **Vision Transformer** (ViT). A key component of the PaLI model is reuse, in which we seed the model with weights from previously-trained uni-modal vision and language models, such as **mT5-XXL** and large **ViTs**. This reuse not only enables the transfer of capabilities from uni-modal training, but also saves computational cost.



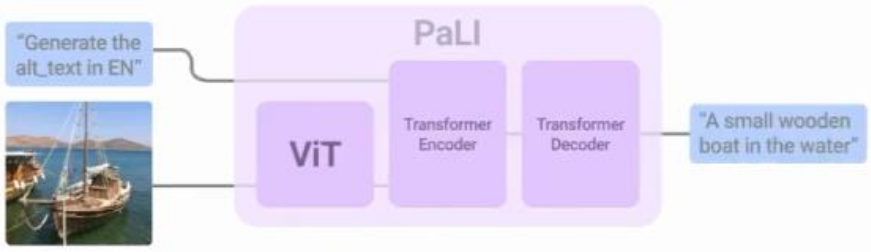
The PaLI model addresses a wide range of tasks in the language-image, language-only and image-only domain using the same API (e.g., visual-question answering, image captioning, scene-text understanding, etc.). The model is trained to support over 100 languages and tuned to perform multilingually for multiple language-image tasks.

Dataset: Language-Image Understanding in 100+ Languages

Scaling studies for deep learning show that larger models require larger datasets to train effectively. To unlock the potential of language-image pretraining, we construct WebLI, a multilingual language-image dataset built from images and text available on the public web.



The PaLI model addresses a wide range of tasks in the language-image, language-only and image-only domain using the same API (e.g., visual-question answering, image captioning, scene-text understanding, etc.). The model is trained to support over 100 languages and tuned to perform multilingually for multiple language-image tasks.



The PaLI model addresses a wide range of tasks in the language-image, language-only and image-only domain using the same API (e.g., visual-question answering, image captioning, scene-text understanding, etc.). The model is trained to support over 100 languages and tuned to perform multilingually for multiple language-image tasks.

