2023 embedded VISION SUMMIT

Vision-Language Representations for Robotics

Dinesh Jayaraman

Assistant Professor,

University of Pennsylvania

Funding agencies:

National Science Foundation

Office of Naval Research Science & Technology

NEC Laboratories America Relentless passion for innovation

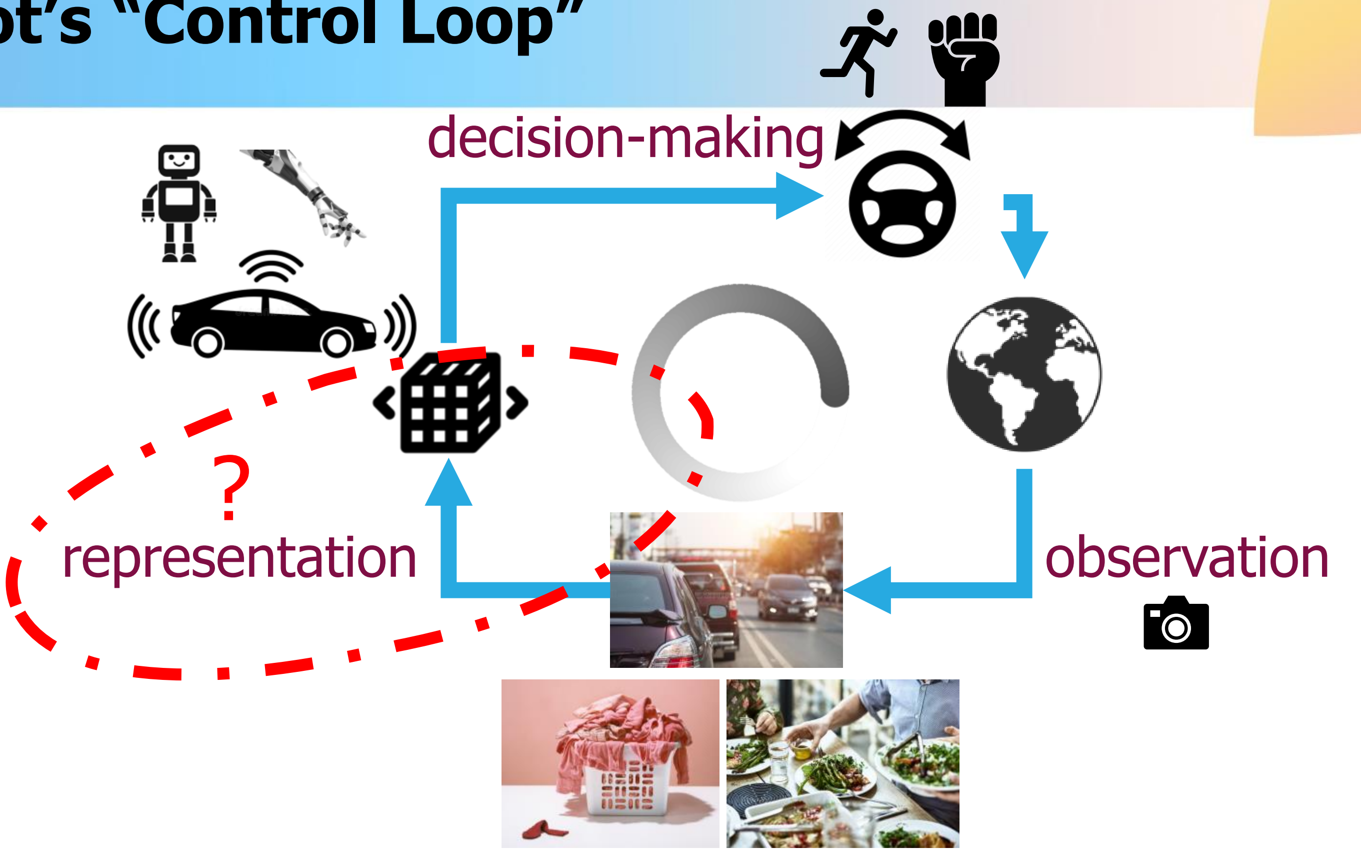
amazon

GE Research

Penn Engineering | GRASP Laboratory

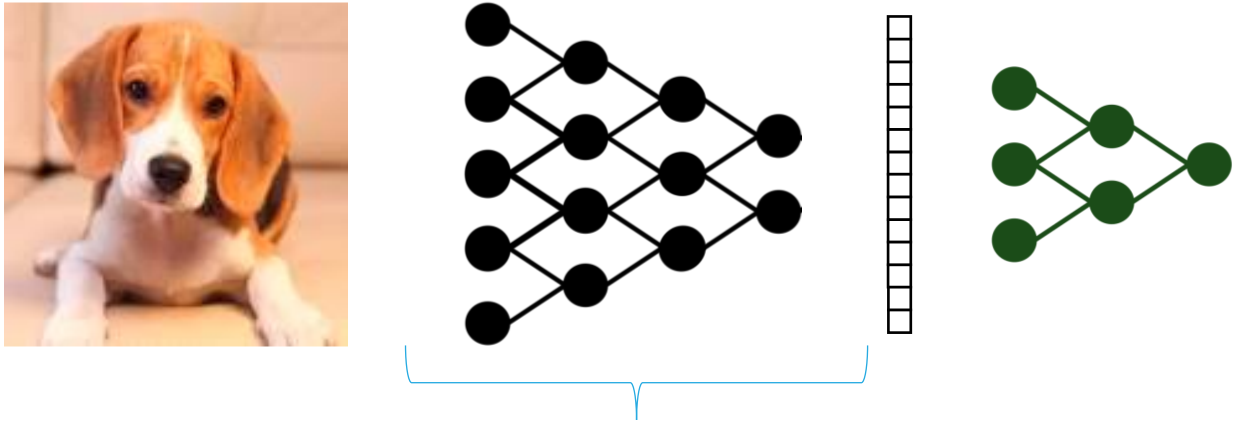
General Robotics, Automation, Sensing & Perception Lab

