

Large models as world representations for robotic downstream tasks

BigMAC workshop
ICCV 2023

Maria Attarian

Research Engineer & PhD Student
Google DeepMind, University of Toronto

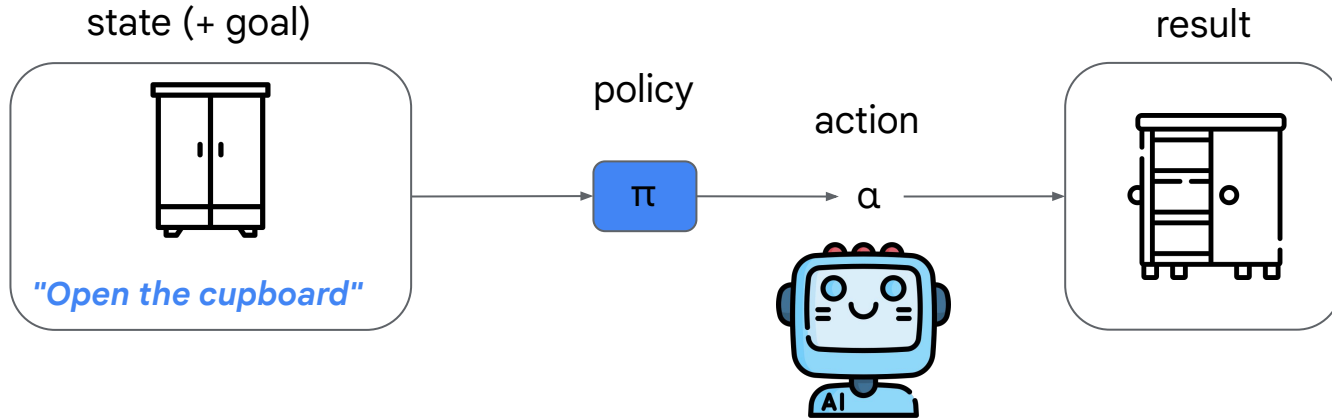
01

Robots acting in the world

How do robots learn how to act anyway?

What is policy learning?

Learning a function that maps observations (and possibly goals) to actions.



Know what your embodiment can do

Exploration

- Random trajectories
- Learning from play

+++ How humans learn early in life

--- Not useful actions in and of themselves



Move with intention

Imitation

- Learn to perform useful tasks
- Goal conditioned Imitation Learning
- Reinforcement Learning

+++ Can successfully perform some tasks

--- Handles only relatively simple tasks



02

LMs as world modelers and planners

Making sense of it all with large models

Why is reasoning and
planning hard?