PaLI-X: On Scaling up a Multilingual Vision and Language Model

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, Radu Soricut

Google Research

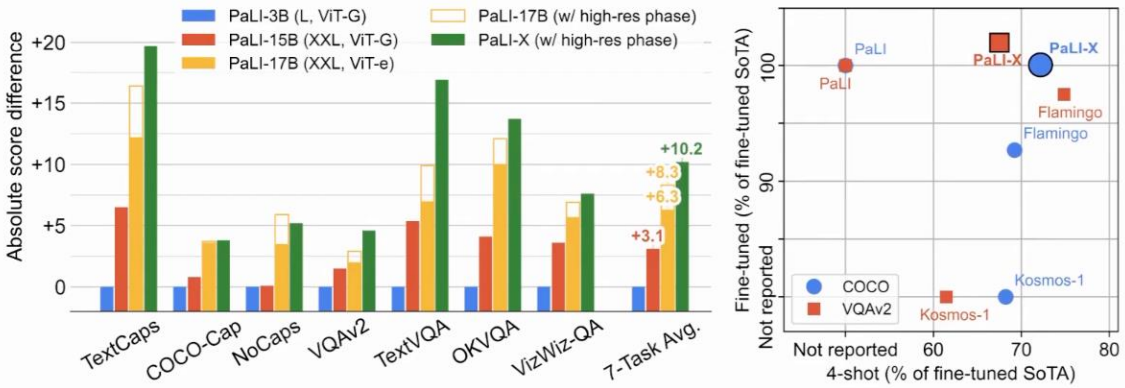
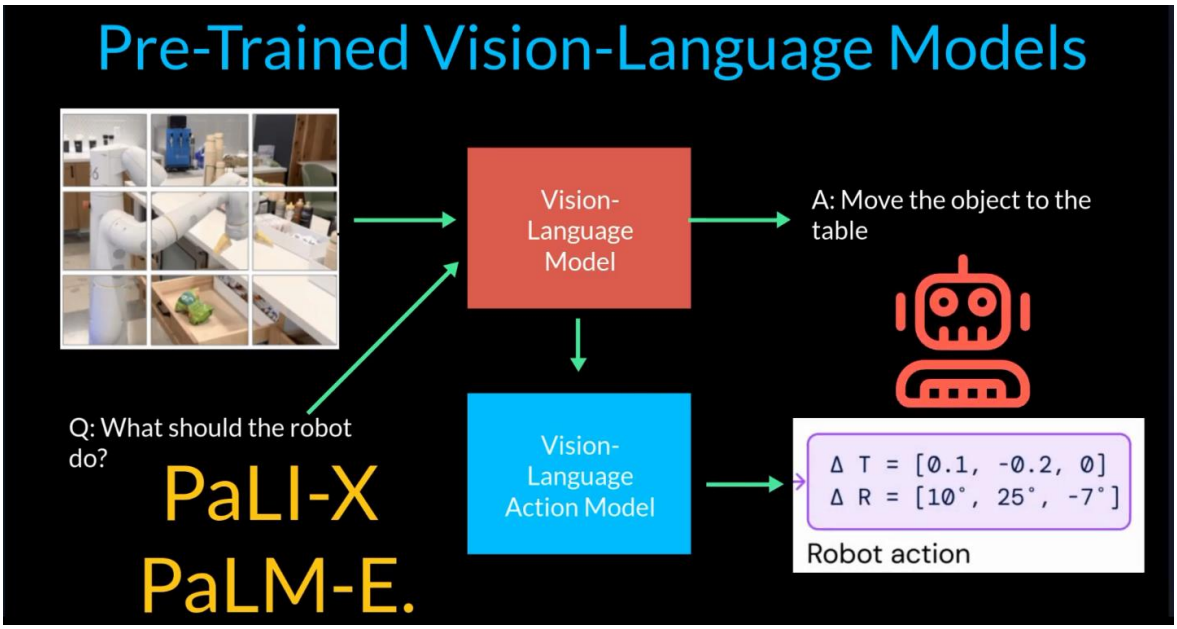
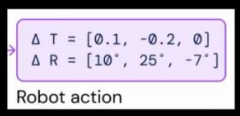


Figure 1: [Left] Comparing PaLI-X against PaLI on image-captioning and VQA benchmarks. [Right] The Pareto frontier between few-shot and fine-tuned performance, comparing PaLI-X with PaLI [5], Flamingo [10], and Kosmos-1 [11].



Robot-Action Fine-tuning



Tokenizer

"1 91 2 128 101 227"

Reserve 256 tokens from tokenizer for action

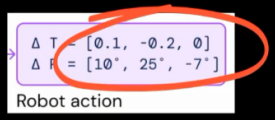
Discretize continuous values of action (numbers) into one of the 256 bins.

For example, for Translation x ranging from 0.0 to 0.2 will be in bin 1.

0.0
0.01
0.02
0.03
...
0.1
} bin 1

0.10
0.11
0.12
0.13
...
0.2
} bin 2

Robot-Action Fine-tuning



Tokenizer

"1 91 2 128 101 227"

Discretize continuous values of action (numbers) into one of the 256 bins.

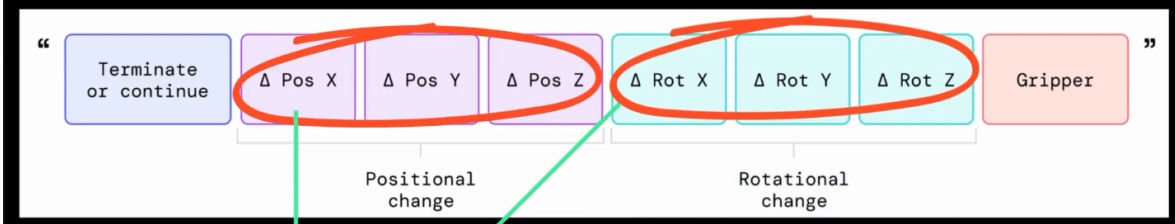
Reserve 256 tokens from tokenizer for action

