Learning from data (experience) without labels (rewards, optimal actions).

Self-Supervised Agents:
Exploring and Learning with Minimal Feedback

Benjamin Eysenbach

Assistant Professor of Computer Science

May 1, 2025

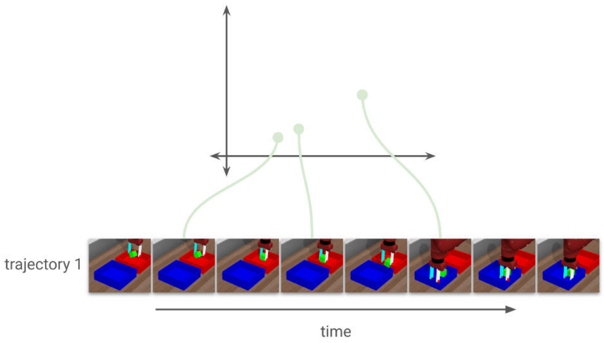
PRINCETON UNIVERSITY

A thought experiment 🤔

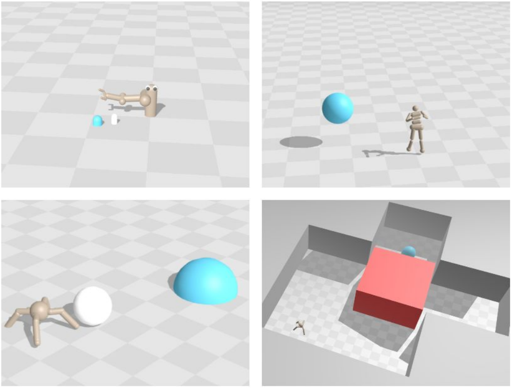
- Can a generative image model generate images different from those seen during training (e.g., a horse on the moon)?
- Can an LLM generate sentences different from those seen during training (e.g., write a poem about Tehran in the style of "twas the night before Christmas")?

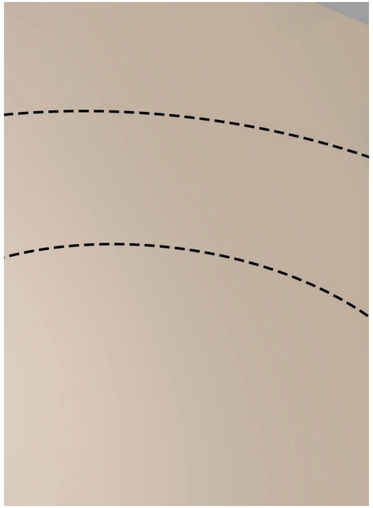
👆What is the right analogy for reinforcement learning (RL)?

Outline for today



1/ Self-supervised RL





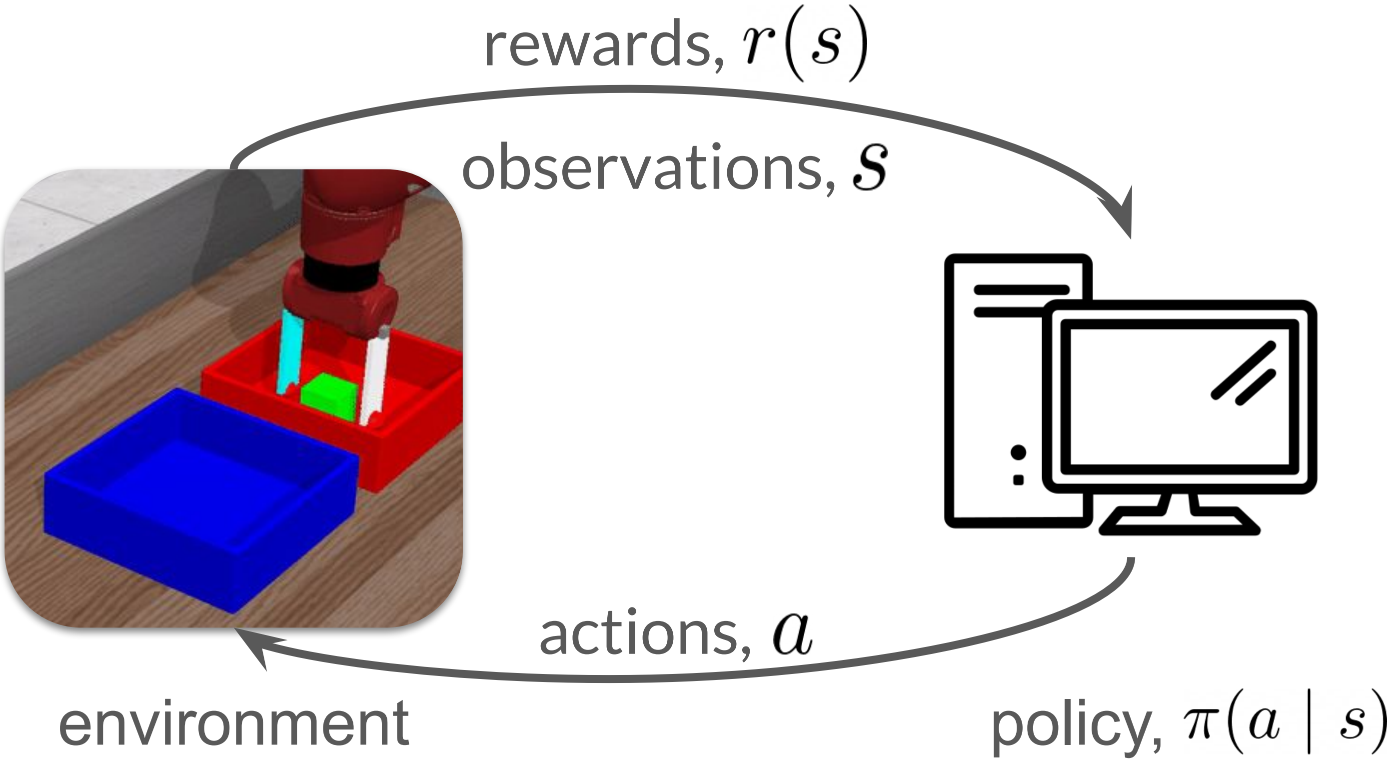
2/ Steps towards RL agents that can learn to do anything with minimal feedback

Papers highlighted:

1. BE, et al. *Contrastive learning as goal-conditioned reinforcement learning*. NeurIPS, 2022.
2. Bortkiewicz, et al. *Accelerating Goal-Conditioned RL Algorithms and Research*. ICLR, 2025.
3. Liu, Tang, BE. *A Single Goal is All You Need: Skills and Exploration Emerge from Contrastive RL without Rewards, Demonstrations, or Subgoals*. ICLR, 2025.
4. Myers, Ji, BE. *Horizon Generalization in Reinforcement Learning.* ICLR, 2025.

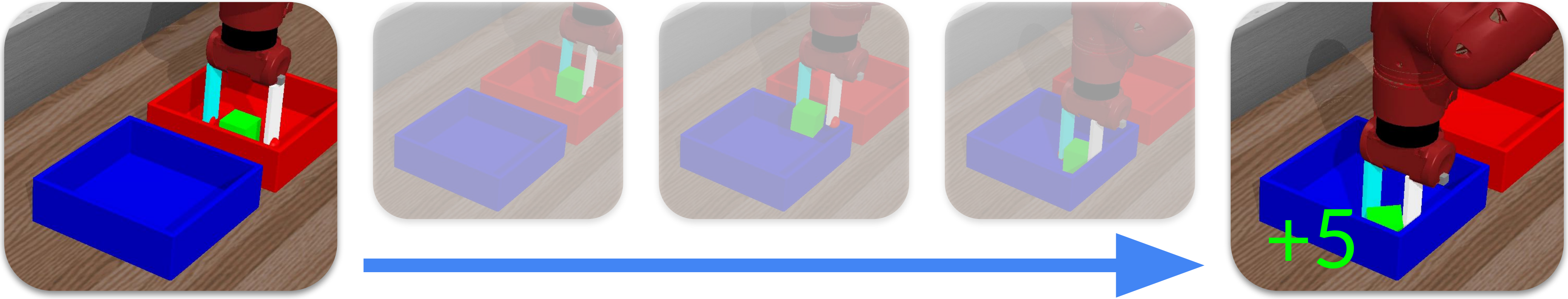
Background: What is RL?