A Robot's "Control Loop"



How should the robot represent the information in its visual observations?

What is a Good Visual Representation?



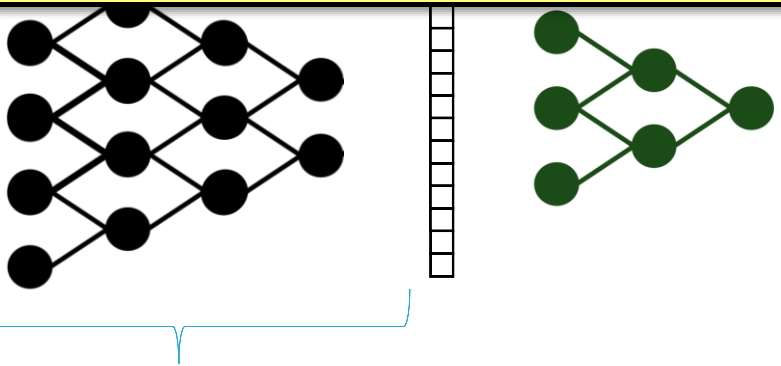
Current state-of-the-art for many computer vision tasks involves learned representations that are:
pretrained without supervision!

What is a Good Visual Representation?

… for Recognition?

Image $\mathbf{x}$ $\phi(\cdot)$ $\phi(\mathbf{x})$

Good representations organize information conveniently *for the task*.



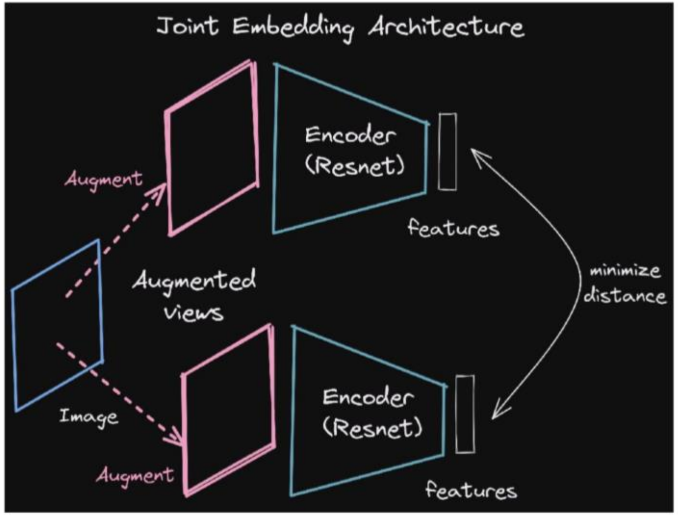
"dog"

representation encoder

Current state-of-the-art for many computer vision tasks involves learned representations that are:
pretrained without supervision!

Background: Contrastive Unsupervised Learning

Pull two views of the same image together in the representation

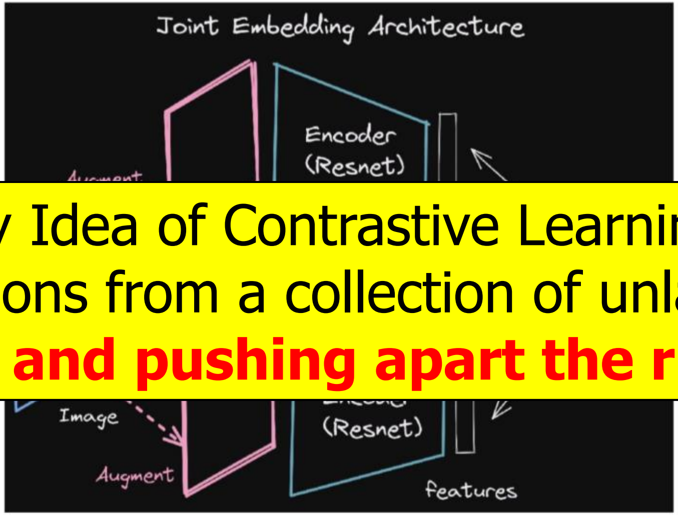


What is to stop the representation from collapsing to $\boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{0}\ \forall \boldsymbol{x}$?

