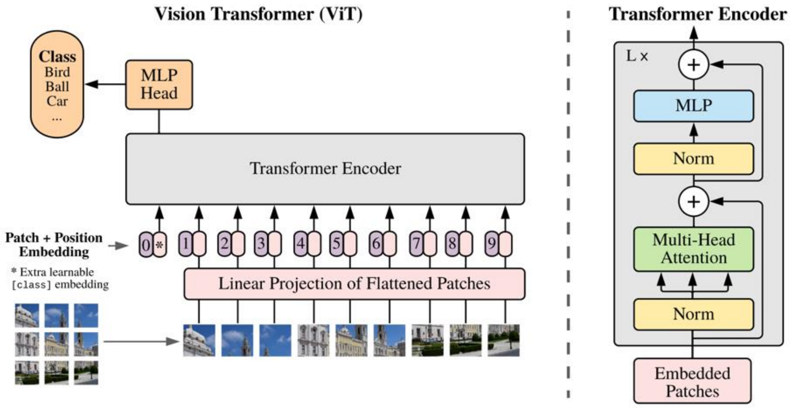
Robotic Foundation Models

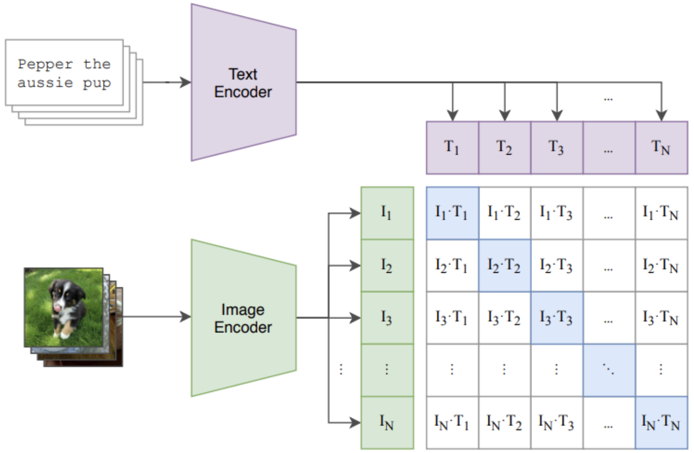
Transfomers for Robotics, Lecture 6, Nikolaus Correll

So far…

- Self-attention has replaced recurrent models as it is easier to train (parallel) and numerically stabler (vanishing gradients)
- Transformers use self-attention for text (lecture 2), images (lecture 3), and multi-modal input (lecture 4). They can also condition output to generate images (lecture 5, sound, trajectories, …)
- Today: robotic foundation models



Encoder for image classification

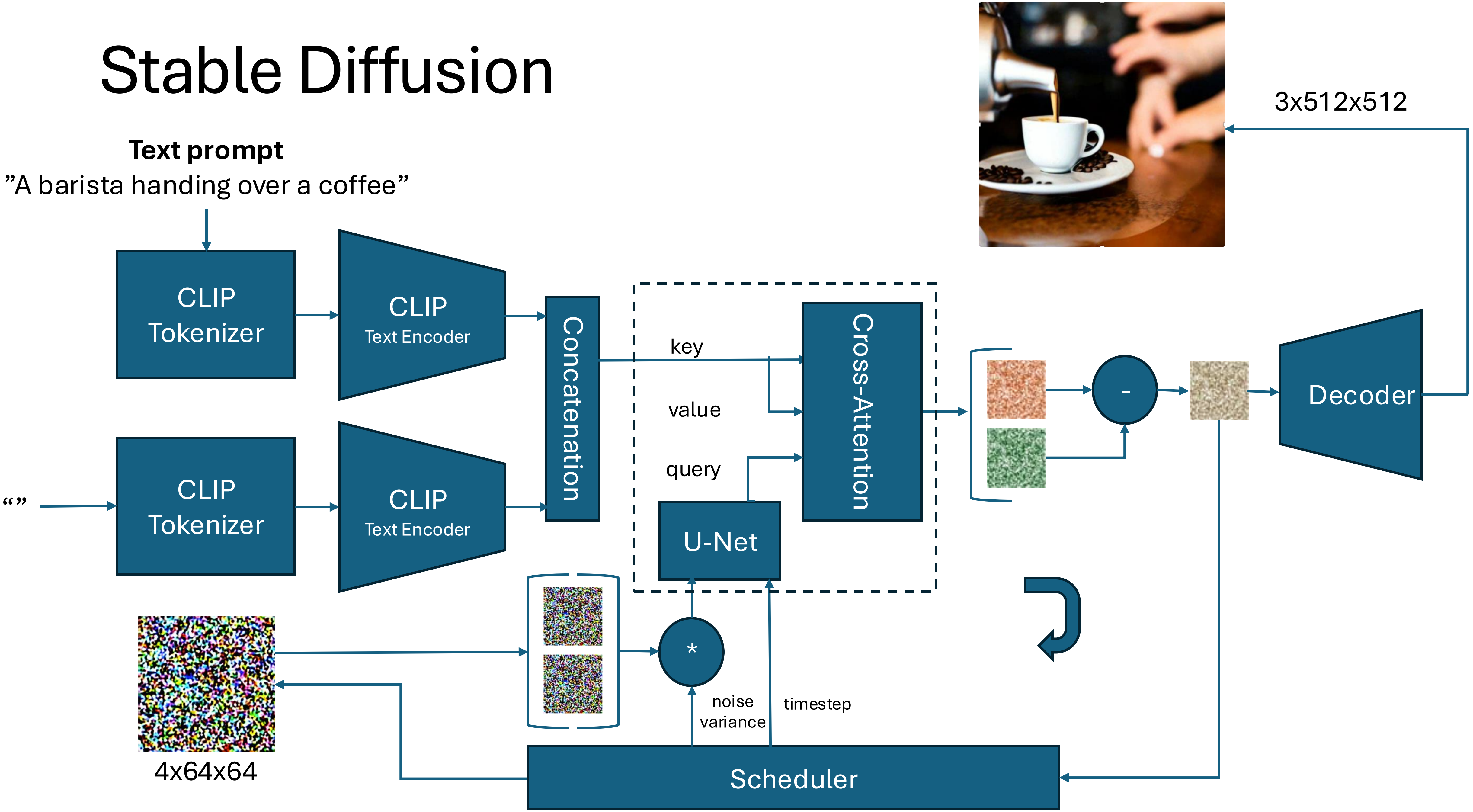


Multi-modal contrastive learning

Last Week: Stable Diffusion

- Prompt: `a barista handing over a coffee`
- Start with random noise
- A trained U-Net predicts the noise that must have been added to an image of the prompt in a latent space
- Cross-attention with CLIP embedding of promp
- An Auto-Encoder scales the image up

Stable Diffusion



Text prompt

”A barista handing over a coffee”

“”