For example, lets take Translation x ranging from 0.0 to 0.2

0.0
0.01
0.02
0.03
...
0.1
} bin 1

0.10
0.11
0.12
0.13
...
0.2
} bin 2

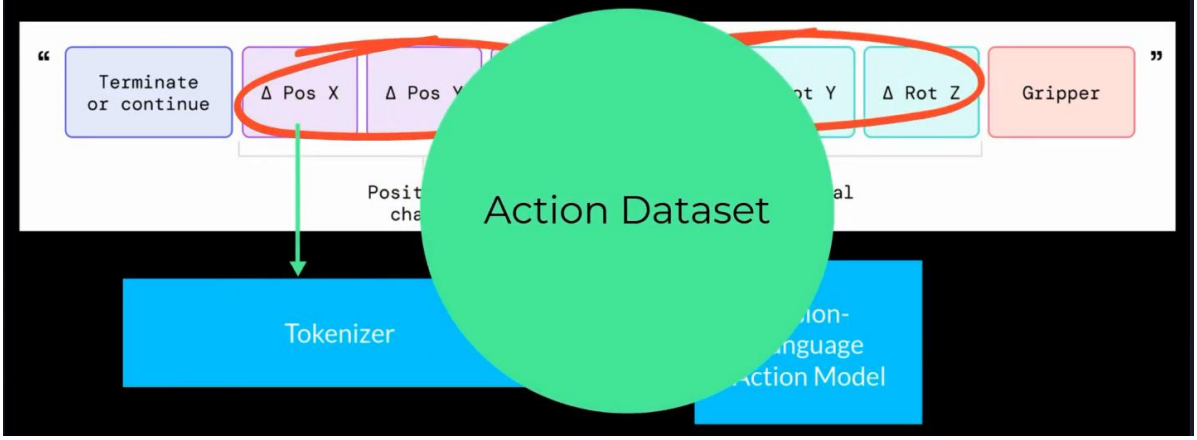
Robot-Action Fine-tuning



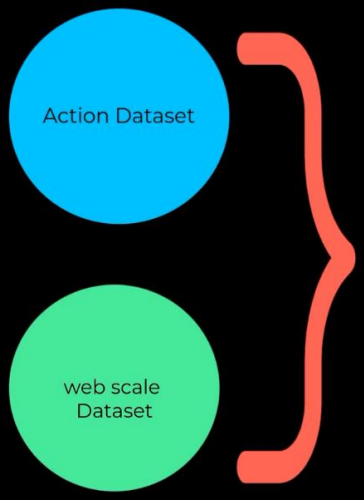
Tokenizer

Vision-Language Action Model

Robot-Action Fine-tuning



Co-fine-tuning

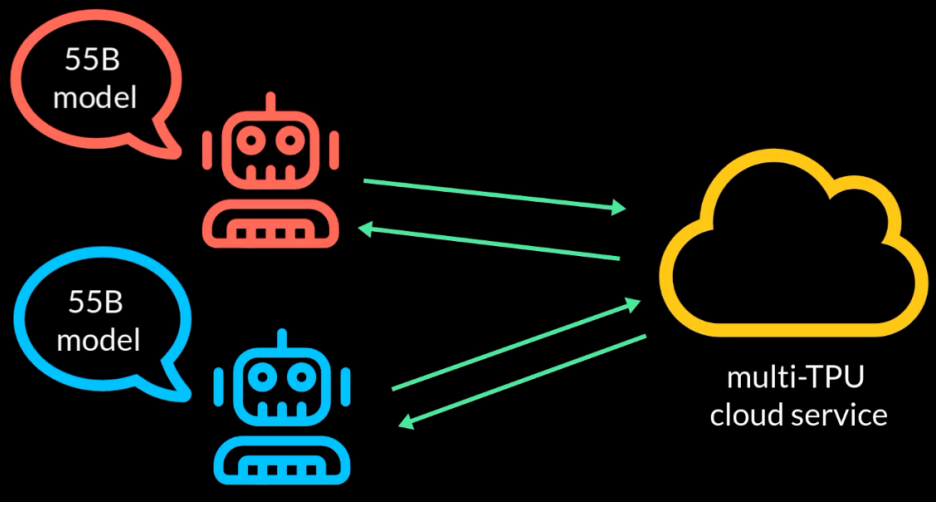


more generalizable policies co mpared to training just with ac tion dataset

For the input, sample more of a ction data to balance

Output only action tokens as w e only need actions from the V LA model

Real-time Inference



We evaluate our approach and several baselines with about 6,000 evaluation trajectories in a variety of conditions, which we describe in the following sections. Unless specified otherwise, we use a 7DoF mobile manipulator with the action space described in Sec. 3.2. We also demonstrate examples of RT-2 execution on the project website: robotics-transformer2.github.io. We 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), and (2) **RT-2-PaLM-E** is built from 12B PaLM-E (Driess et al., 2023).

For training, we 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. We combine it with the 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") (see Appendix B for more details on the used datasets). For all RT-2 training runs we 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. More training details can be found in Appendix E.