To prevent this, push different images to have different representations

Background: Contrastive Unsupervised Learning

Pull two views of the same image together in the representation



Key Idea of Contrastive Learning:
Train representations from a collection of unlabeled images by
pulling together and pushing apart the right image pairs

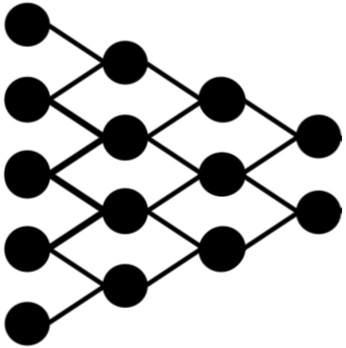
What is to stop the representation from collapsing to $\boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{0}\ \forall \boldsymbol{x}$?

To prevent this, push different images to have different representations

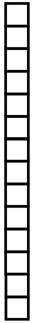
What is a Good Visual Representation for Robotics?

Image $\mathbf{o}$

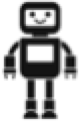
$\phi(\cdot)$



$\phi(\mathbf{o})$



$\pi(\cdot)$



$a = \pi(\phi(o))$

“rotate gripper +4°”

representation encoder

action “policy”

Overview: The Reinforcement Learning (RL) Formalism

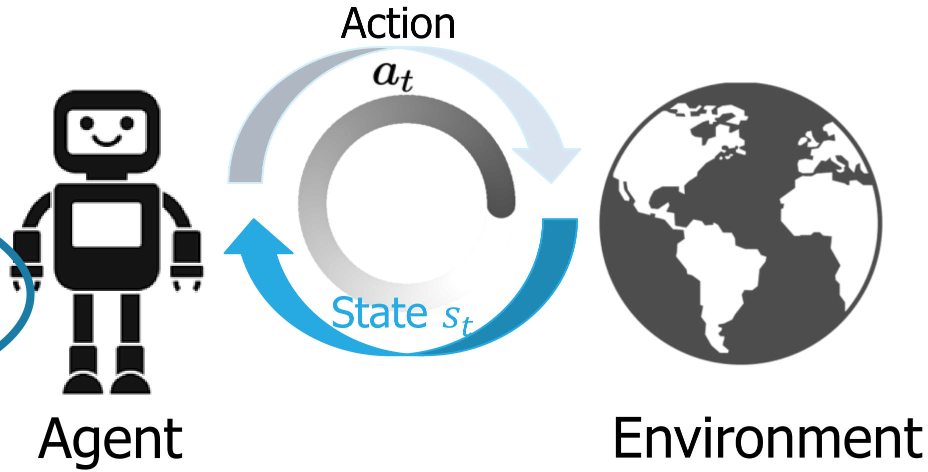
States $s \in S$

Actions $a \in A$

Transition function $P(s'|s,a)$

Task Reward function $r(s,a,s')$

unknown



Agent's objective: maximize the discounted sum of "reward" over time by executing a good action sequence $a_1, a_2, \ldots,$

$$\max_{\pi} \; R(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1})\right]$$

Task-Conditioned "Universal" Value Functions

- **Optimal Value Function of A State** … Conditioned On A Task g [Schaul et al 2015]

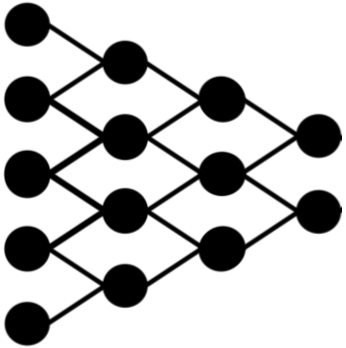
$$V^*(s_0; g) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}; g)\right]$$

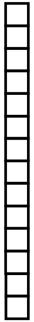
"How good is this state for completing the task g (if acting optimally)"?

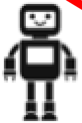
- V Value functions are a useful abstraction:
  - Can guide policy improvement such as through RL
  - Well-known "Bellman equation" constraints connecting V values at consecutive steps, permitting easy dynamic programming-style learning.
  - Don't require known actions

Key Idea: Representations as Value Functions

Image $\mathbf{o}$ $\quad$ $\phi(\cdot)$ $\quad$ $\phi(\mathbf{o})$ $\quad$ $\pi(\cdot)$ $\quad$ $a = \pi(\phi(o))$







"rotate gripper +4°"

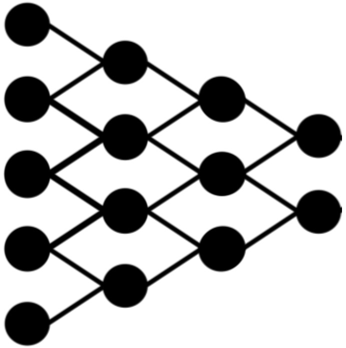
representation encoder

policy

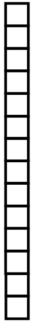
Key Idea: Representations as Value Functions

Image $\mathbf{o}$

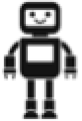
$\boldsymbol{\phi}(\cdot)$



$\boldsymbol{\phi}(\mathbf{o})$



$\boldsymbol{V}^*(\cdot\,; \boldsymbol{g})$



$\boldsymbol{V}^*(\boldsymbol{o}; \boldsymbol{g})$

0.8

task specification g

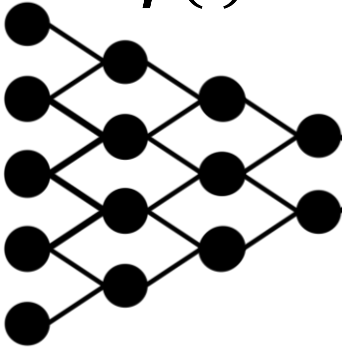
or

"squeeze the brush dry"