CLIP Tokenizer

CLIP Text Encoder

CLIP Tokenizer

CLIP Text Encoder

Concatenation

key

value

query

Cross-Attention

U-Net

-

*

noise variance

timestep

Scheduler

Decoder

4x64x64

3x512x512

Today

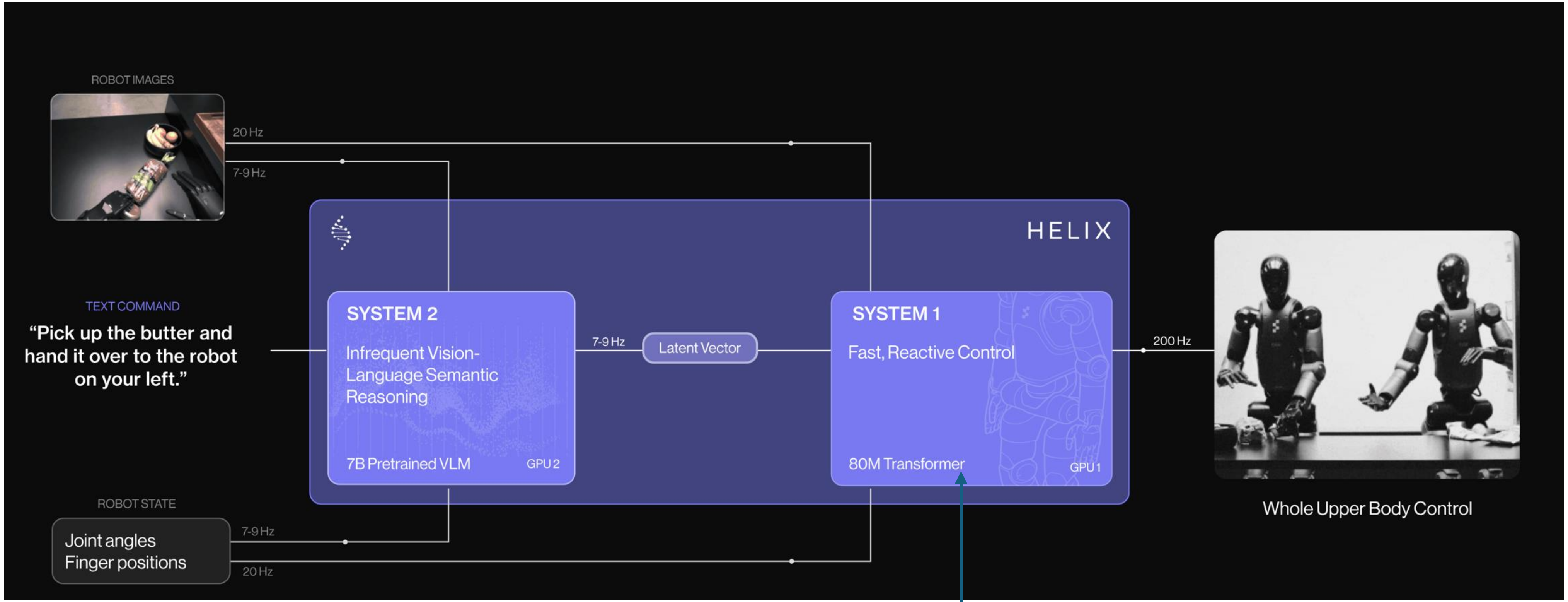
- Homework (30 min)
- LLMs + Foundation Models (60 min)
- Paper (30 min)
- Project planning (30 min)

Homework

- Please send me your medium username
- Let's work through some submissions…
  - https://medium.com/@scampbell2792/solving-mazes-with-an-encoder-decoder-transformer-2bfffcf2ab54
  - https://medium.com/@noahmuthler/fine-tuned-encoder-decoder-for-english-to-spanish-conjugation-translation-e41c2bac32cf
  - https://medium.com/@besp2426/understanding-the-transformer-encoder-and-its-use-in-translation-c8d4cc44d671
- Really awesome work!

Figure AI’s Helix (Feb 22, 2025)

- https://www.figure.ai/news/helix



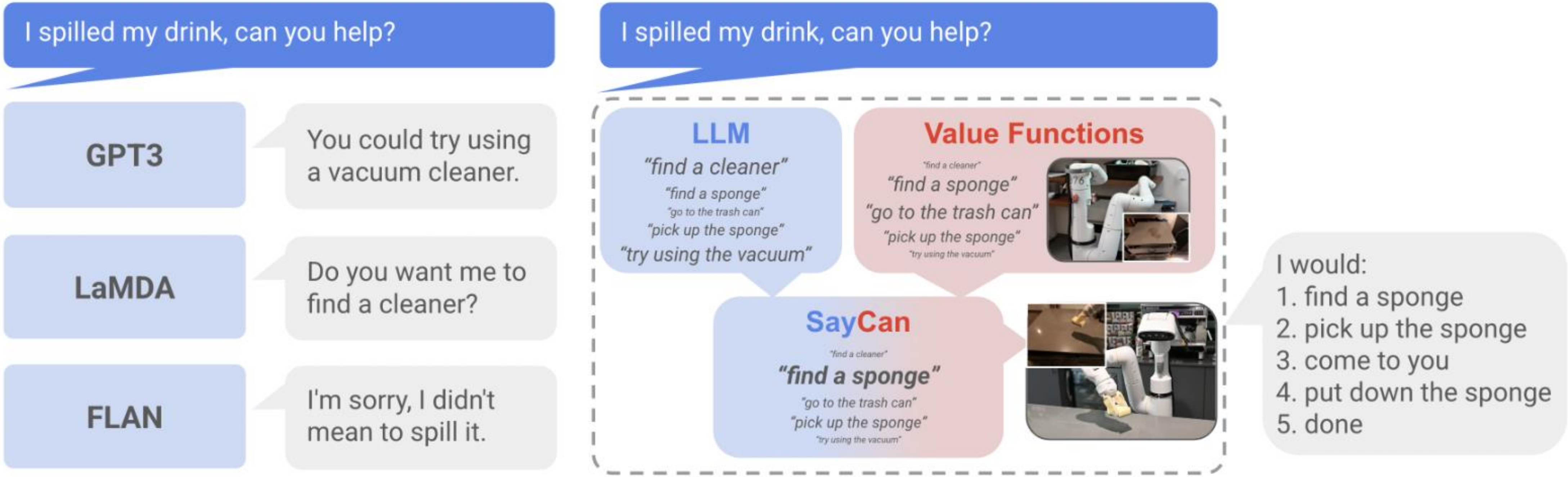
“Robotic Foundation Model”

Helix Training

- 500h of training
- Auto-labeling using VLM “What instruction would you have given the robot to get the action seen in this video?”
- S1: 80M parameter cross-attention encoder-decoder transformer, pretrained in simulation
- My thoughts:
  - splitting into high/low bandwidth definitely the way to go
  - Do we really need a foundation model for ”System 1”?
  - How can we generalize to “pick up paper airplane” vs. “pick up orange”

Do As I Can, Not As I Say: Grounding Language in Robotic Affordances

Google / Everyday Robotics, 2022



LLM reasoning x "Foundation Models"

