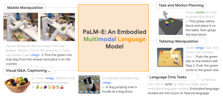
PaLM-E: An Embodied Multimodal Language Model

Danny Driess[1,2] Fei Xia[1] Mehdi S. M. Sajjadi[3] Corey Lynch[1]
Aakanksha Chowdhery[3]
Brian Ichter[1] Ayzaan Wahid[1] Jonathan Tompson[1] Quan Vuong[1]
Tianhe Yu[1] Wenlong Huang[1]
Yevgen Chebotar[1] Pierre Sermanet[1] Daniel Duckworth[3]
Sergey Levine[1] Vincent Vanhoucke[1]
Karol Hausman[1] Marc Toussaint[2] Klaus Greff[3] Andy Zeng[1]
Igor Mordatch[3] Pete Florence[1]

[1] Robotics at Google (http://g.co/robotics) [2] TECHNISCHE UNIVERSITÄT BERLIN (https://www.tu.berlin/en/) [3] Google Research (https://research.google/teams/brain/)

Paper

(assets/palm-e.pdf)

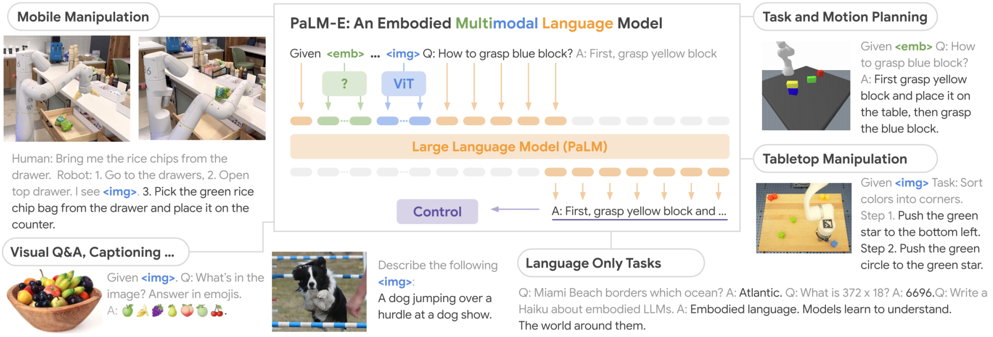


Demo

Abstract

Large language models have been demonstrated to perform complex tasks. However, enabling general inference in the real world, e.g. for robotics problems, raises the challenge of grounding. We propose embodied language models to directly incorporate real-world continuous sensor modalities into language models and thereby establish the link between words and percepts. Input to our embodied language model are multi-modal sentences that interleave visual, continuous state estimation, and textual input encodings. We train these encodings end-to-end, in conjunction with a pre-trained large language model, for multiple embodied tasks, including sequential robotic manipulation planning, visual question answering, and captioning. Our evaluations show that PaLM-E, a single large embodied multimodal model, can address a variety of embodied reasoning tasks, from a variety of observation modalities, on multiple embodiments, and further, exhibits *positive transfer*: the model benefits from diverse joint training across internet-scale language, vision, and visual-language domains. Our largest model, PaLM-E-562B with 562B parameters, in addition to being trained on robotics tasks, is a visual-language generalist with state-of-the-art performance on OK-VQA, and retains generalist language capabilities with increasing scale.

Approach


Mobile Manipulation
PaLM-E: An Embodied Multimodal Language Model
?
ViT
Large Language Model (PaLM)
Control
Task and Motion Planning
Tabletop Manipulation
Visual Q&A, Captioning ...
Language Only Tasks

Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block

A: First, grasp yellow block and ...

Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Given <emb> Q: How to grasp blue block? A: First grasp yellow block and place it on the table, then grasp the blue block.

Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.

Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.

Describe the following <img>: A dog jumping over a hurdle at a dog show.

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372 x 18? A: 6696.Q: Write a Haiku about embodied LLMs. A: Embodied language. Models learn to understand. The world around them.