Planning with an LLM

Language only

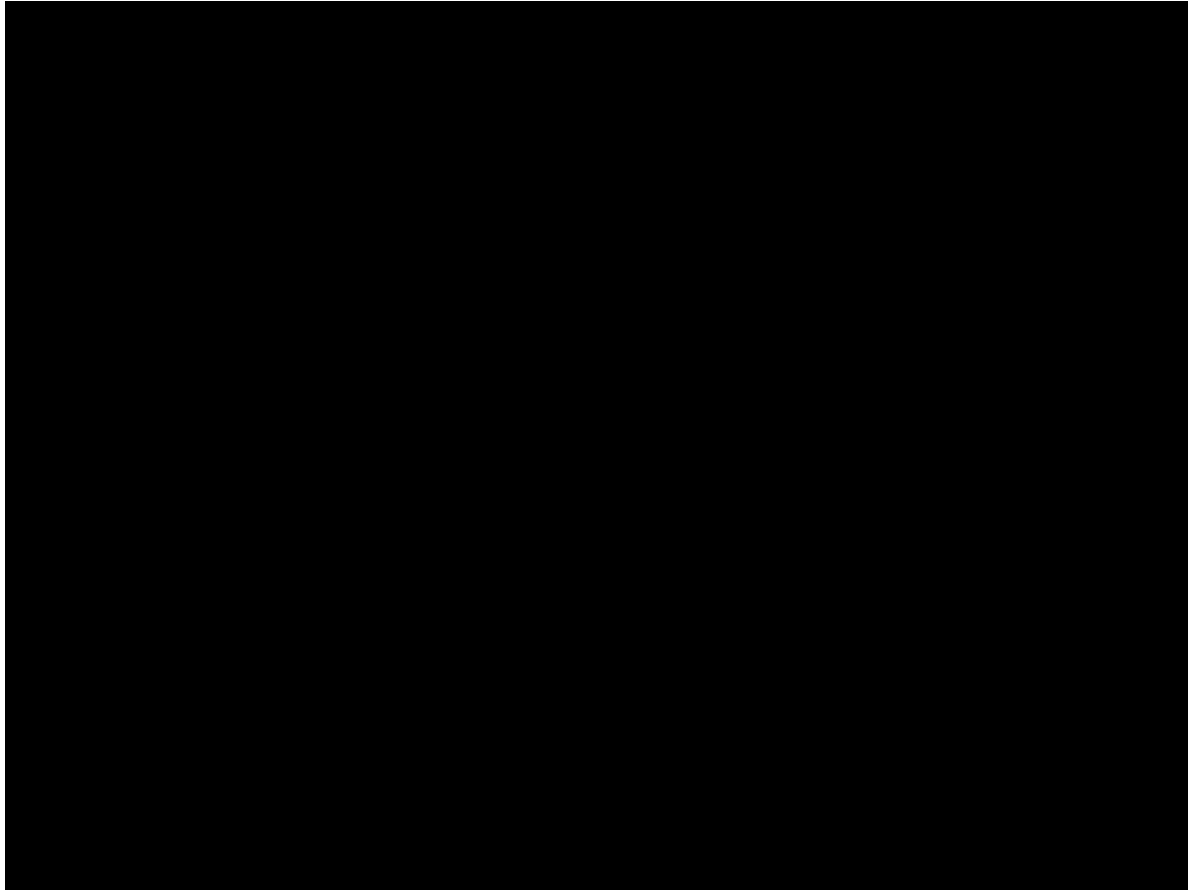
- LLMs have a lot of apriori world knowledge
- Visual grounding is missing
- In the physical world, you can't do something that isn't feasible



How do I make an omelette with just the ingredients in front of me. Please be succinct.



Of course! First, let me know which ingredients you have in front of you, and I'll guide you on how to make an omelette with them.






Full video: https://www.youtube.com/watch?v=cDA3_5982h8



Planning with a VLM


Vision + Language

- Vision grounding now exists
- Still this does not emit robot commands
- Even if it did, there is no notion of physical constraints

 Tell me how to make a cake with only the items I have on my counter. Please be succinct. 



[View other drafts](#)  

 To make a cake with the ingredients in your image (eggs, flour, sugar, and a spoon), you will need to:

1. Preheat your oven to 350 degrees Fahrenheit (175 degrees Celsius).
2. Grease and flour a small baking pan.
3. In a bowl, whisk together the eggs and sugar until light and fluffy.
4. Gradually add the flour, whisking until just combined.
5. Pour the batter into the prepared baking pan and bake for 20-25 minutes, or until a toothpick inserted into the center comes out clean.
6. Let the cake cool completely before frosting and serving.

Optional: You can add a teaspoon of vanilla extract to the batter for extra flavor.

LMs for Direct Action Prediction

Vision + Language \rightarrow Actions

- LMs trained to output actions
- Physical constraints are still not encoded
- Common sense is still not encoded
- Environments and embodiments can vary a lot

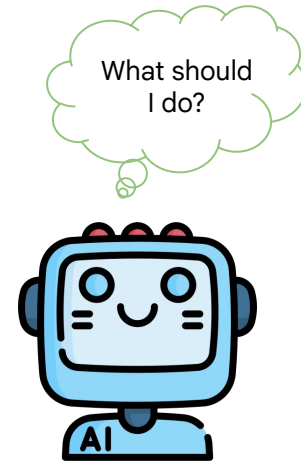


Why is a perfect
planner not enough?