SayCan

- https://www.youtube.com/watch?v=ysFav0b472w

Formalism

- Language model: $p(W) = \Pi_{j=0}^{n} p(w_j | w_{<j})$ $\quad W = \{w_0, w_1, w_2, ..., w_n\}$
- Markov Decision Process (MDP): $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ $\quad Q^{\pi}(s, a)$
- Which action *a* should I take in state *s* to maximize reward?
  (The amount of actions is given by pre-trained behaviors)

$$p(c_i | i, s, \ell_\pi)$$

ℓ_π : language description of skill pi, e.g. “Find a sponge”

s : current state

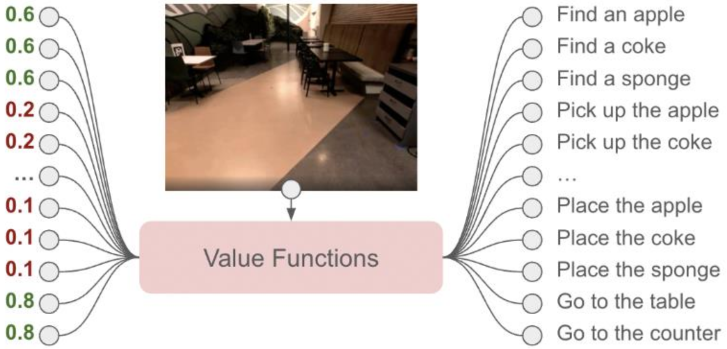
c : completion probability

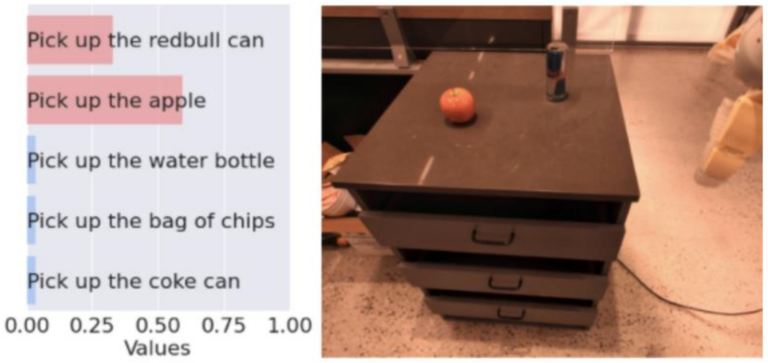
i : language instruction, e.g. “clean up the mess”

$$p(c_i | i, s, \ell_\pi) \propto \overset{\text{World Grounding}}{p(c_\pi | s, \ell_\pi)}\underset{\text{Task Grounding}}{p(\ell_\pi | i)}$$

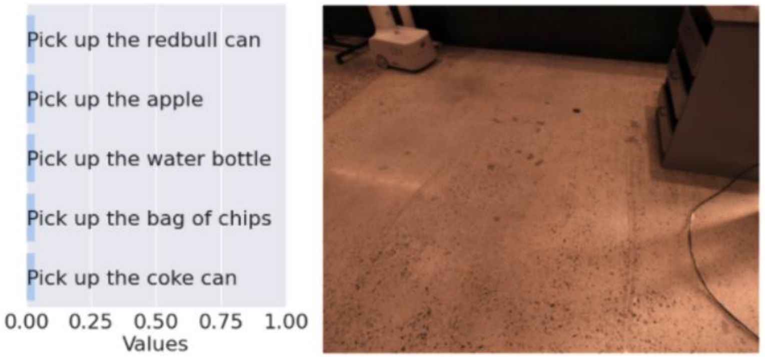
SayCan vs. Prompt Engineering

- Prompt engineering can bias the model to adhere to a specific structure
- Possibility to select unavailable / impossible actions
- Solution: Selection of limited number of actions using a value function
- Task grounding: how likely is the language model to select this completion?
- World grounding: how likely is the action to be feasible



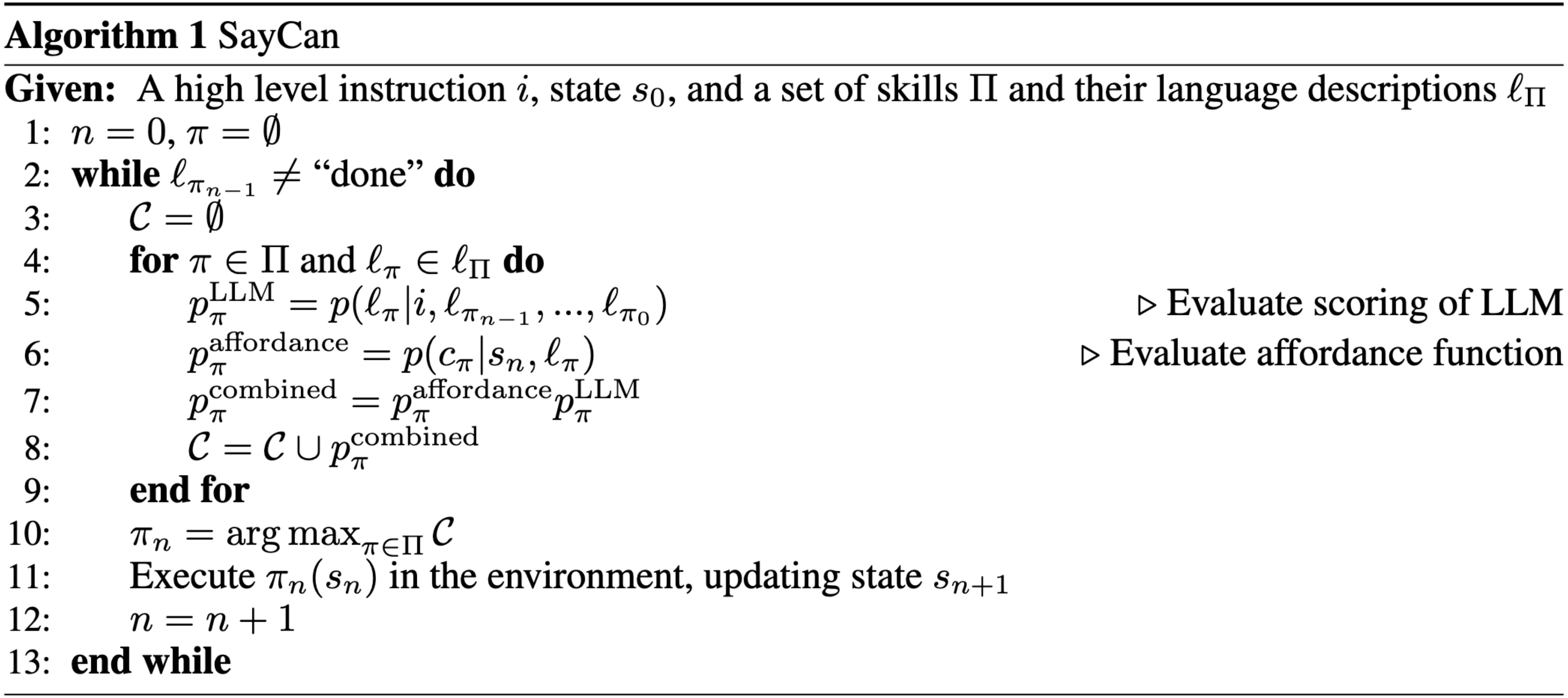


Feasible actions



Infeasible actions

Algorithm



Algorithm 1 SayCan

```
Given: A high level instruction $i$, state $s_0$, and a set of skills $\Pi$ and their language descriptions $\ell_\Pi$
$n = 0, \pi = \emptyset$
while $\ell_{\pi_{n-1}} \neq$ "done" do
    $\mathcal{C} = \emptyset$
    for $\pi \in \Pi$ and $\ell_\pi \in \ell_\Pi$ do
        $p_\pi^{\text{LLM}} = p(\ell_\pi | i, \ell_{\pi_{n-1}}, ..., \ell_{\pi_0})$        ▷ Evaluate scoring of LLM
        $p_\pi^{\text{affordance}} = p(c_\pi | s_n, \ell_\pi)$        ▷ Evaluate affordance function
        $p_\pi^{\text{combined}} = p_\pi^{\text{affordance}} p_\pi^{\text{LLM}}$
        $\mathcal{C} = \mathcal{C} \cup p_\pi^{\text{combined}}$
    end for
    $\pi_n = \arg\max_{\pi \in \Pi} \mathcal{C}$
    Execute $\pi_n(s_n)$ in the environment, updating state $s_{n+1}$
    $n = n + 1$
end while
```