Baselines. We compare our method to multiple state-of-the-art baselines that challenge different aspects of our method. All of the baselines use the exact same robotic data. To compare against a state-of-the-art policy, we use **RT-1** (Brohan et al., 2022), a 35M parameter transformer-based model. To compare against state-of-the-art pretrained representations, we use **VC-1** (Majumdar et al., 2023a)

4.1. How does RT-2 perform on seen tasks and more importantly, generalize over new objects, backgrounds, and environments?

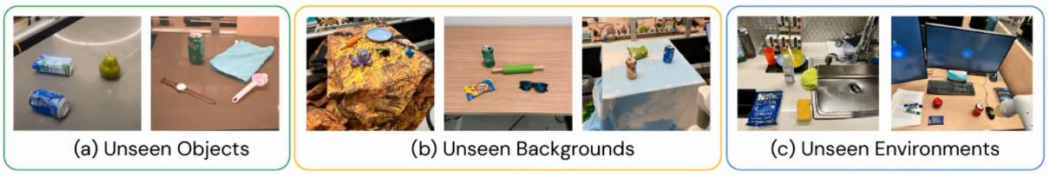


Figure 3 | Example generalization scenarios used for evaluation in Figures 4 and 6b and Tables 4 and 6.

To evaluate in-distribution performance as well as generalization capabilities, we compare the RT-2-PaLI-X and RT-2-PaLM-E models to the four baselines listed in the previous sections. For the *seen tasks* category, we use the same suite of seen instructions as in RT-1 (Brohan et al., 2022), which 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. Note, however, that these "in-distribution" evaluations still vary the placement of objects and factors such as time of day and robot position, requiring the skills to generalize to realistic variability in the environment.

primarily on pick and placing skills in many diverse scenarios. The list of instructions for unseen categories is specified in Appendix F.2.