Key Idea: Representations as Value Functions

Image $\mathbf{o}$

$\boldsymbol{\phi}(\cdot)$

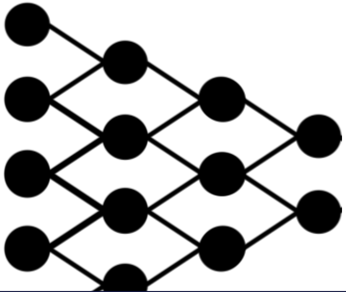


$\boldsymbol{\phi}(\mathbf{o})$

$\boldsymbol{V^*}(\cdot\,; \boldsymbol{g})$

$\boldsymbol{V^*}(\boldsymbol{o}; \boldsymbol{g})$

Goal image $\mathbf{g}$



$\boldsymbol{\phi}(\mathbf{g})$

distance

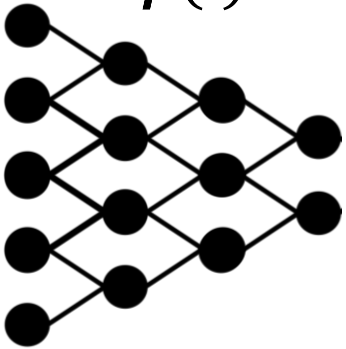
0.8

Representation $\boldsymbol{\phi}(\cdot)$ should be rich enough so that it easily expresses V^*

Key Idea: Representations as Value Functions

Image $\mathbf{o}$

$\boldsymbol{\phi}(\cdot)$



$\boldsymbol{\phi}(\mathbf{o})$

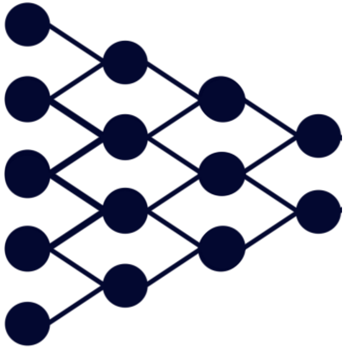
$V^*(\cdot\,; \boldsymbol{g})$ $\quad$ $V^*(\boldsymbol{o}; \boldsymbol{g})$

distance $\quad$ 0.8

Goal language $\mathbf{g}$

"squeeze the brush dry"

$\boldsymbol{\phi}^l(\cdot)$



$\boldsymbol{\phi}^l(\mathbf{g})$

Train ϕ, ϕ^l through training the value function:

$$V^*\left(\boldsymbol{\phi}(\boldsymbol{o}), \boldsymbol{\phi}^l(\boldsymbol{g})\right) = \boldsymbol{d}(\boldsymbol{\phi}(\boldsymbol{o}), \boldsymbol{\phi}^l(\boldsymbol{g}))$$

Pre-Train on Pre-Recorded In-the-Wild Human Videos

Ego4D dataset (Grauman 2022)

EpicKitchens (Damen 2021)

Human videos are goal-directed, and abundant!

- Treat the final frame of any video as the goal
- Reward function? r = 1 for last step of video, 0 elsewhere.
- Actions not available, but no problem: we only care for $V^*(s)$

Offline RL Value Function Training Objective

Pulls every frame o preceding g to have high $V^*(o\,;g)$

Encourages $V^*(\cdot\,;g) = ||\phi(\cdot) - \phi(g)||_2$ to become a valid (“Bellman-consistent”) value function.

$$\mathbb{E}_{p(g)}\left[(1-\gamma)\mathbb{E}_{\mu_0(o;g)}\left[\|\phi(o)-\phi(g)\|_2\right] + \log \mathbb{E}_{(o,o';g)\sim D}\left[\exp\left(\|\phi(o)-\phi(g)\|_2 - \tilde{\delta}_g(o) - \gamma\,\|\phi(o')-\phi(g)\|_2\right)\right]\right]$$

All frames leading up to the goal should be close to goal

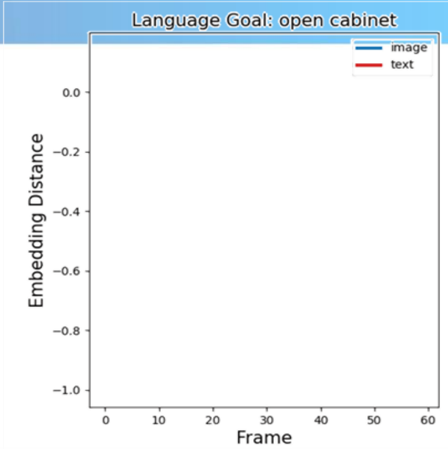
– <u>pulls frames together</u>

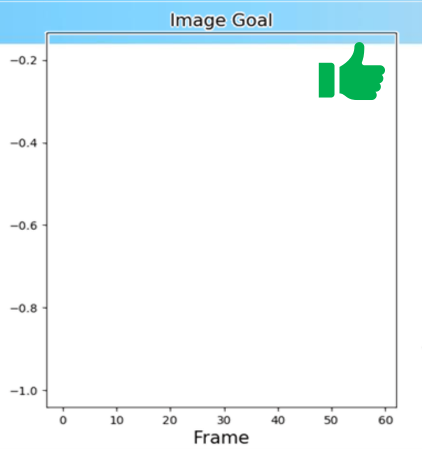
Consecutive frames should be at different distances from the goal