Learning individual tasks: Approach 1

- Jang E, Irpan A, Khansari M, Kappler D, Ebert F, Lynch C, Levine S, Finn C. **Bc-z: Zero-shot task generalization with robotic imitation learning**. In Conference on Robot Learning 2022 Jan 11 (pp. 991-1002). PMLR.

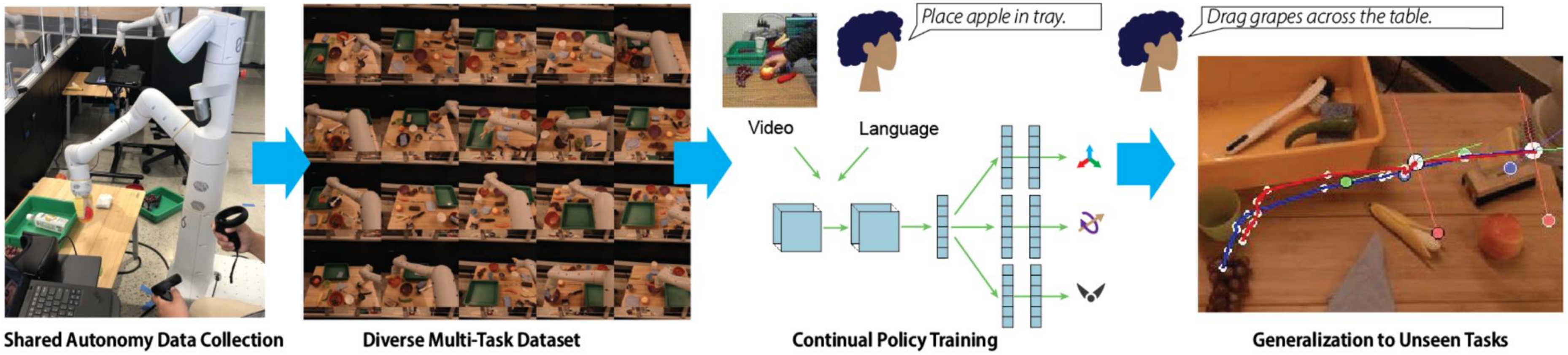
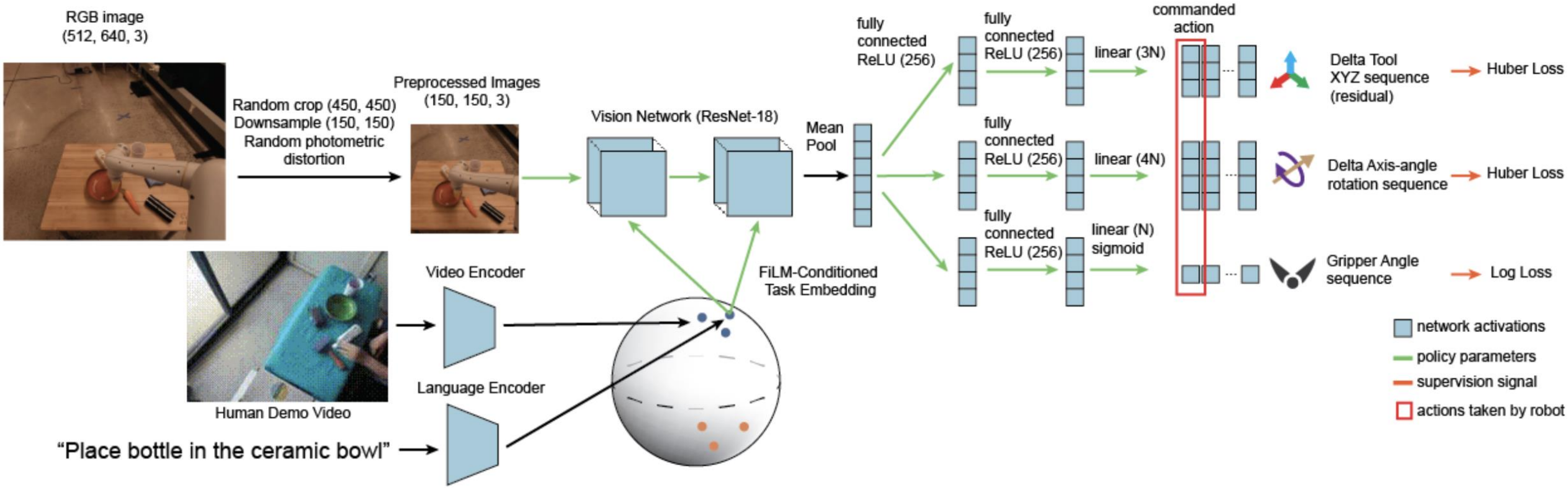
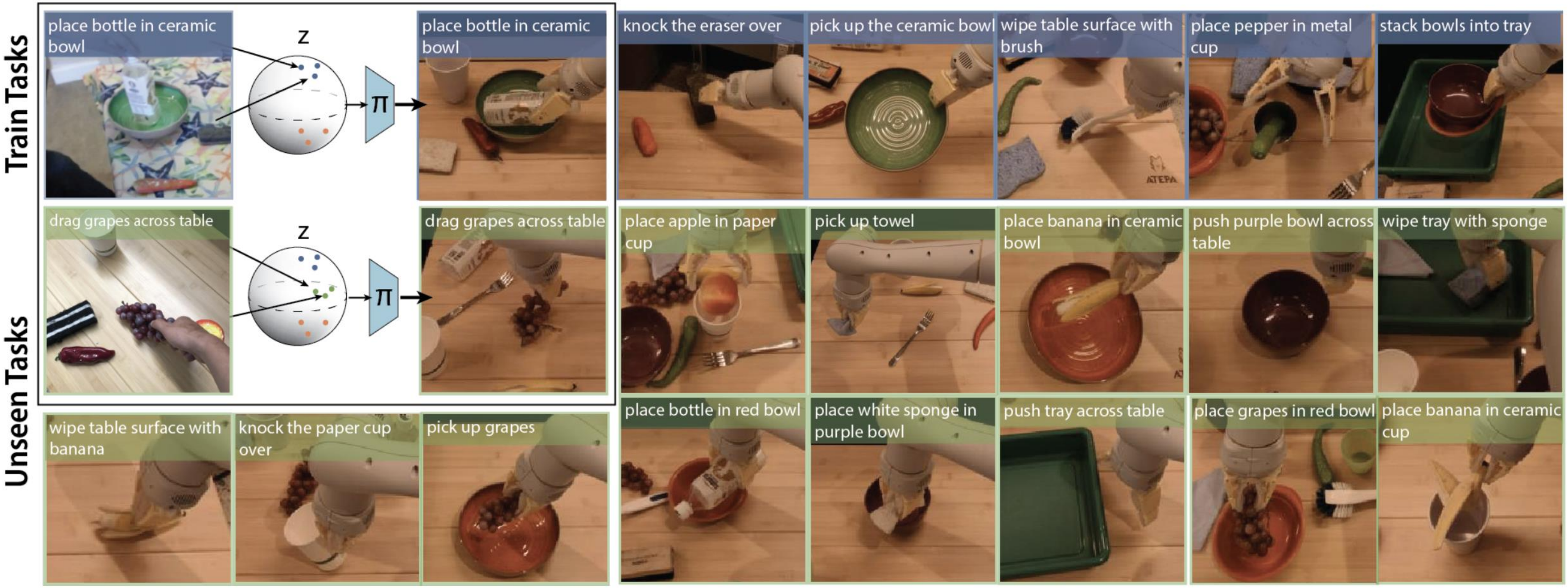