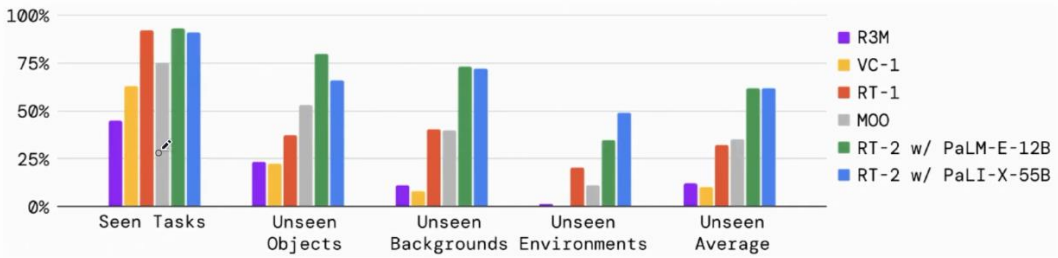
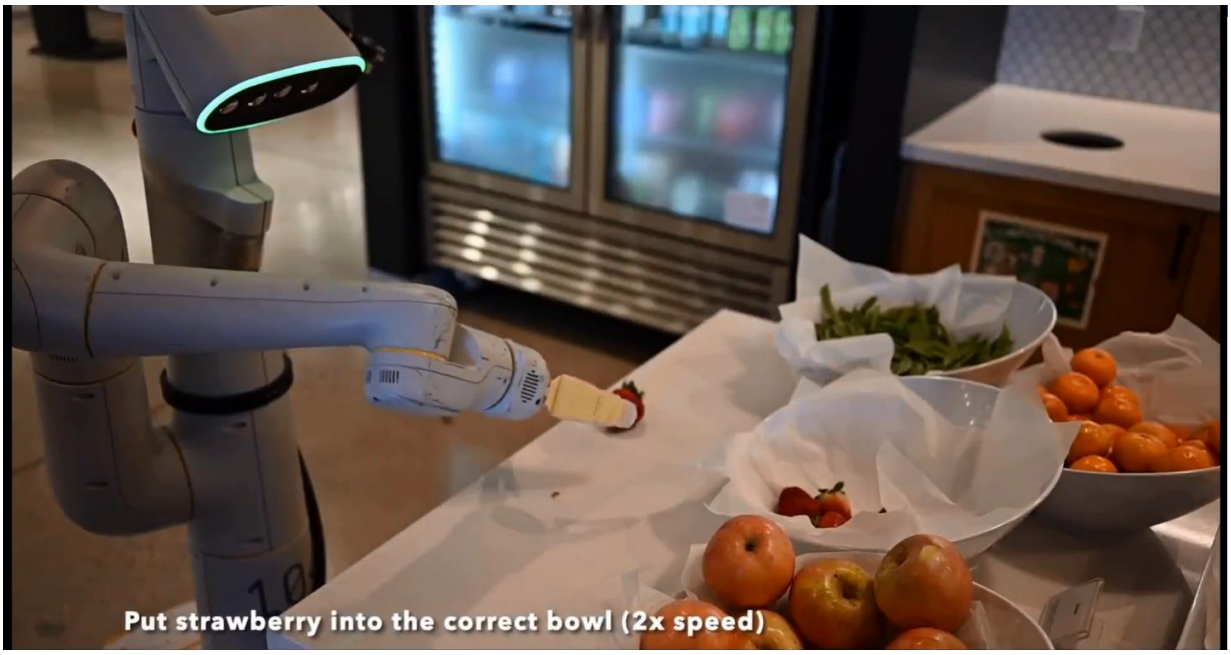


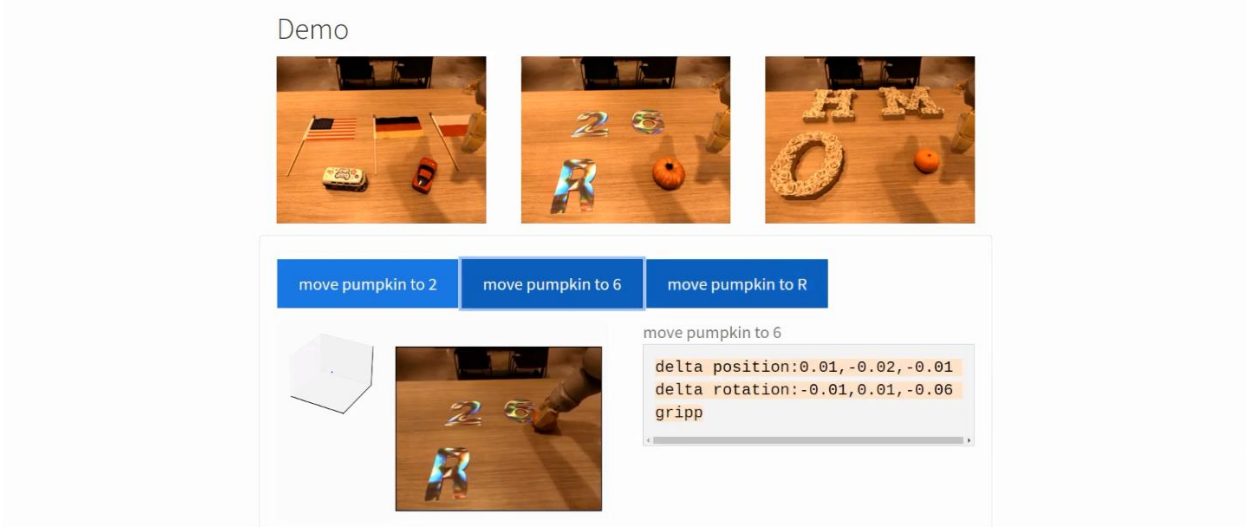
Figure 4 | Overall performance of two instantiations of RT-2 and baselines across seen training tasks as well as unseen evaluations measuring generalization to novel objects, novel backgrounds, and novel environments. Appendix Table 4 details the full results.