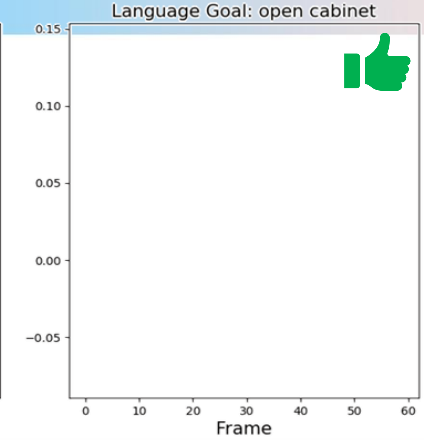
– <u>pushes frames apart</u>

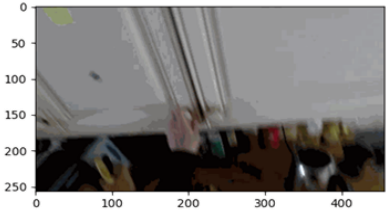
Training representations as value functions with offline RL generates a new control-aware contrastive learning objective!

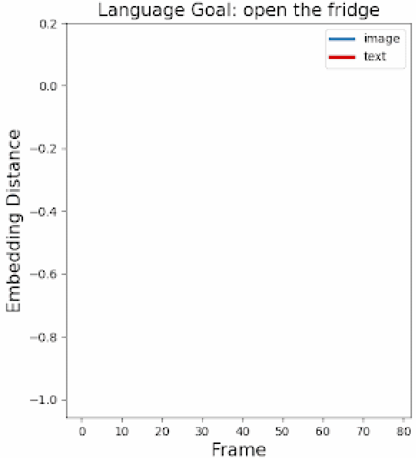
Results: Image ↔ Language-Goal Distance $d(\phi(o), \phi^l(g))$