Defining the reinforcement learning problem.



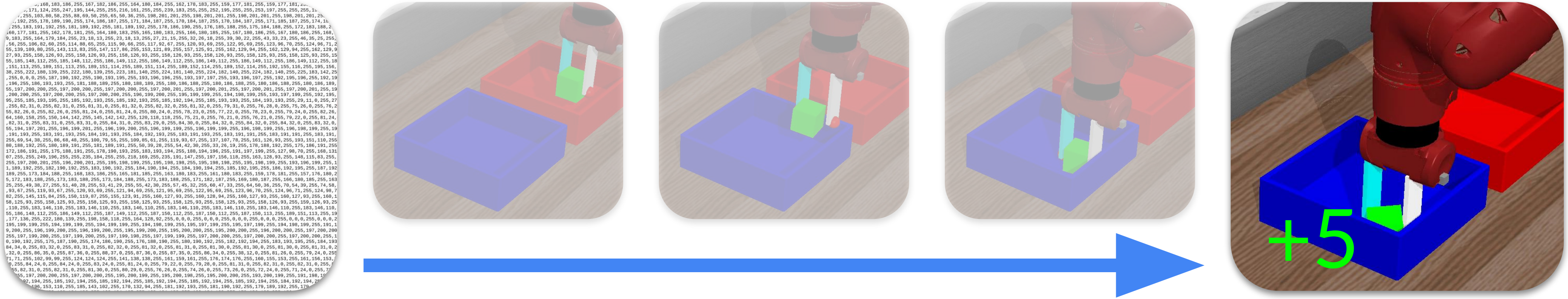
RL is hard because of limited feedback (rewards).

1 Only get feedback many steps into the future.



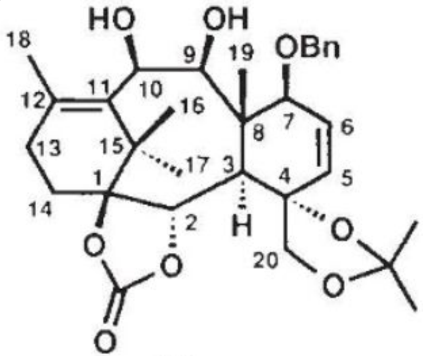
RL is hard because of limited feedback (rewards).

1. Only get feedback many steps into the future.
2. Limited feedback makes it challenging to learn from high-dimensional data.

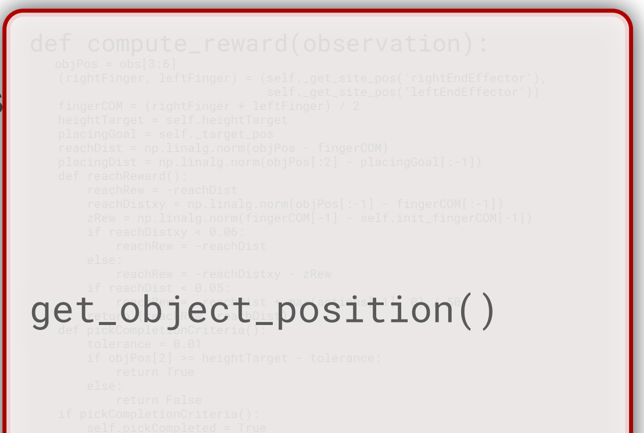


RL is hard because of limited feedback (rewards).

1. Only get feedback many steps into the future.
2. Limited feedback makes it challenging to learn from high-dimensional data.
3. Practically, rewards and optimal actions are hard to s



13



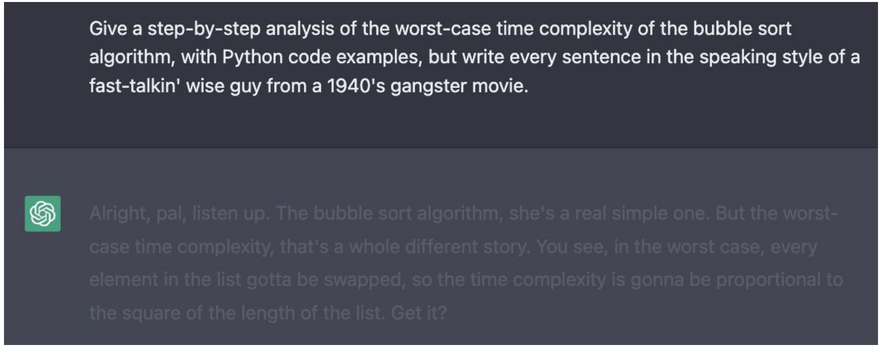
How can we get the benefits of RL without the challenge of reward engineering?

Background: Self-Supervised Learning

Self-supervised learning in other areas of ML: Learning from high-dimensional data with limited feedback.