The main architectural idea of PaLM-E is to inject continuous, embodied observations such as images, state estimates, or other sensor modalities into the language embedding space of a pre-trained language model. This is realized by encoding the continuous observations into a sequence of vectors with the same dimension as the embedding space of the language tokens. The continuous information is hence injected into the language model in an analogous way to language tokens. PaLM-E is a decoder-only LLM that generates textual completions autoregressively given a prefix or prompt. We call our model PaLM-**E**, since we use PaLM (Chowdhery et al., 2022) (https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html) as the pre-trained language model, and make it **E**mbodied.

Results

We show a few example videos showing how PaLM-E can be used to plan and execute long horizon tasks on two different real embodiments. Please note, that all of these results were obtained using the same model trained on all data. In the first video, we execute a long-horizon instruction "bring me the rice chips from the drawer" that includes multiple planning steps as well as incorporating visual feedback from the robot's camera. Finally, show another example on the same robot where the instruction is "bring me a green star". Green star is an object that this robot wasn't directly exposed to.

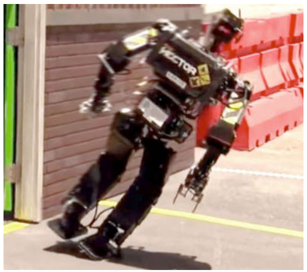
In the following part, we show PaLM-E controlling a table top robot arranging blocks. We show the PaLM-E can successfully plan over multiple stages based on visual and language input. Our model is able to successfully plan a long-horizon task "sort blocks by colors into different corners" . Another example of planning over multiple stages and incorporating visual feedback over long time horizons. Finally, we demonstrate another example of long-horizon pushing tasks on this robot. The first instruction is "move remaining blocks to the group". PaLM-E sequences step-by-step commands to the low-level policy such as "move the yellow hexagon to the green star", and "move the blue triangle to the group".

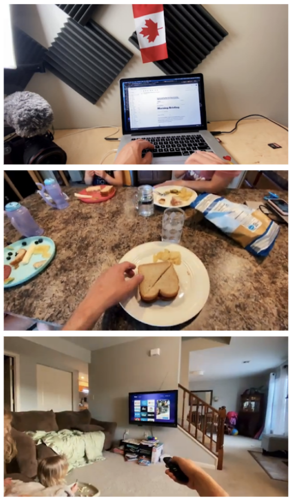
Next, we demonstrate two examples of generalization. In the case below the instruction is "push red blocks to the coffee cup". The dataset contains only three demonstrations with the coffee cup in them, and none of them included red blocks. We show another generalization example, where the instruction is "push green blocks to the turtle". The robot is able to successfully execute this task even though it has never seen the turtle before.

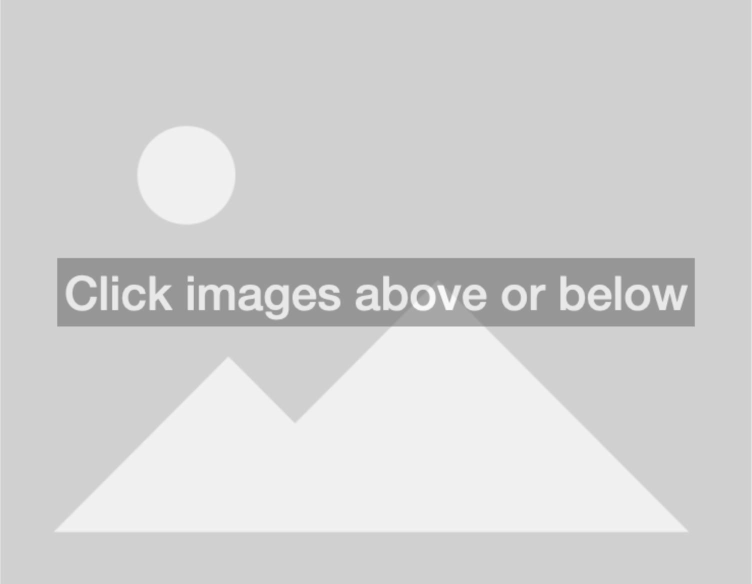
In addition to unlocking new capabilities in robot planning. PaLM-E is a competent Vision-Language Model. Please check out our paper (https://arxiv.org/abs/2303.03378) for more details and see the dmeo below.