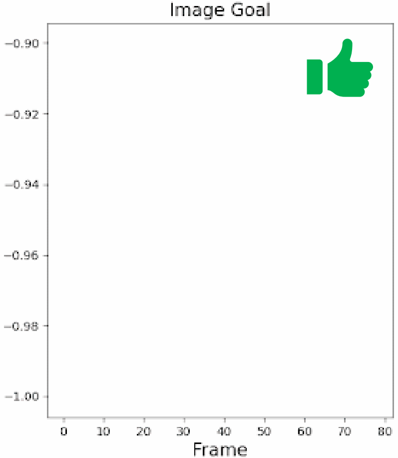


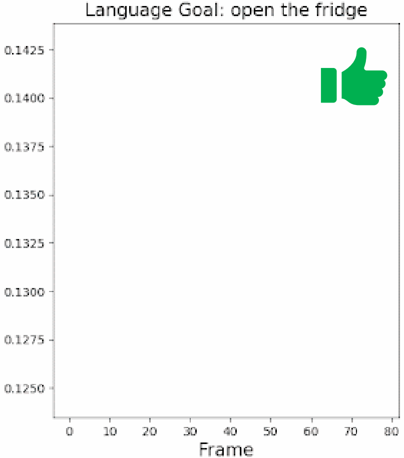


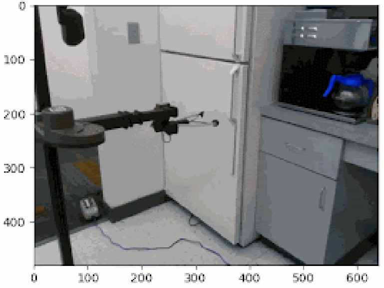




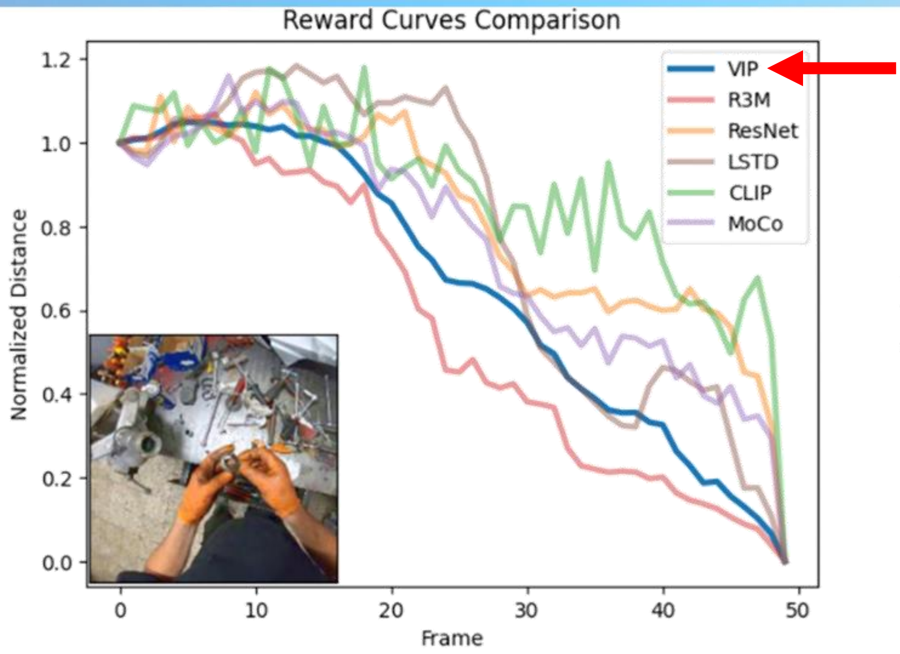


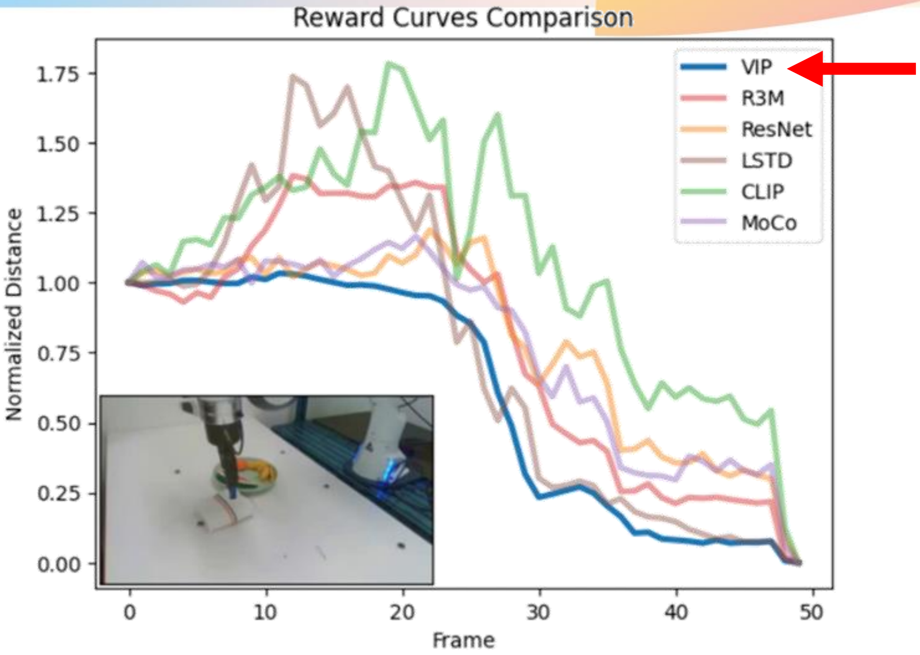






Results: Image ↔ Image-Goal Distance $d(\phi(o), \phi^l(g))$





On demo data, our representations predict smooth goal-conditioned V* on human and robot videos.

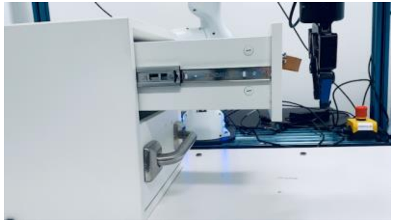
What Can We Do With $\phi(\cdot)$ and $\phi^l(\cdot)$?

- **Use as representations for robot learning:**
  - Training robot policies on image representation with:
    - behavior cloning
    - language-conditioned behavior cloning [Lynch '20]
- **Use as dense reward functions to guide reinforcement policy learning:**
  - $R(o, a, o'; g) = V^*(o', g) - V^*(o, g) = ||\phi(o') - \phi(g)||_2 - ||\phi(o) - \phi(g)||_2$
    - offline RL (reward-weighted regression [Peters '07]) for policy learning from noisy demos
    - online policy improvement with trajectory optimization and RL (natural policy gradient [Kakade '01])

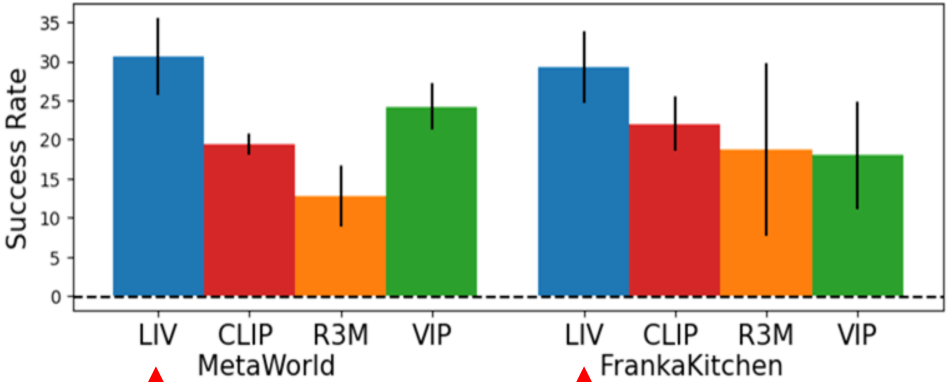
Quantitative Results Summary

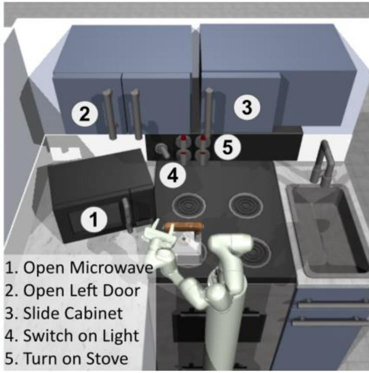
Results: Real-World BC / Offline RL From 20 Demos