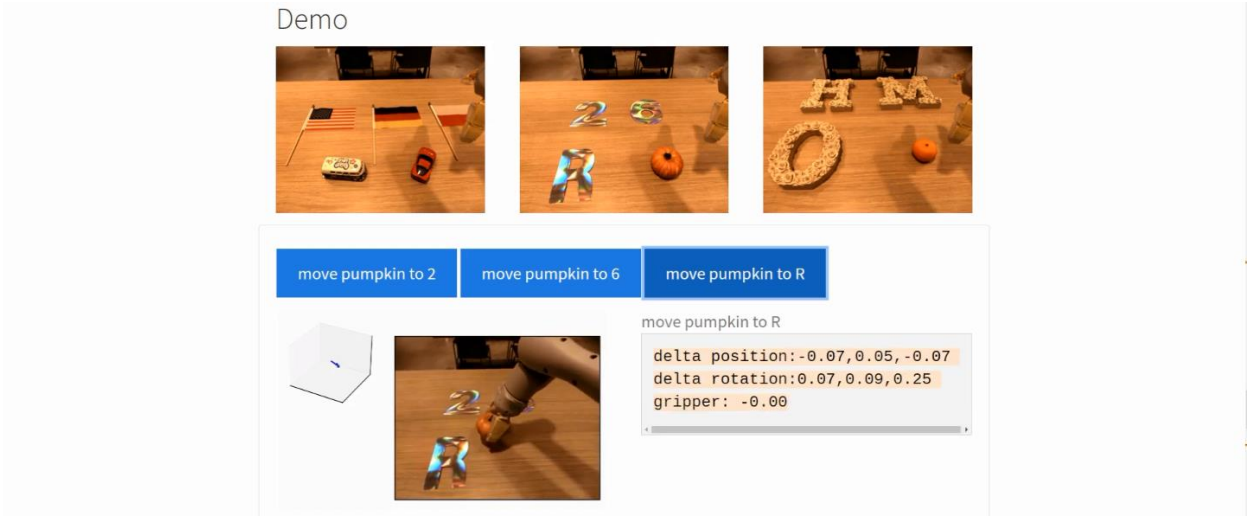
The evaluation results are shown in Figure 4 and Appendix Table 4. The performance on seen tasks is similar between the RT-2 models and RT-1, with other baselines attaining a lower success rate. The difference between the RT-2 models and the baseline is most pronounced in the various generalization experiments, suggesting that the strength of vision-language-action models lies in transferring more generalizable visual and semantic concepts from their Internet-scale pretraining