Demo

The examples below are all example completions (in orange) from PaLM-E. The prompt is the one or more images and the text in gray.

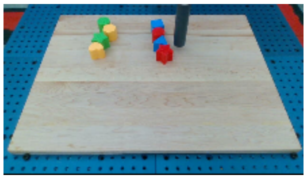


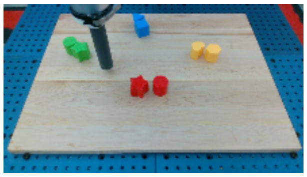


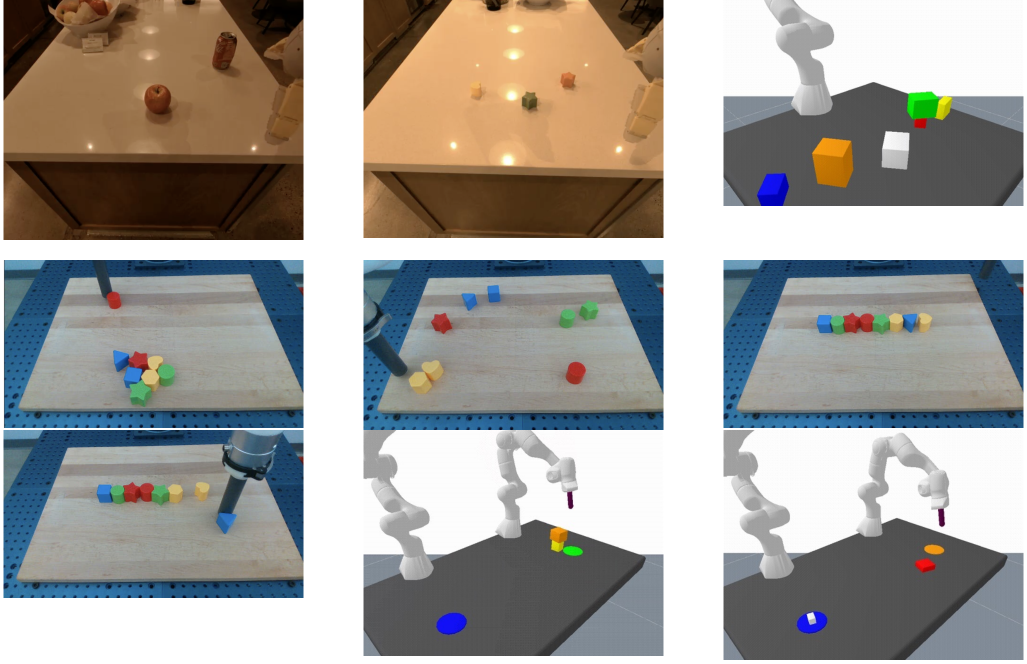


Prompt text in gray.

PaLM-E response in orange shade.







Citation

[arxiv version] (https://arxiv.org/abs/2303.03378)

```
@inproceedings{driess2023palme,
    title={PaLM-E: An Embodied Multimodal Language Model},
    author={Driess, Danny and Xia, Fei and Sajjadi, Mehdi S. M.
and Lynch, Corey and Chowdhery, Aakanksha and Ichter, Brian and
Wahid, Ayzaan and Tompson, Jonathan and Vuong, Quan and Yu,
Tianhe and Huang, Wenlong and Chebotar, Yevgen and Sermanet,
Pierre and Duckworth, Daniel and Levine, Sergey and Vanhoucke,
Vincent and Hausman, Karol and Toussaint, Marc and Greff, Klaus
and Zeng, Andy and Mordatch, Igor and Florence, Pete},
    booktitle={arXiv preprint arXiv:2303.03378},
    year={2023}
}
```

Acknowledgements

The authors would like to thank, for their advice, help and support: Xi Chen, Etienne Pot, Sebastian Goodman, Ted Xiao, Keerthana Gopalakrishnan, Kehang Han, Henryk Michalewski, Neil Houlsby, Basil Mustafa, Justin Gilmer, Yonghui Wu, Erica Moreira, Victor Gomes, Tom Duerig, and Kendra Byrne.