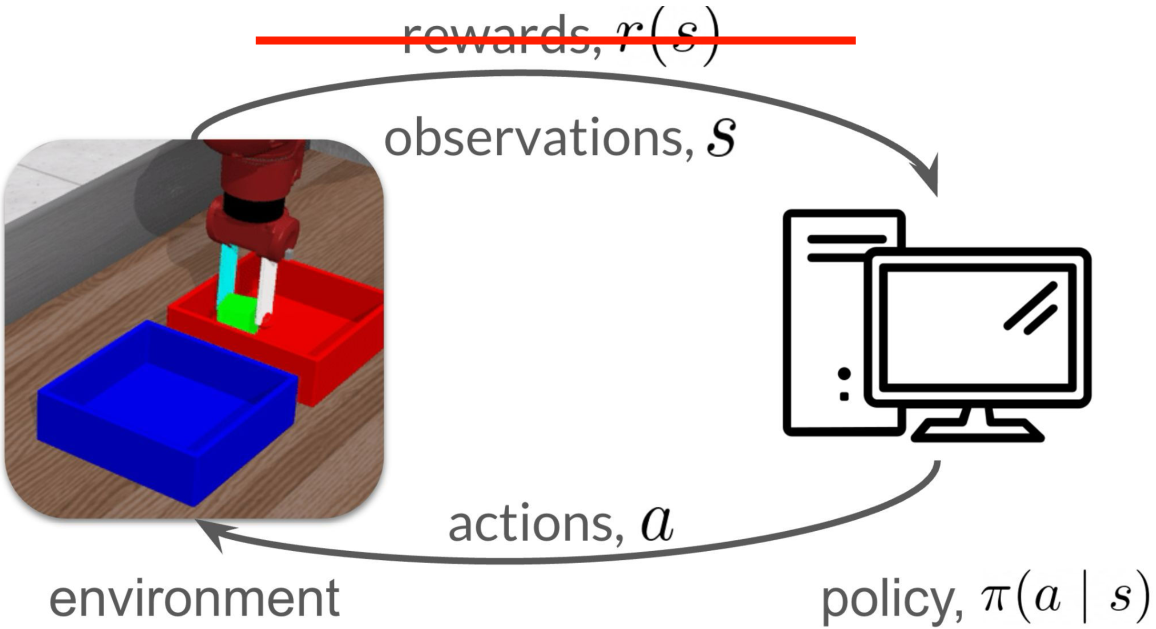
Computer Vision

NLP



Self-Supervised **Reinforcement** Learning

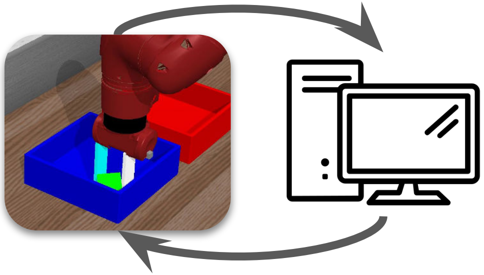
Self-supervised RL



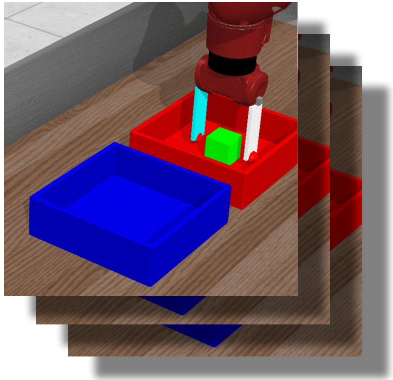
Goal-Conditioned RL: one type of self-supervised RL

Inputs:

a) Online setting OR b) Offline setting



data: videos + actions



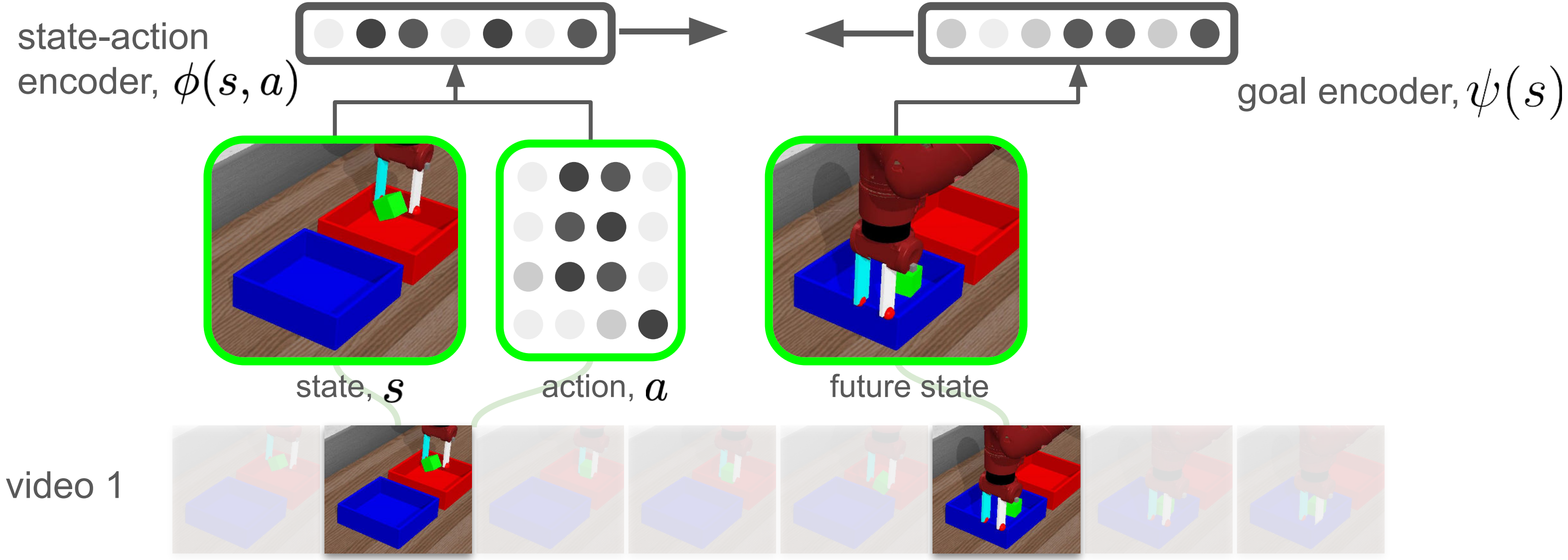
Output:

Goal-conditioned policy

$$\pi(a \mid s, g)$$

No reward or action labels required!

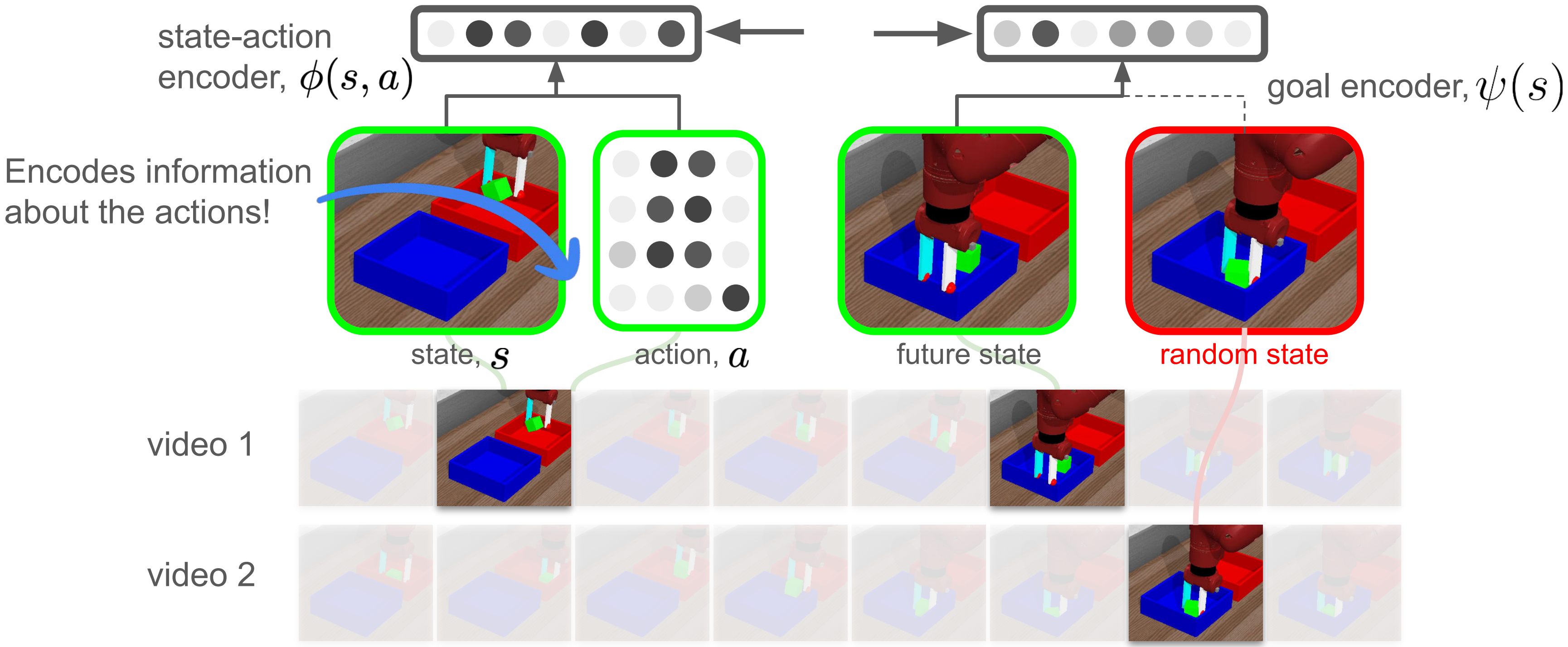
Learn representations via temporal contrastive learning