Policy learning is hard...

...even with a perfect planner

- What are the gaps?
 - Adaptation to changing dynamics
 - Error recovery
 - Safety
 - Generalization to new tasks and environments
 - Generation of novel motions



03

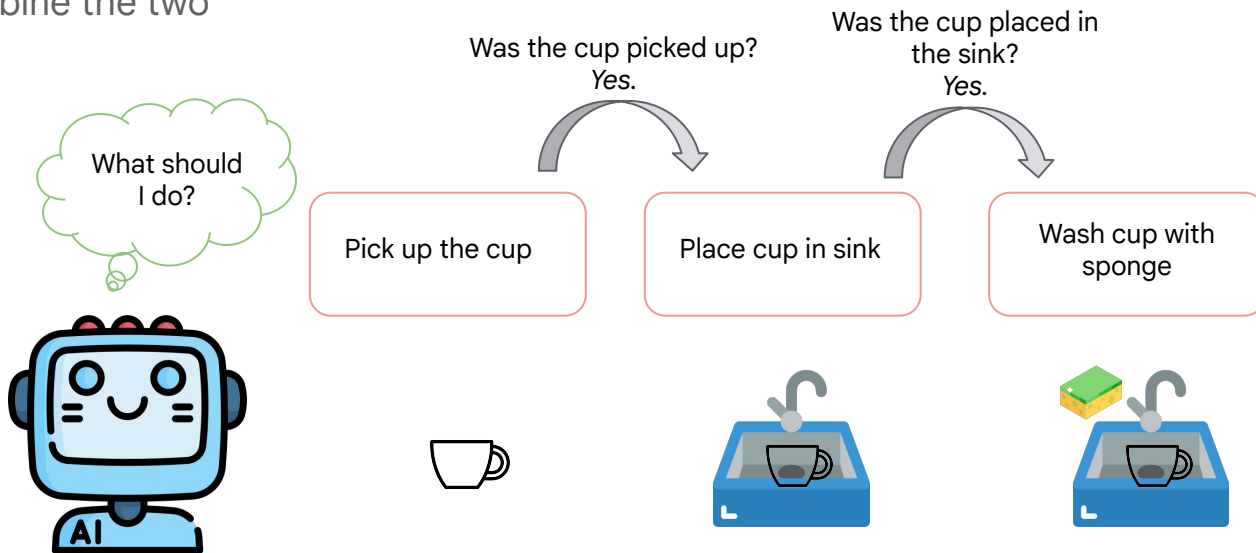
What's next?

Cool, interesting research directions!

Improving planning abilities

Make VLMs better at planning

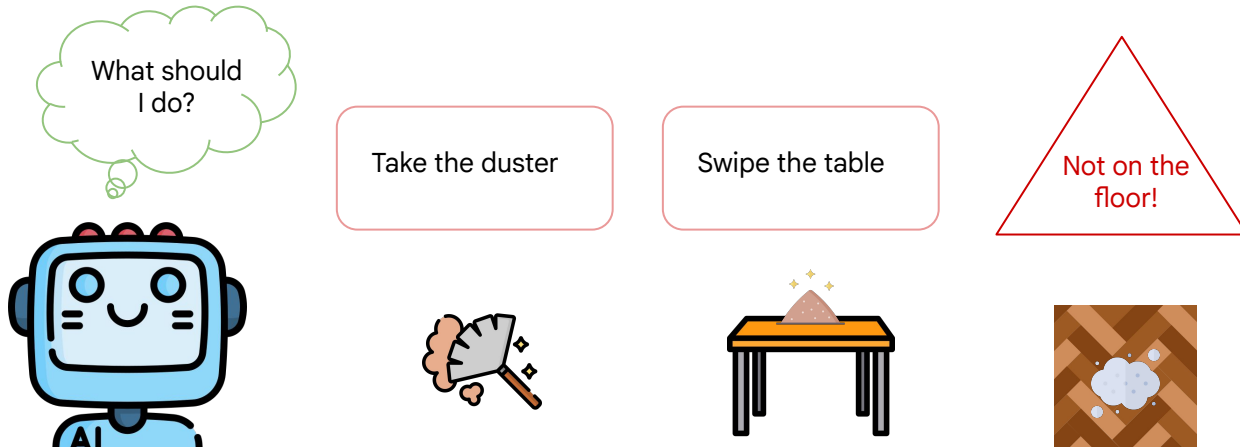
- Cast the planning problem into a VQA problem
- Finetune specifically for planning required by a robot agent
- Combine the two