Place apple in tray.
Drag grapes across the table.
Video
Language
Shared Autonomy Data Collection
Diverse Multi-Task Dataset
Continual Policy Training
Generalization to Unseen Tasks

Figure 1: Overview of BC-Z. We collect a large-scale dataset (25,877 episodes) of 100 diverse manipulation tasks, and train a 7-DoF multi-task policy that conditions on task language strings or human video. We show this system produces a policy that is capable of generalizing zero-shot to new unseen tasks.

Network architecture



More examples



Learning individual tasks: Approach 2

- Kalashnikov, D., Varley, J., Chebotar, Y., Swanson, B., Jonschkowski, R., Finn, C., Levine, S. and Hausman, K., 2021. **Mt-opt: Continuous multi-task robotic reinforcement learning at scale**. *arXiv preprint arXiv:2104.08212*.
- Multi-task Reinforcement Learning framework
- Action-space wrist position, orientation, and gripper
- State from camera
- Task embedding
- **Both approaches use value functions learned via Temporal Differences Backup (TD Backup)**

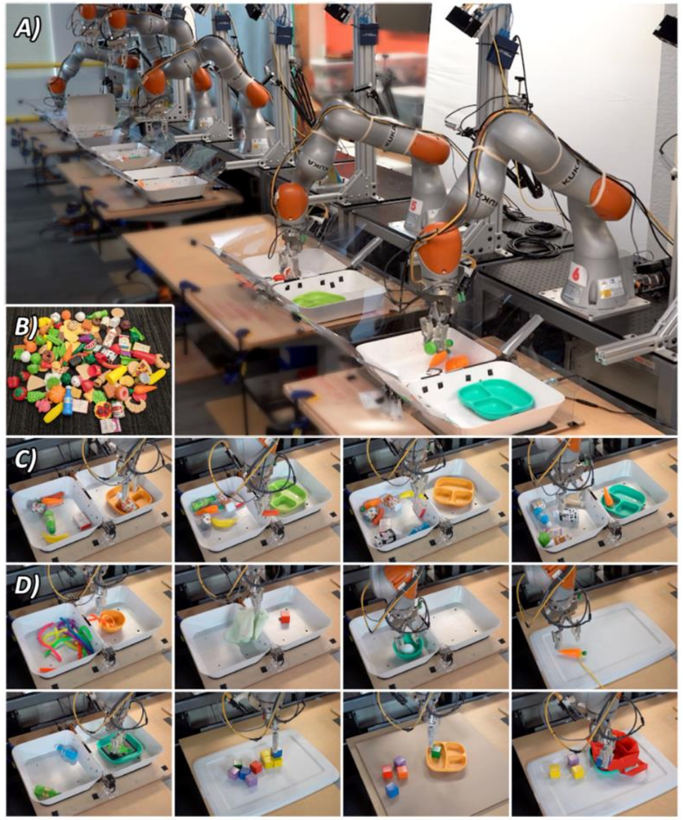


Fig. 1: A) Multi-task data collection. B) Training objects. C) Sample of tasks that the system is trained on. D) Sample of behaviorally and visually distinct tasks such as covering, chasing, alignment, which we show our method can adapt to. MT-Opt learns new tasks faster (potentially zero-shot if there is sufficient overlap with existing tasks), and with less data compared to learning the new task in isolation.

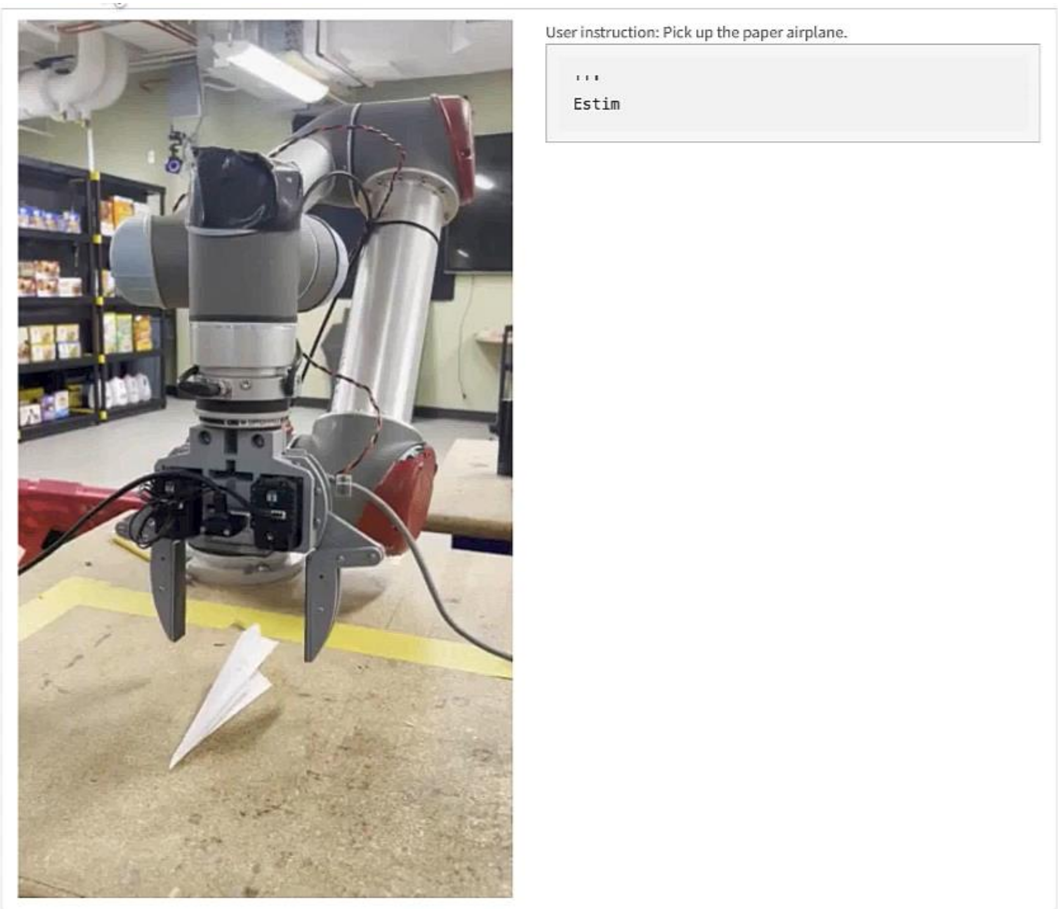
SayCan vs. State-of-the-Art

- PaLM was a very powerful model, but LLMs did improve
- Agent frameworks can deal with “tools”
- “Value function” in SayCan might be replaced by Vision Language Models (VLM)
- General observation: robot agnostic LLM/VLMs can be trained on internet-scale data, robots can not