Much prior work in other domains: E.g., [Sermanet et al, 2018; van den Oord et al, 2018]

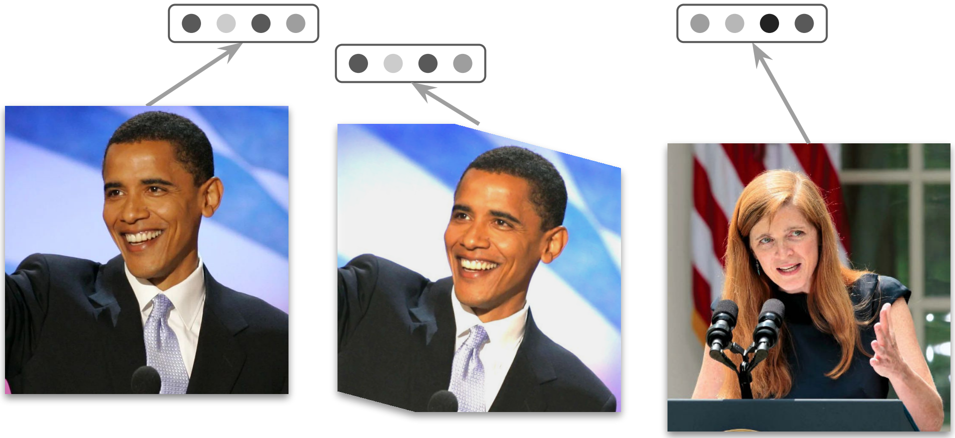
BE et al. *Contrastive learning as goal-conditioned reinforcement learning*, NeurIPS, 2022.

Learn representations via temporal contrastive learning



Much prior work in other domains: E.g., [Sermanet et al, 2018; van den Oord et al, 2018]

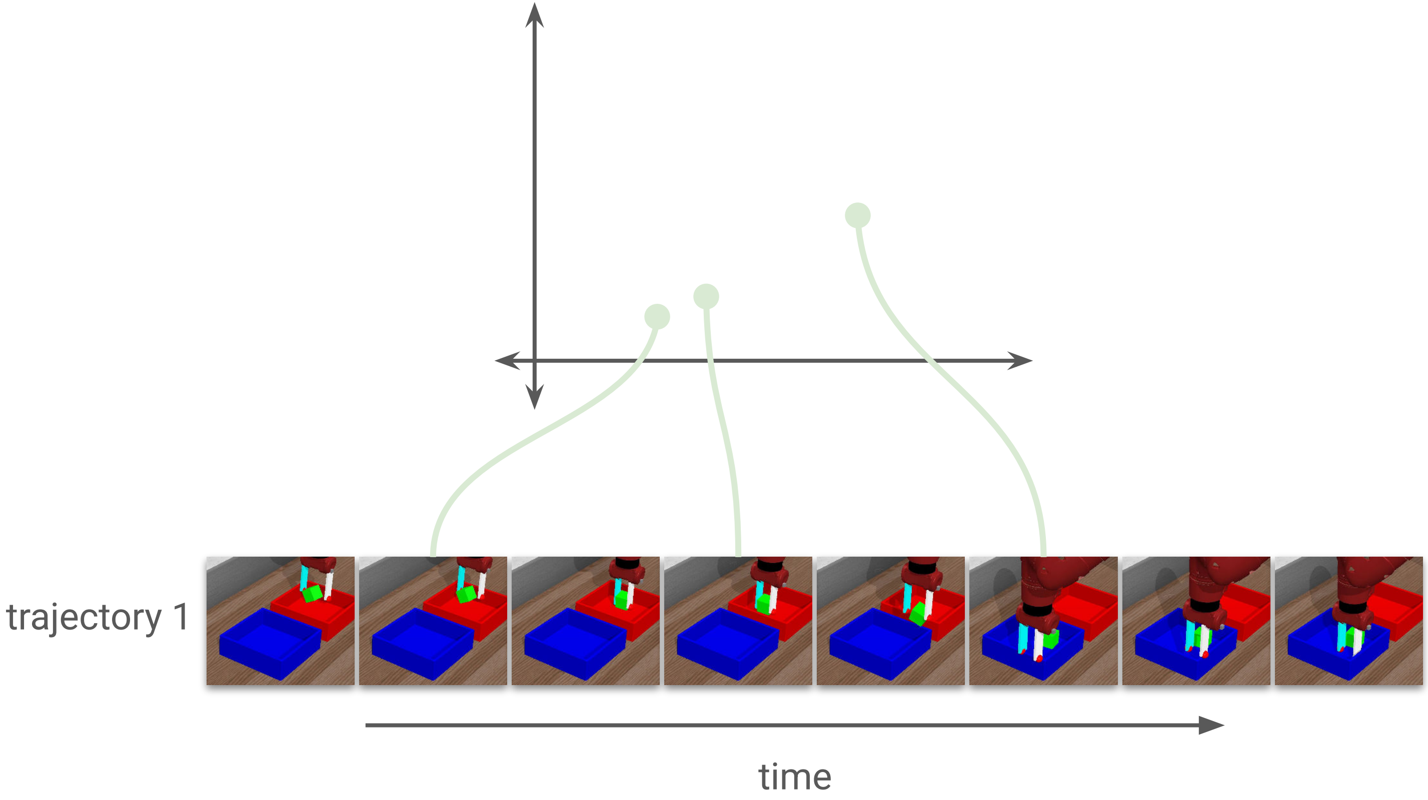
Compare with contrastive learning for vision



Augmentation: $p(x' \mid x)$

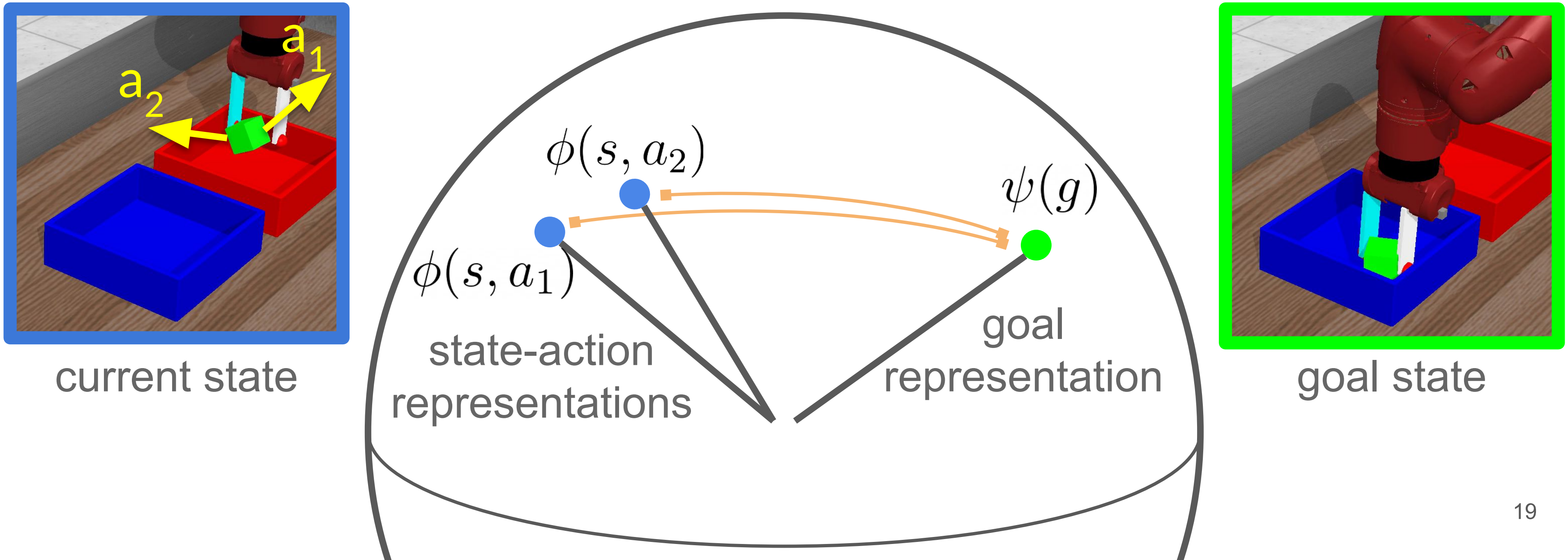
Schroff, et al. A Unified Embedding for Face Recognition and Clustering, 2015.
Chen, et al. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020