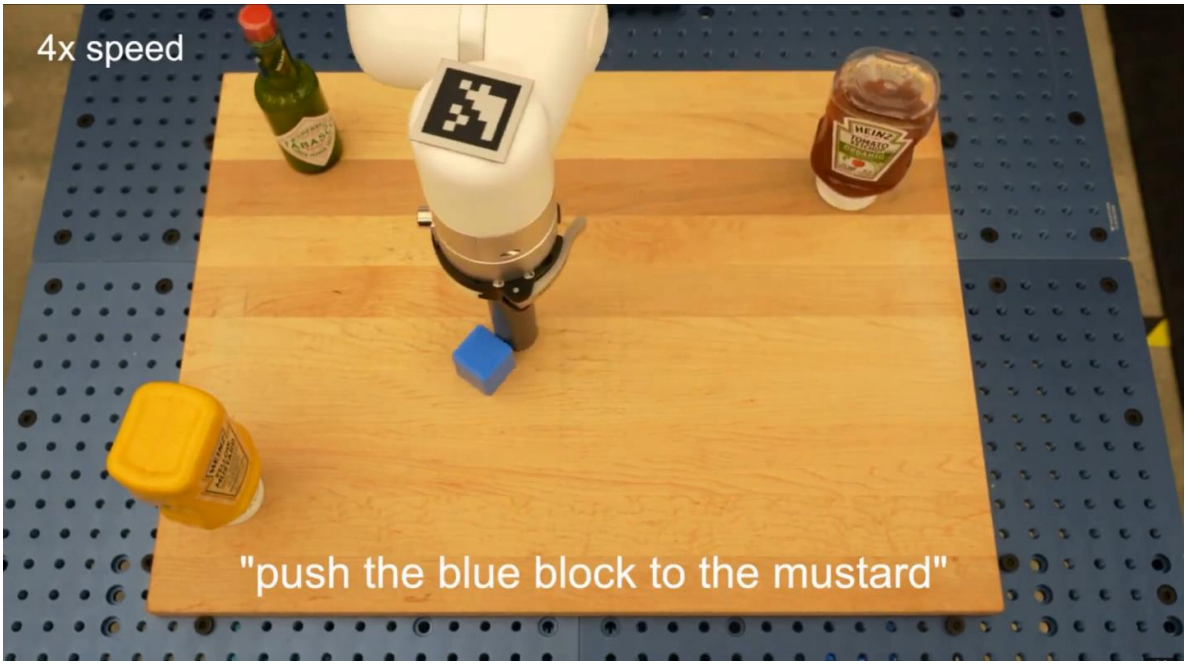
4.2. Can we observe and measure any emergent capabilities of RT-2?

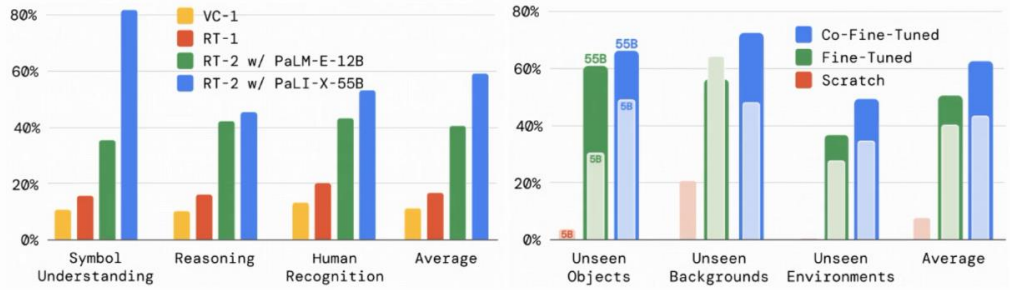
In addition to evaluating the generalization capabilities of vision-language-action models, we also aim to evaluate the degree to which such models can enable new capabilities beyond those demonstrated











(a) Performance comparison on various emergent skill evaluations (Figure 8) between RT-2 and two baselines.

(b) Ablations of RT-2-PaLI-X showcasing the impact of parameter count and training strategy on generalization.

Figure 6 | Quantitative performance of RT-2 across (6a) emergent skills and (6b) size and training ablations. Appendix Tables 5 and 6 detail the full numerical results.

4.3. How does the generalization vary with parameter count and other design decisions?

For this comparison, we use RT-2-PaLI-X model because of its flexibility in terms of the model size (due to the nature of PaLM-E, RT-2-PaLM-E is restricted to only certain sizes of PaLM and ViT models). In particular, we compare two different model sizes, 5B and 55B, as well as three different training the fact that keeping the original data around the fine-tuning part of training, allows the model to not forget its previous concepts learned during the VLM training. Lastly, somewhat unsurprisingly, we notice that the increased size of the model results in a better generalization performance.