LLM-based grasping tool: DeliGrasp

- Goal: grasp *unknown* objects
- Find “paper airplane” using CLIP-like model
- Use LLM (GPT-4) to estimate mass, friction constant, and spring constant
- Outperforms adaptive strategies for a variety of objects

https://deligrasp.github.io/



User instruction: Pick up the paper airplane.

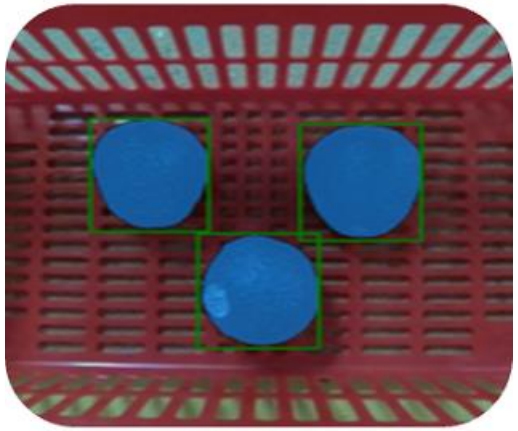
```
'''
Estim
```

Xie, William, Maria Valentini, Jensen Lavering, and Nikolaus Correll. "DeliGrasp: Inferring Object Properties with LLMs for Adaptive Grasp Policies." In *8th Annual Conference on Robot Learning*. 2024.

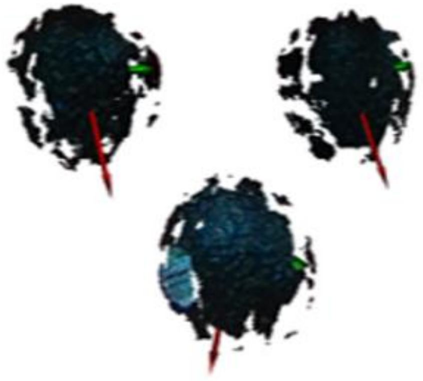
Open-world grasping

- OWL-ViT for finding objects (zero shot)
- “Segment Anything” to segment objects (zero shot)
- Principal component analysis on point cloud for grasp generation
- Align gripper perpendicular to the smallest of the first two Eigenvectors

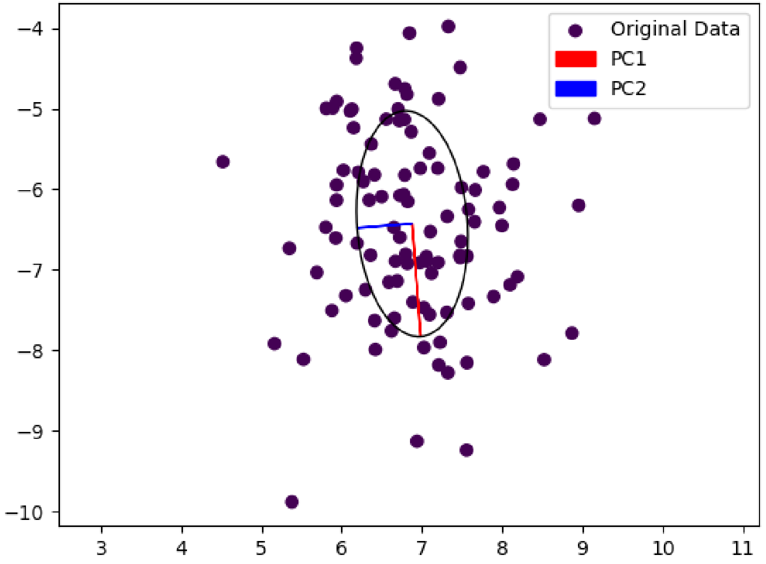
“avocado”



OWL-ViT + SAM

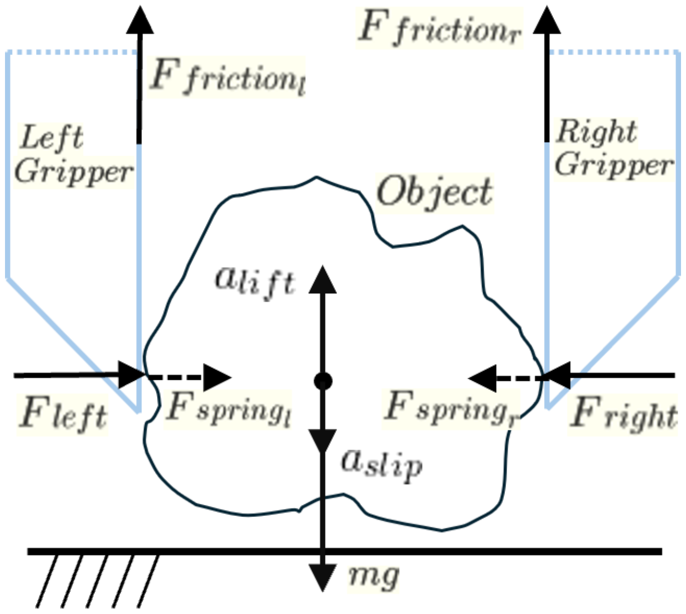


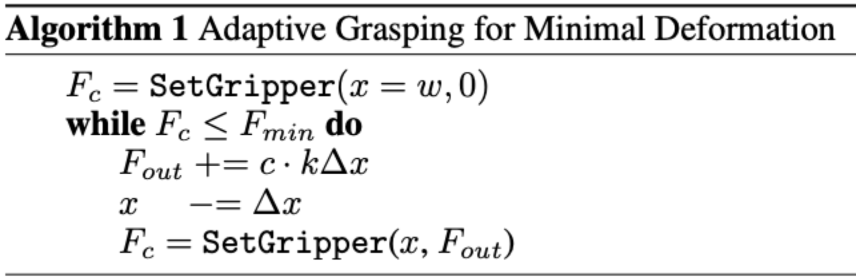
Point Cloud + PCA



Adaptive grasp policy

- Desired: minimally deforming grasp
- Decrease aperture and force at the same time
- Increase gripper output force F_{out} and decrease gripper aperture *x* until sensing a contact force F_c greater than the target F_{min}
- Two prompt “chain-of-thought”:
    - Estimate object parameters
    - Generate adaptive controller code