Intuition: representations encode temporal information

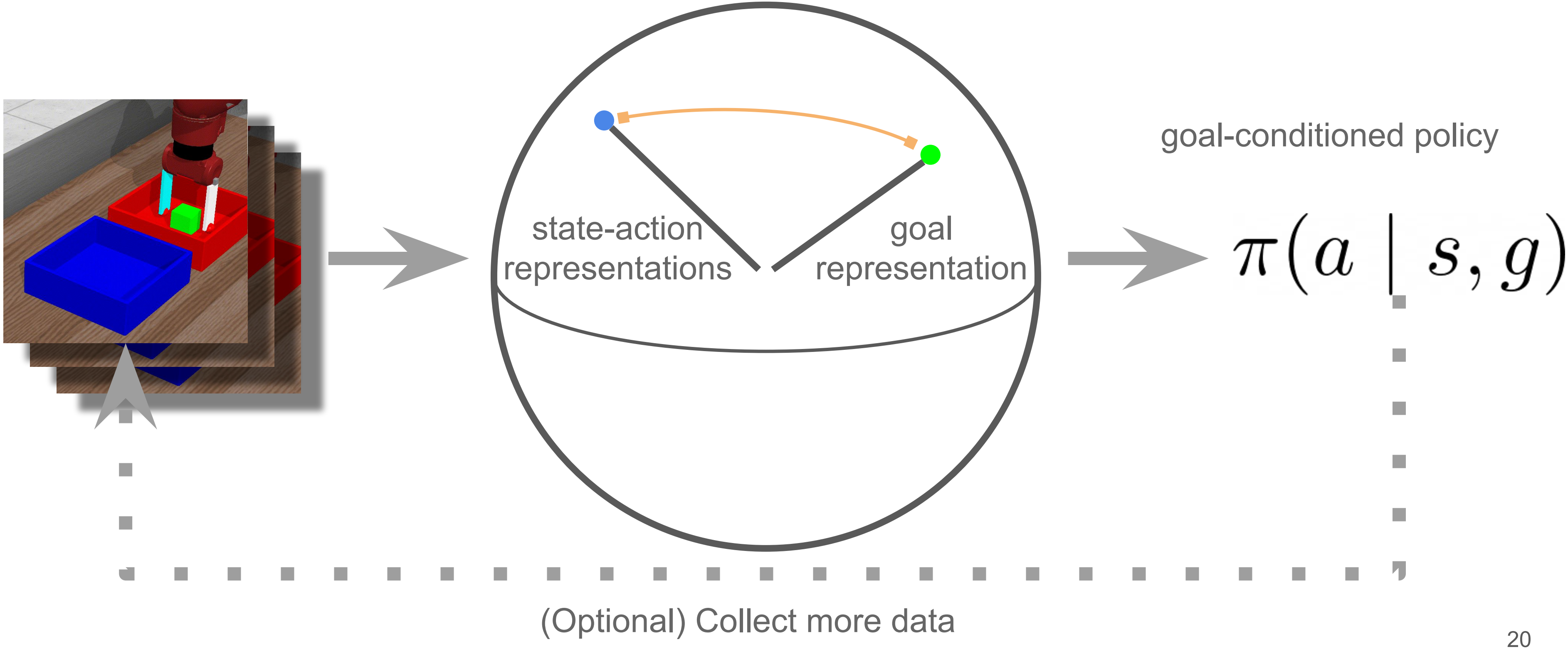


Use representations to extract goal-reaching skills.

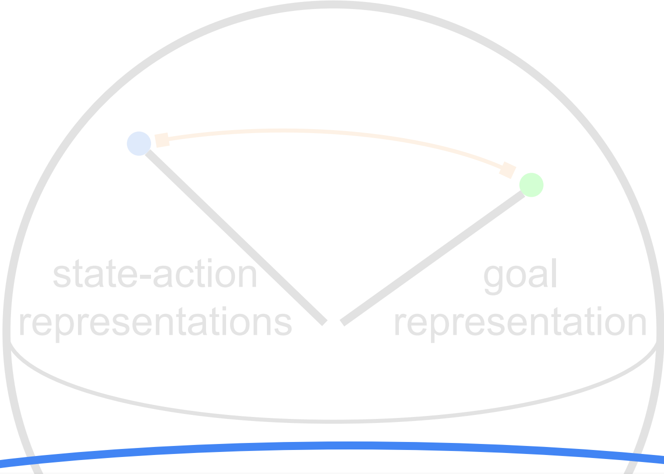
$$\left\{\pi(a \mid s, g) = \arg\max_a \phi(s,a)^T \psi(g), \cdots\right\}$$



A complete method for learning (goal-reaching) skills from data.



Our complete method for learning (goal-reaching) skills from data.