World priors

Encode "laws" of the physical world

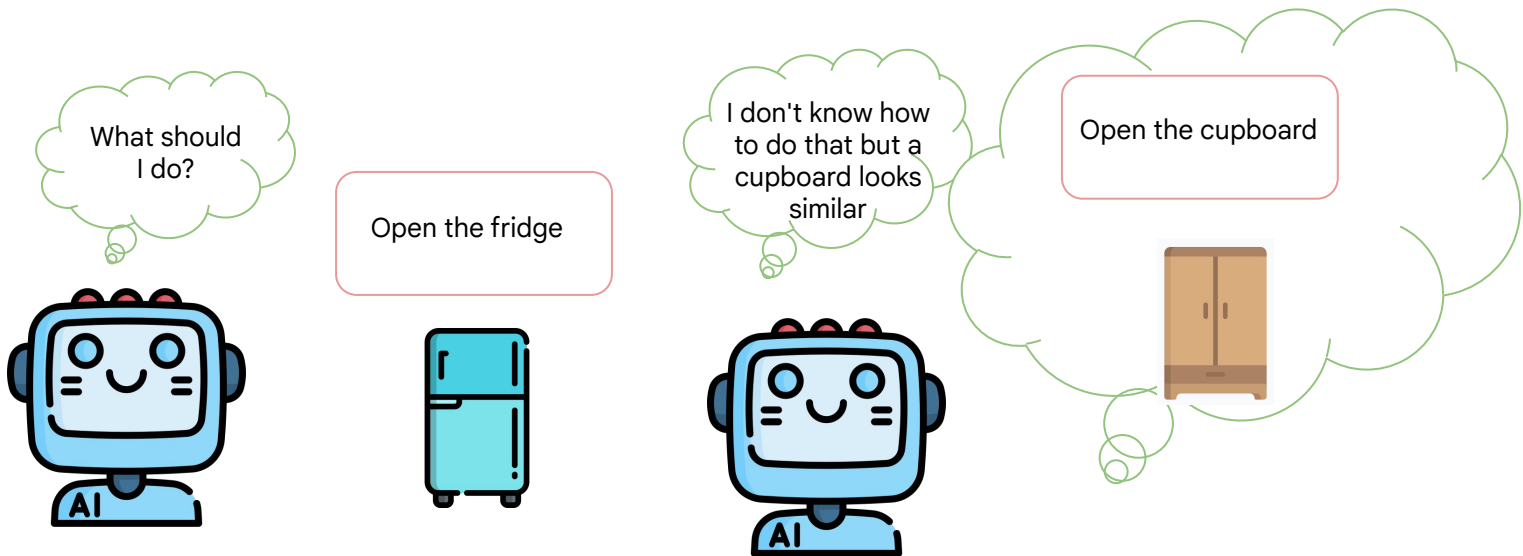
- Maybe full video generation is not necessary but encoding in a latent space things that we, humans, treat as implied
- Chain-of-thought on what is physically possible and/or implied
- Finetune on constraints of the environment for a given scene



Skill and action spaces

Look at the problem from an "action-centric" perspective

- Skills/tasks can be categorized based on the type of motion needed to be performed
- This could be used as a novel similarity metric
- Can we generate novel motions for unseen tasks if we know which category those tasks fall into?



Summary

One day our robots will be able to make a PB&J sandwich!





Thank you!