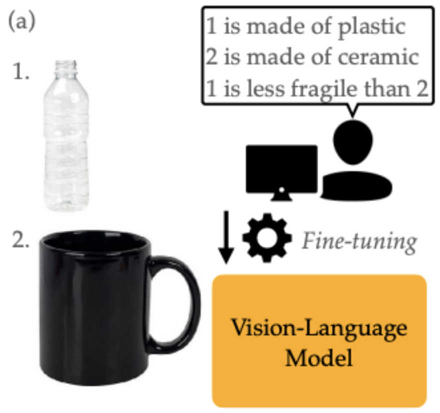


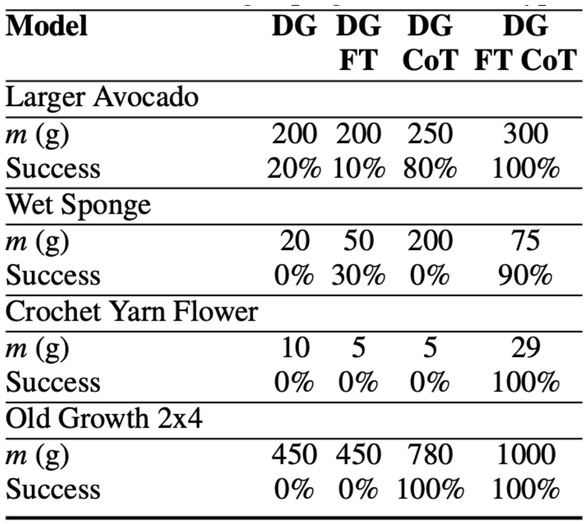

Algorithm 1 Adaptive Grasping for Minimal Deformation

```
$F_c$ = SetGripper($x = w, 0$)
while $F_c \leq F_{min}$ do
    $F_{out}$ += $c \cdot k\Delta x$
    $x$ -= $\Delta x$
    $F_c$ = SetGripper($x$, $F_{out}$)
```

Finetuning

- Finetune GPT-3.5-Turbo on 6000 captions of PhysObjects* images
- 276 household objects, only two are in Deligrasp (mandarin and plastic bottle)
- Statements about material properties and relative mass, size, and fragility
- **Improves grasping success rate**
- Potential for unsupervised, life-long learning





Model	DG	DG	DG	DG
		FT	CoT	FT CoT
Larger Avocado				
m (g)	200	200	250	300
Success	20%	10%	80%	100%
Wet Sponge				
m (g)	20	50	200	75
Success	0%	30%	0%	90%
Crochet Yarn Flower				
m (g)	10	5	5	29
Success	0%	0%	0%	100%
Old Growth 2x4				
m (g)	450	450	780	1000
Success	0%	0%	100%	100%

Gao, J., Sarkar, B., Xia, F., Xiao, T., Wu, J., Ichter, B., Majumdar, A. and Sadigh, D., 2024, May. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 12462-12469). IEEE.

Experiments

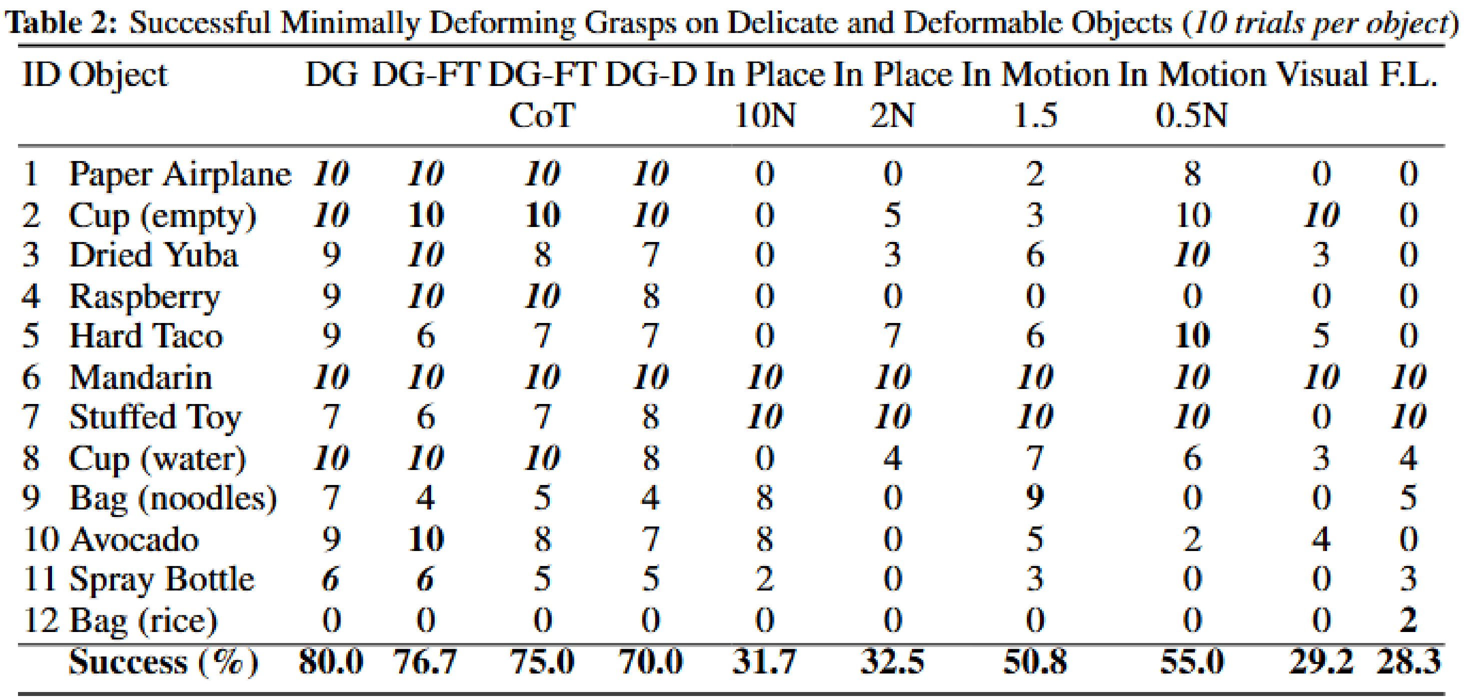
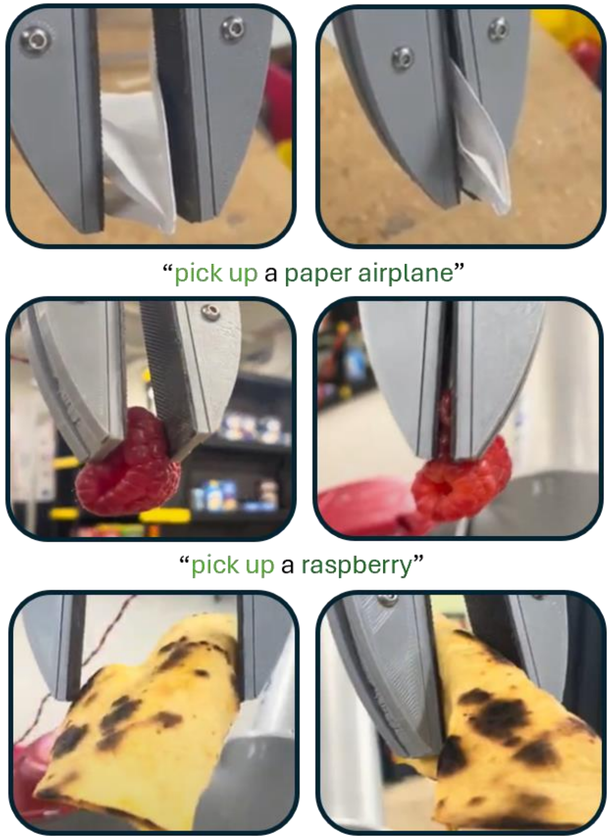