Environment	VIP-RWR	*Pre-Trained* VIP-BC	R3M-RWR	R3M-BC	Scratch-BC	*In-Domain* VIP-RWR	VIP-BC
CloseDrawer	**100** ± 0	50 ± 50	80 ± 40	10 ± 30	30 ± 46	0 ± 0	0* ± 0
PushBottle	**90** ± 30	50 ± 50	70 ± 46	50 ± 50	40± 48	0* ± 0	0* ± 0
PlaceMelon	**60** ± 48	10 ± 30	0 ± 0	0 ± 0	0 ± 0	0* ± 0	0* ± 0
FoldTowel	**90** ± 30	20 ± 40	0 ± 0	0 ± 0	0 ± 0	0* ± 0	0* ± 0



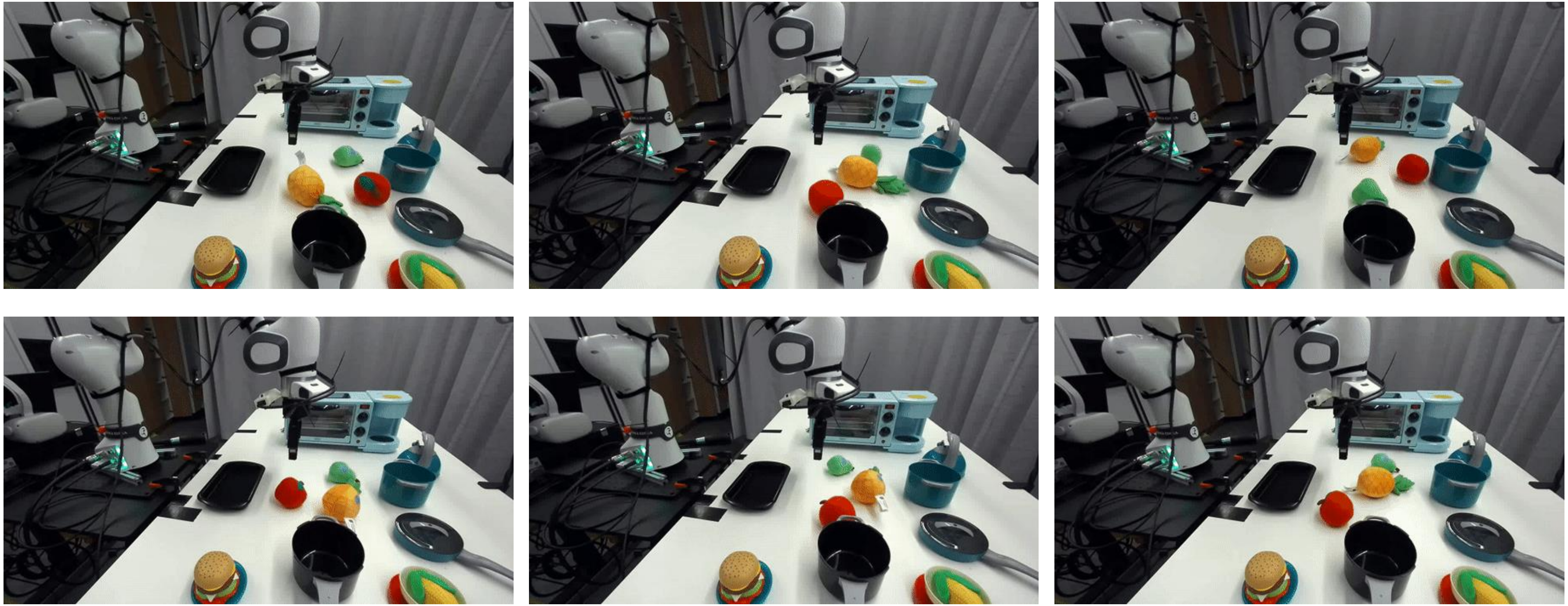
Results: Language-Conditioned Behavior Cloning