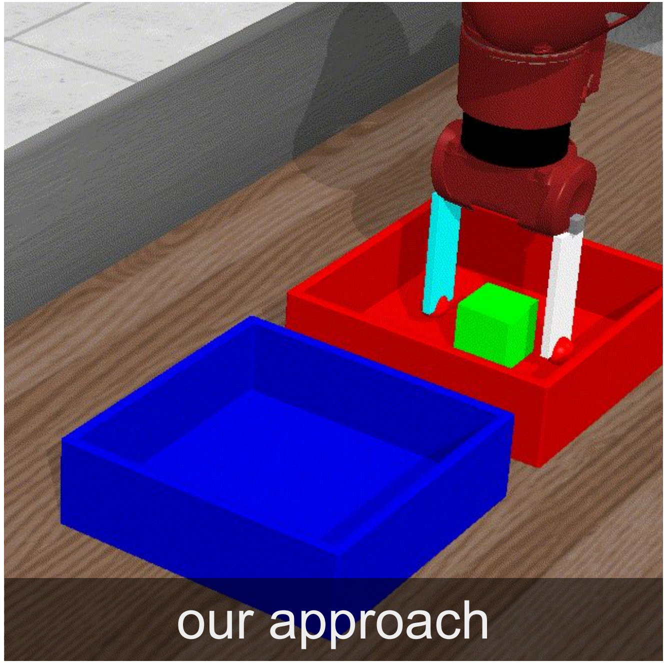
goal-conditioned policy

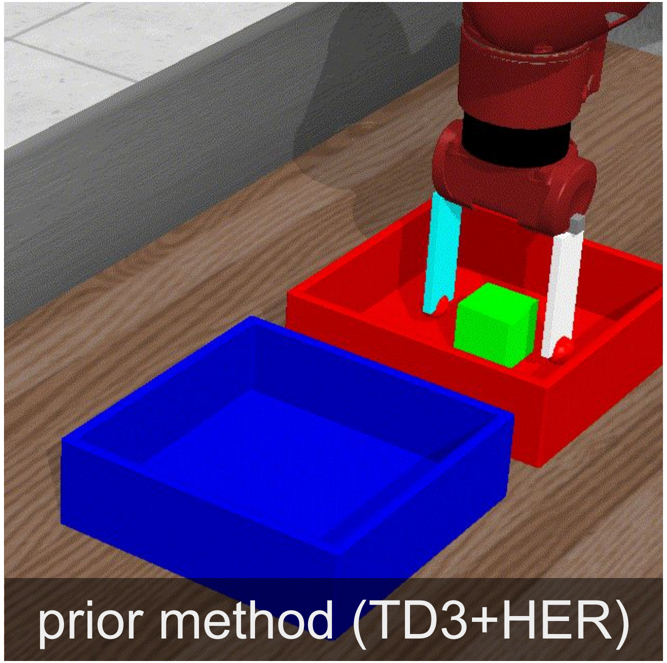
$$\pi(a \mid s, g)$$

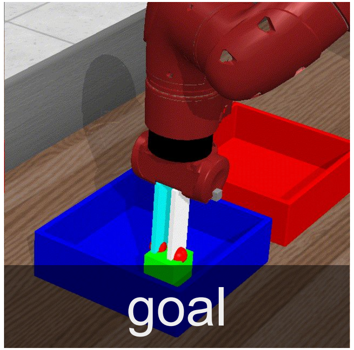
self-supervised RL RL goal-conditioned RL

[Newell et al, 1961; Laird et al, 1987; Kaelbling, 1993; Andrychowicz et al, 2017.]

0 amount of rewards required







self-supervised RL | RL | goal-conditioned RL

[Newell et al, 1961; Laird et al, 1983; Kaelbling, 1993; Andrychowicz et al, 2017.]

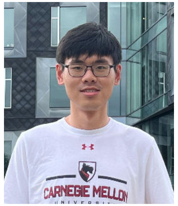
0

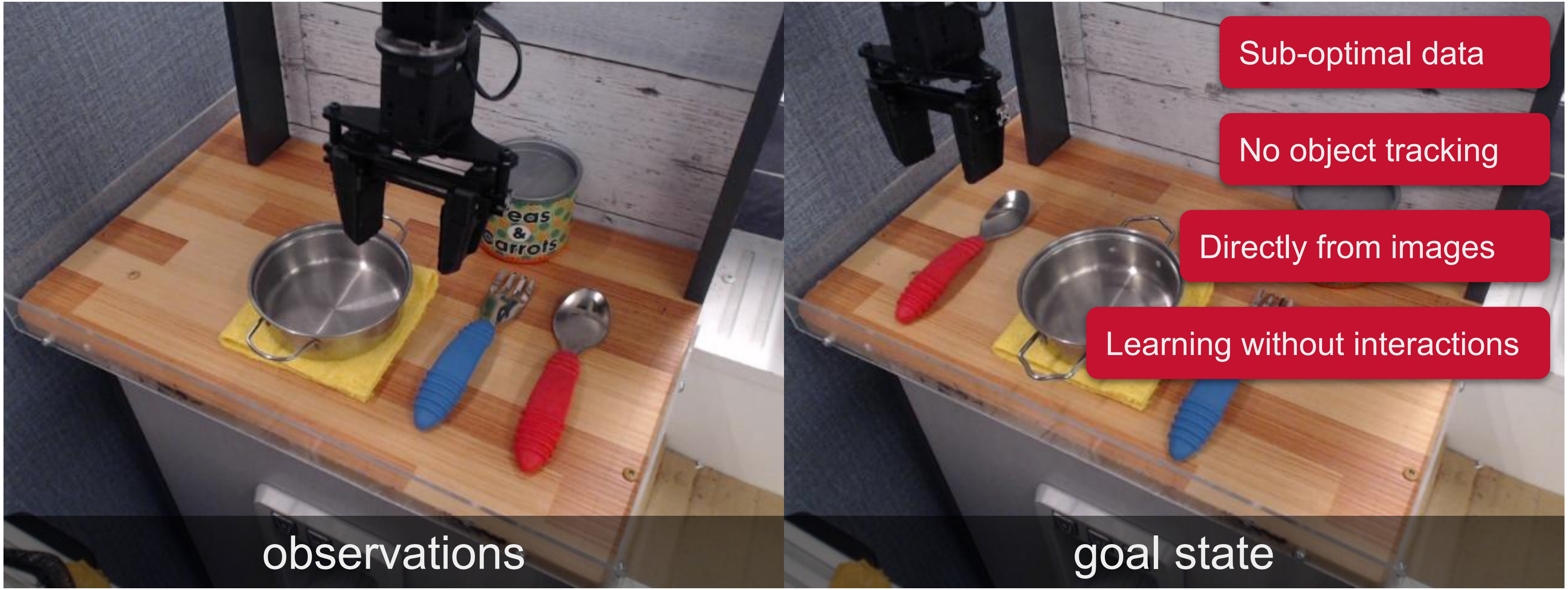
amount of rewards required

A couple real-world applications of self-supervised RL

Evaluating the goal-reaching on a real robot.