Table 2: Successful Minimally Deforming Grasps on Delicate and Deformable Objects (*10 trials per object*)

ID	Object	DG	DG-FT	DG-FT CoT	DG-D	In Place 10N	In Place 2N	In Motion 1.5	In Motion 0.5N	Visual	F.L.
1	Paper Airplane	***10***	***10***	***10***	***10***	0	0	2	8	0	0
2	Cup (empty)	***10***	**10**	**10**	***10***	0	5	3	10	***10***	0
3	Dried Yuba	9	***10***	8	7	0	3	6	***10***	3	0
4	Raspberry	9	***10***	***10***	8	0	0	0	0	0	0
5	Hard Taco	9	6	7	7	0	7	6	**10**	5	0
6	Mandarin	***10***	***10***	***10***	***10***	***10***	***10***	***10***	***10***	***10***	***10***
7	Stuffed Toy	7	6	7	8	***10***	***10***	***10***	***10***	0	***10***
8	Cup (water)	***10***	***10***	***10***	8	0	4	7	6	3	4
9	Bag (noodles)	7	4	5	4	8	0	**9**	0	0	5
10	Avocado	9	**10**	8	7	8	0	5	2	4	0
11	Spray Bottle	***6***	***6***	5	5	2	0	3	0	0	3
12	Bag (rice)	0	0	0	0	0	0	0	0	0	**2**
	Success (%)	**80.0**	**76.7**	**75.0**	**70.0**	**31.7**	**32.5**	**50.8**	**55.0**	**29.2**	**28.3**

Non-deforming ✓ Traditional Grasp ✗



"pick up a paper airplane"

"pick up a raspberry"

"pick up a hard taco"

Paper Line-up

- Monday, February 24th, 2025
  - Paper 1:
- Monday, March 3rd, 2025
  - Paper 3: Litrico, Mattia, et al. "Tadm: Temporally-aware diffusion model for neurodegenerative progression on brain mri." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2024. **Nolan Brady**
  - Paper 4:
- Monday, March 10th, 2025
  - Paper 5: TinyVLA Florian Frick
  - Paper 6:
- Monday, March 17th, 2025
  - Paper 7:
  - Paper 8:
- Monday, March 24th, 2025
  - Paper 2:

Paper candidates:

https://code-as-policies.github.io
https://sayplan.github.io/
https://voxposer.github.io/

Project

- Your final project is a 8-page paper following the format of the "Conference on Robotic LearningLinks to an external site." (CoRL).
- A valid final project for this class will require:
  - Articulate a clear hypothesis at the intersection of robotics and learning that is grounded in course content
  - Grounding of the proposed idea in the existing literature ("Introduction")
  - A detailed description of algorithms and tools used ("Materials and Methods"). These do not need to be original work, but sufficient for a beginning graduate student (you!) to reproduce the work
  - Experimental results (in-)validating your hypothesis ("Results")
  - A critical analysis of your experiments and outcomes ("Discussion")

Project: Heilmeier Catechism

- **What are you trying to do?**
  - Articulate your objectives using absolutely no jargon.
- **How is it done today, and what are the limits of current practice?**
  - This question addresses the state of the art and highlights the shortcomings of existing approaches.
- **What is new in your approach, and why do you think it will be successful?**
  - You need to explain what is innovative about your solution and why it is likely to succeed where others have failed.
- **Who cares?**
  - Identify the stakeholders who would benefit from the solution and why it matters to them.
- **If you're successful, what difference will it make?**
  - Explain the broader impact of your success, both practically and strategically.
- **What are the risks?**
  - Acknowledge the potential challenges and uncertainties that might prevent success.
- **How much will it cost?**
  - Estimate the financial investment required to complete the project.
- **How long will it take?**
  - Provide a realistic timeline for the project’s completion.
- **What are the midterm and final "exams" to check for success?**
  - Define the measurable milestones and metrics that will be used to evaluate progress and ultimate success.

Project: Heilmeier Catechism - Constraints

- **What are you trying to do?**
  - Articulate your objectives using absolutely no jargon.
- **How is it done today, and what are the limits of current practice?**
  - This question addresses the state of the art and highlights the shortcomings of existing approaches.
- **What is new in your approach, and why do you think it will be successful?**
  - You need to explain what is innovative about your solution and why it is likely to succeed where others have failed.
- **Who cares? OTHER RESEARCHERS**
  - Identify the stakeholders who would benefit from the solution and why it matters to them.
- **If you're successful, what difference will it make?**
  - Explain the broader impact of your success, both practically and strategically.
- **What are the risks? IF THERE ARE NO RISKS, ITS NOT RESEARCH**
  - Acknowledge the potential challenges and uncertainties that might prevent success.
- **How much will it cost? IT HAS TO BE FREE**
  - Estimate the financial investment required to complete the project.
- **How long will it take? YOU ONLY HAVE A FEW WEEKS. IF YOUR IDEA IS TOO BIG, WHAT IS THE FIRST STEP?**
  - Provide a realistic timeline for the project’s completion.
- **What are the midterm and final "exams" to check for success? WHERE WILL YOU BE IN 4 WEEKS, WHERE IN 8?**
  - Define the measurable milestones and metrics that will be used to evaluate progress and ultimate success.

Project: Heilmeier Catechism - Questions

- **What are you trying to do?**
  - Articulate your objectives using absolutely no jargon.
- **How is it done today, and what are the limits of current practice?**
  - This question addresses the state of the art and highlights the shortcomings of existing approaches.
- **What is new in your approach, and why do you think it will be successful?**
  - You need to explain what is innovative about your solution and why it is likely to succeed where others have failed.
- **Who cares?**
  - Identify the stakeholders who would benefit from the solution and why it matters to them.
- **If you're successful, what difference will it make? CAN YOU TRAIN FASTER, USE LESS MEMORY, IMPROVE ACCURACY? WILL IT BECOME MORE EXPLAINABLE, MORE COMPACT, EASIER TO UNDERSTAND?**
  - Explain the broader impact of your success, both practically and strategically.
- **What are the risks?**
  - Acknowledge the potential challenges and uncertainties that might prevent success.
- **How much will it cost?**
  - Estimate the financial investment required to complete the project.
- **How long will it take?**
  - Provide a realistic timeline for the project’s completion.
- **What are the midterm and final "exams" to check for success?**
  - Define the measurable milestones and metrics that will be used to evaluate progress and ultimate success.

Next week

- Think about a final project using the Heilmeier Catechism
- Form teams (up to three students)
- Select appropriate paper to present