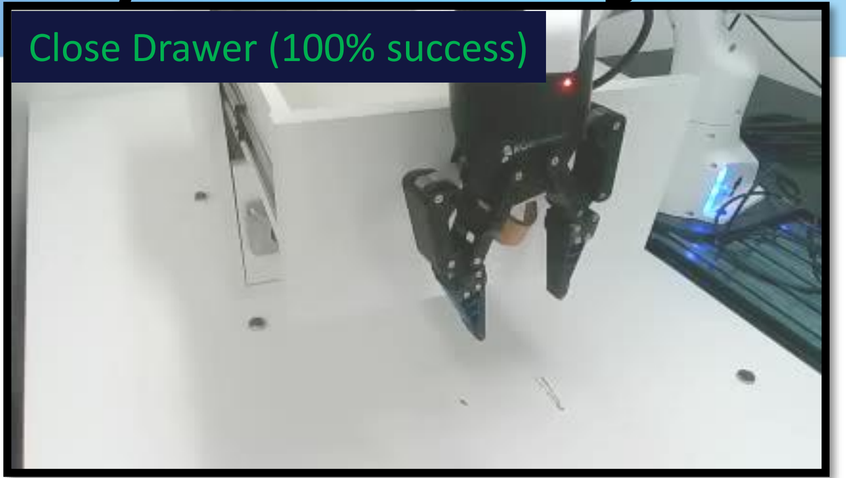


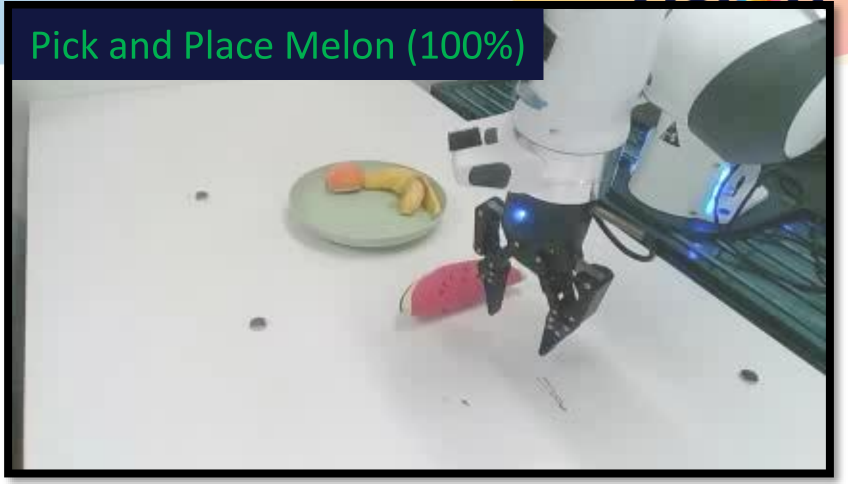
Noisy demos → Language Goal-Based Policies

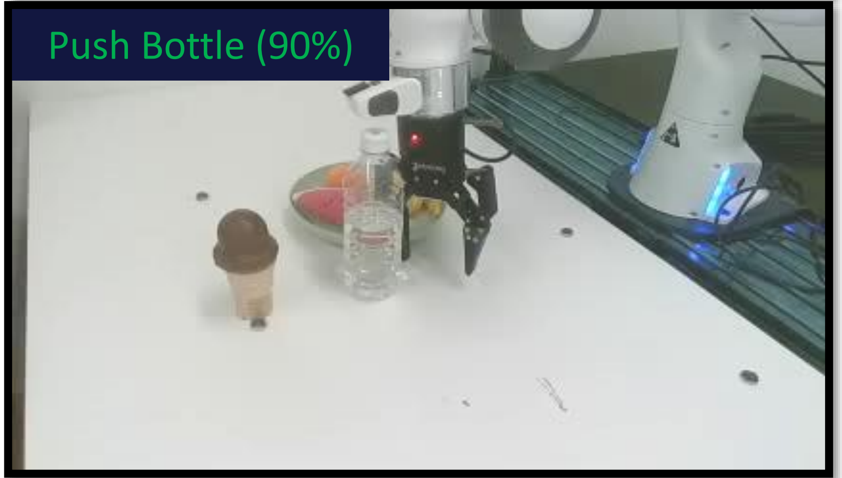
Goal: “Place the pineapple in the pot”

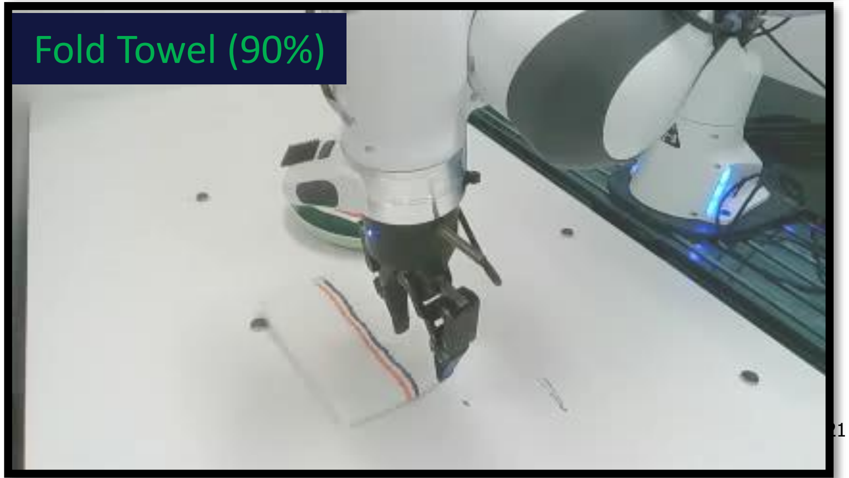


Noisy demos → Image Goal-Based Policies









Takeaways

For a robot to make decisions about good actions, in what format should it internally represent the images from its camera stream?

- Modern visual representations leverage deep neural networks self-supervised from large unlabeled datasets of images, largely focus on visual recognition use cases.
- No explicitly **robot-focused** representations before the work presented here.
- **Training representations as goal-conditioned “universal value functions”**: a powerful new way to learn *control-aware* vision, language, (and other?) representations.

Resources

The work presented here is covered the following papers:

- Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, Dinesh Jayaraman. LIV: Language-Image Representations and Rewards for Robotic Control. ICML 2023.

- Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, Amy Zhang. VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training. ICLR 2023.

For further reading on self-supervised representations:

- Balestriero et al, A Cookbook of Self-Supervised Learning. arXiv 2023.