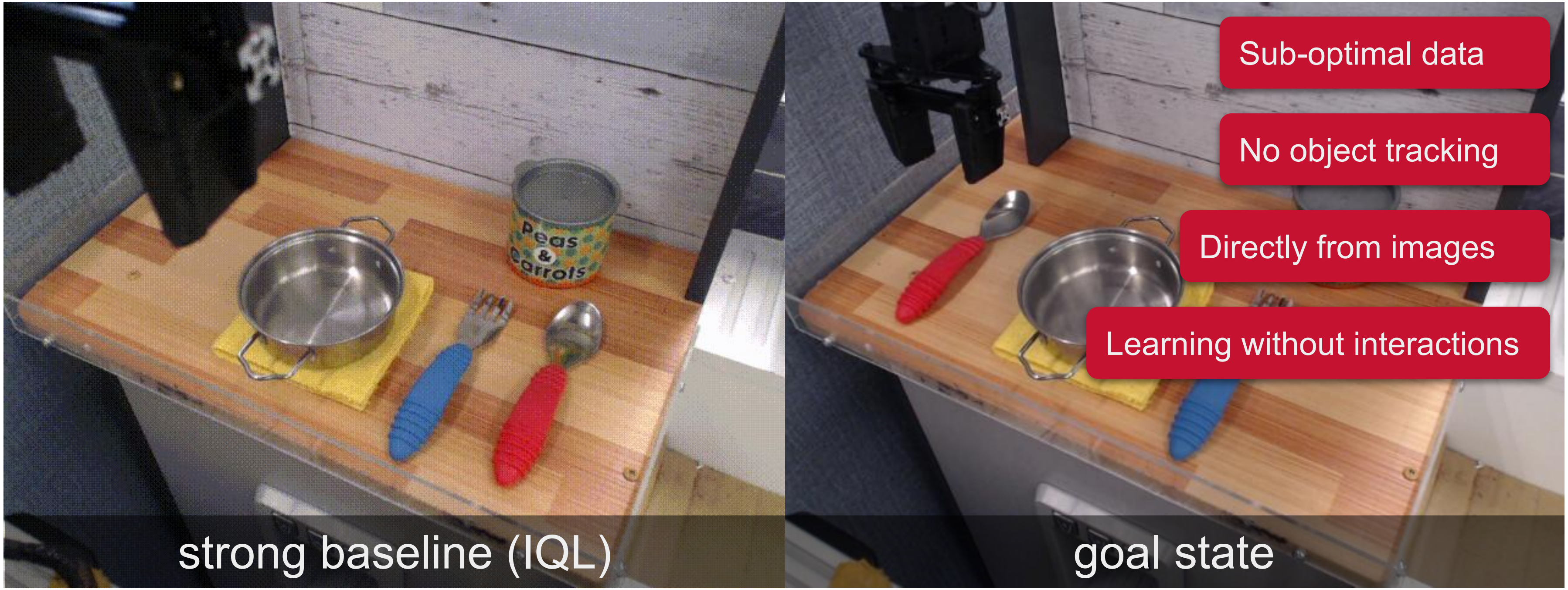
Chongyi Zheng,
Homer Walke





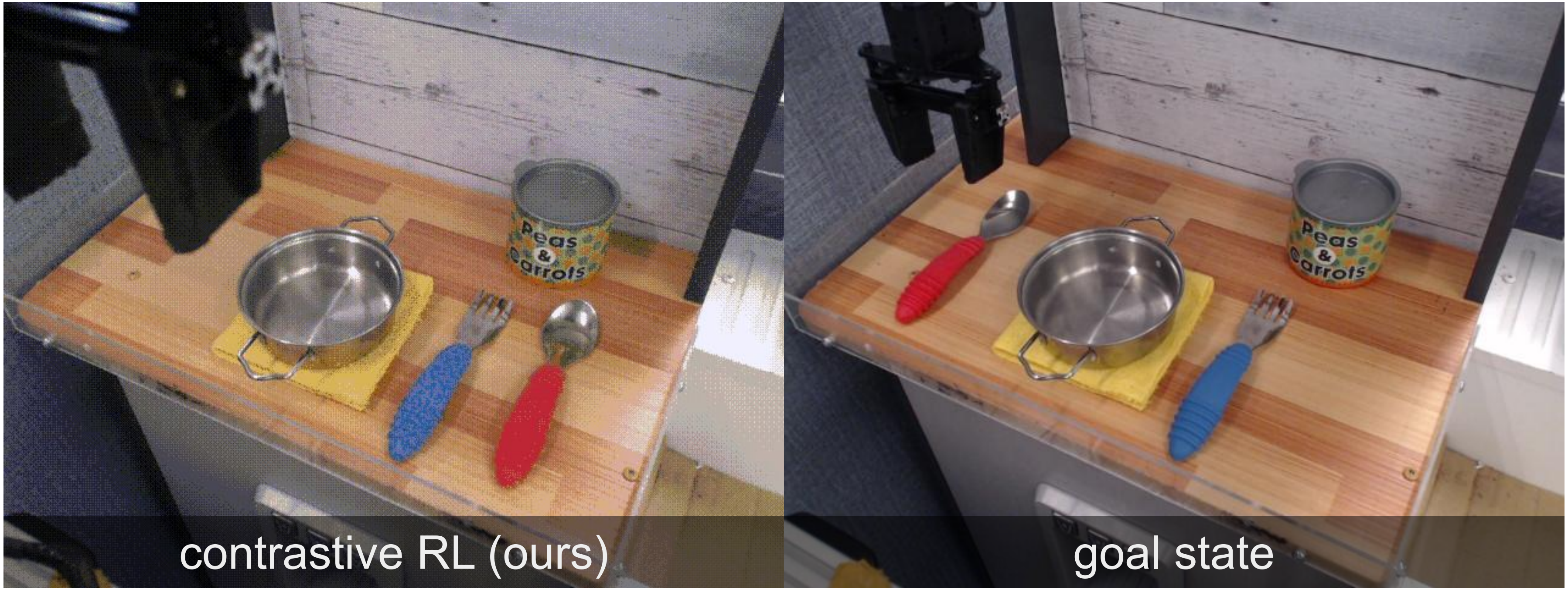
Zheng, et al. *Stabilizing Contrastive RL: Techniques for Robotic Goal Reaching from Offline Data*. ICLR, 2024

Evaluating the (goal-reaching) skills on a real robot.



Zheng, et al. *Stabilizing Contrastive RL: Techniques for Robotic Goal Reaching from Offline Data*. ICLR, 2024

Our goal-conditioned skills solve a real-world robotic task.

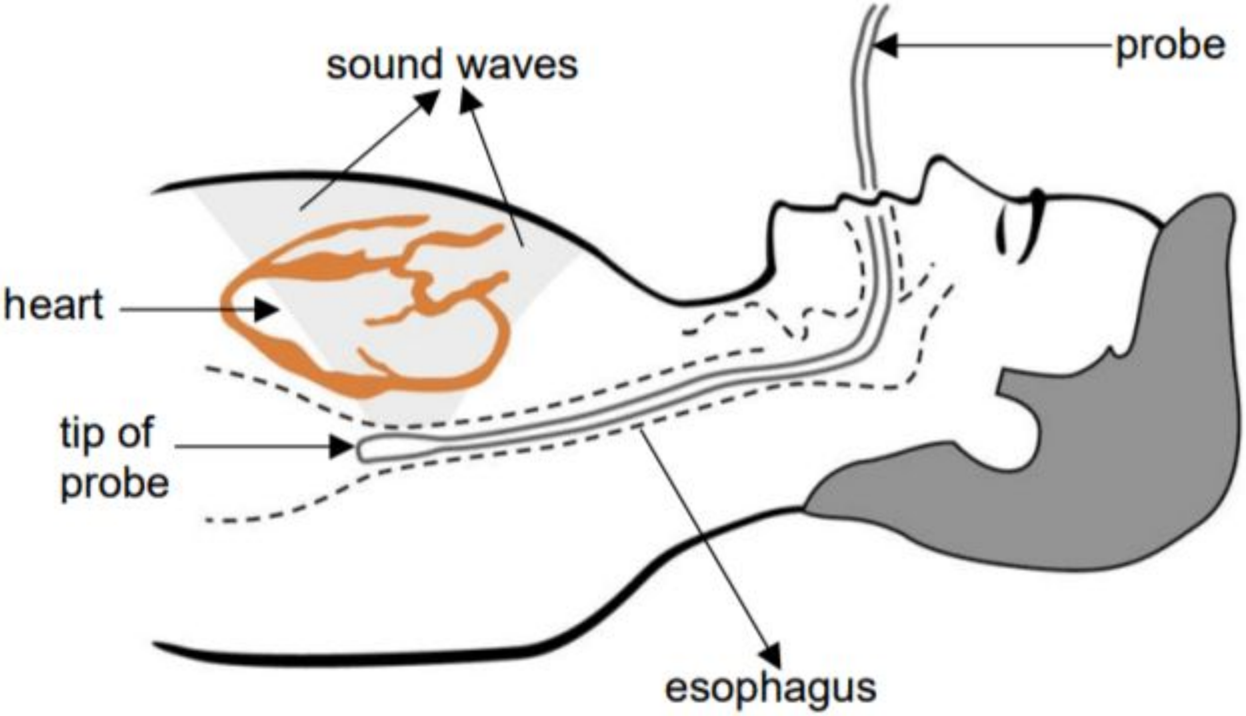


Zheng, et al. *Stabilizing Contrastive RL: Techniques for Robotic Goal Reaching from Offline Data*. ICLR, 2024

Solve new tasks by just "prompting" with a new goal

Zheng, et al. *Stabilizing Contrastive RL: Techniques for Robotic Goal Reaching from Offline Data*. ICLR, 2024