4.4. Can RT-2 exhibit signs of chain-of-thought reasoning similarly to vision-language models?

Inspired by the chain-of-thought prompting method in LLMs (Wei et al., 2022), we fine-tune a variant of RT-2 with PaLM-E for just a few hundred gradient steps to increase its capability of utilizing language and actions jointly with the hope that it will elicit a more sophisticated reasoning behavior. We augment the data to include an additional "Plan" step, which describes the purpose of the action that the robot is about to take in natural language first, which is then followed by the actual action tokens, e.g. "Instruction: I'm hungry. Plan: pick rxbar chocolate. Action: 1 128 124 136 121 158 111 255." This data augmentation scheme acts as a bridge between VQA datasets (visual reasoning) and manipulation datasets (generating actions).

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 (Ahn et al., 2022; Driess et al., 2023) can be combined with low-level policies in a single VLA model. Rollouts of RT-2 with chain-of-thought reasoning are shown in Figure 7 and in Appendix I.

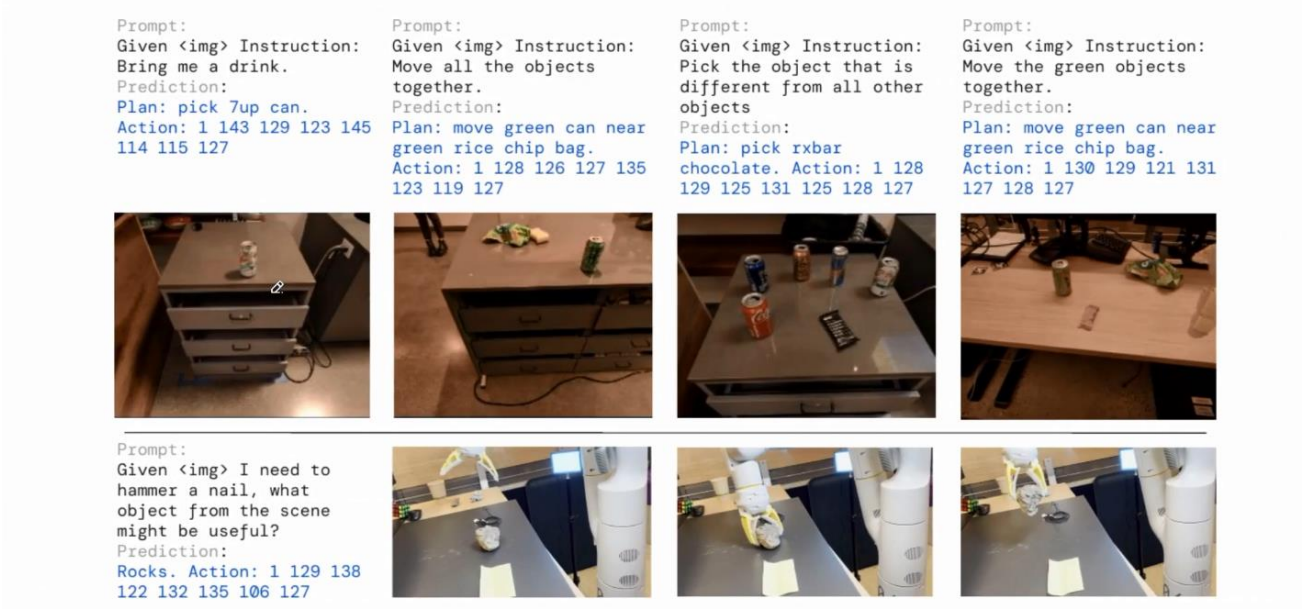


Figure 7 | Rollouts of RT-2 with chain-of-thought reasoning, where RT-2 generates both a plan and an action.

5. Limitations

Even though RT-2 exhibits promising generalization properties, there are multiple limitations of this approach. First, although we show that including web-scale pretraining 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 (see Appendix G), but it learns to deploy those skills in new ways. We believe this is a result of the dataset not being varied enough along the axes of skills. An exciting direction for future work is to study how new skills could be acquired through new data collection paradigms such as videos of humans.

Second, although we showed we 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. This is also connected to another current limitation in that there are only a small number of generally available VLM models that can be used to create RT-2. We hope that more open-sourced models will become available (e.g. https://llava-vl.github.io/) and the proprietary ones will open up their fine-tuning APIs, which is a sufficient requirement to build VLA models.