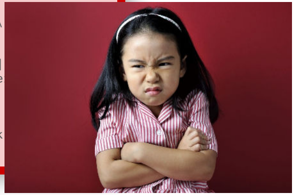
Application to Esophageal Ultrasound

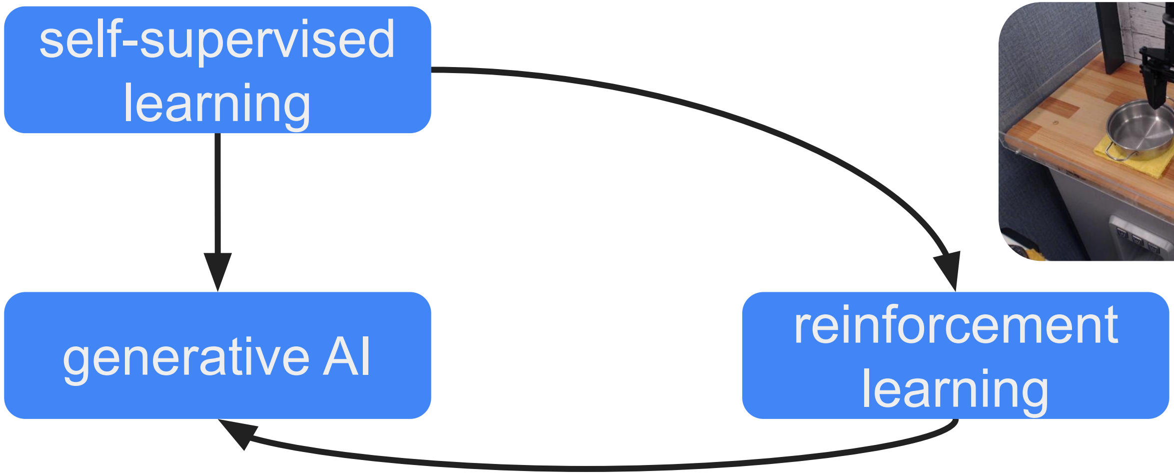


Amadou, Abdoul Aziz, et al. "Goal-conditioned reinforcement learning for ultrasound navigation guidance." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024.

How can we get the benefits of RL without the challenge of reward engineering?

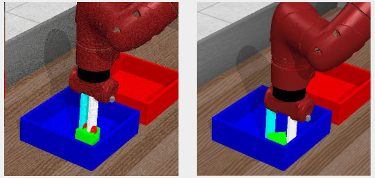
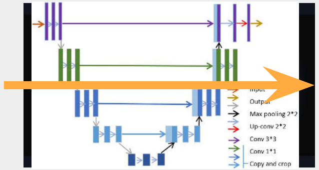
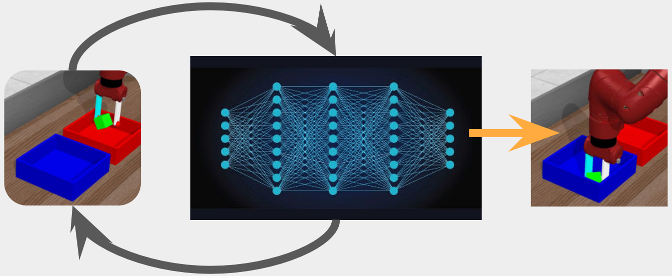
```
def compute_reward(observation):
    objPos = obs[3:6]
    (rightFinger, leftFinger) = (self._get_site_pos('rightEndEffector'),
                                 self._get_site_pos('leftEndEffector'))
    fingerCOM = (rightFinger + leftFinger) / 2
    heightTarget = self.heightTarget
    placingGoal = self._target_pos
    reachDist = np.linalg.norm(objPos - fingerCOM)
    placingDist = np.linalg.norm(objPos[:2] - placingGoal[:-1])
    def reachReward():
        reachRew = -reachDist
        reachDistxy = np.linalg.norm(objPos[:-1] - fingerCOM[:-1])
        zRew = np.linalg.norm(fingerCOM[-1] - self.init_fingerCOM[-1])
        if reachDistxy < 0.06:
            reachRew = -reachDist
        else:
            reachRew = -reachDistxy - zRew
        if reachDist < 0.05:
            reachRew = -reachDist + max(actions[-1], 0) / 50
        return (reachRew, reachDist)
    def pickCompletionCriteria():
        tolerance = 0.01
        if objPos[2] >= heightTarget - tolerance:
            return True
        else:
            return False
    if pickCompletionCriteria():
        self.pickCompleted = True
    def objDropped():
        return objPos[2] < self.objHeight + 0.005 and placingDist \
            > 0.02 and reachDist > 0.02
    def placeCompletionCriteria():
        if abs(objPos[0] - placingGoal[0]) < 0.05 and abs(objPos[1]
              - placingGoal[1]) < 0.05 and objPos[2] < self.objHe
            return True
        else:
            return False
// Yu et al. "Meta-world: A benchmark and evaluation for multi-task
reinforcement learning." CoRL, 2020
```



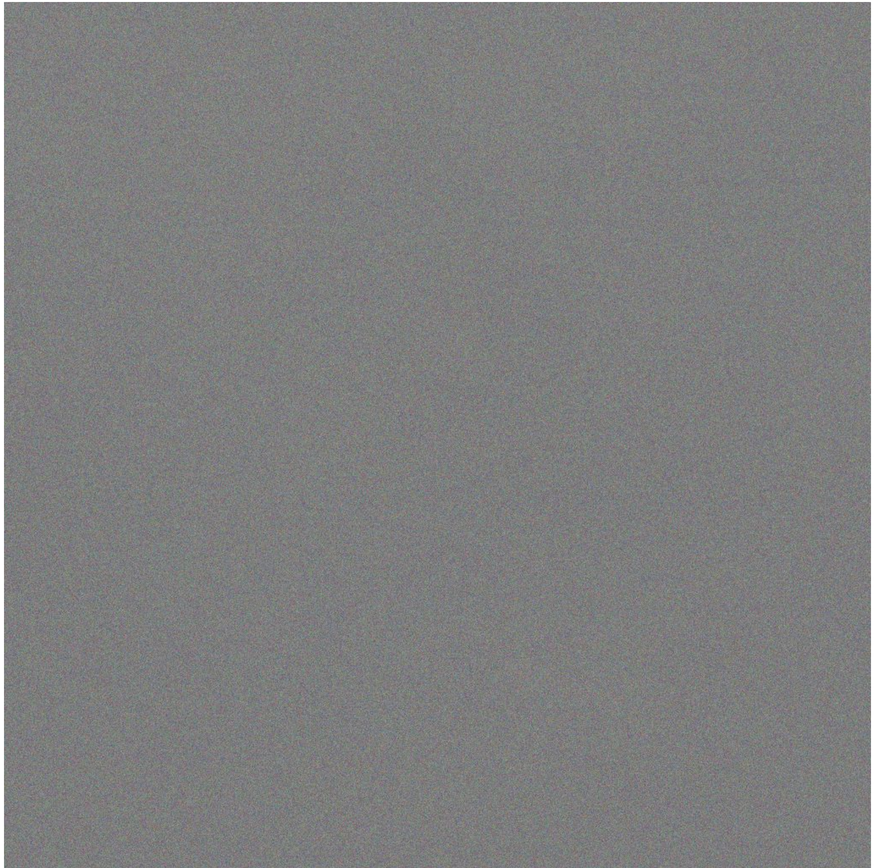


Self-supervised RL is generative AI

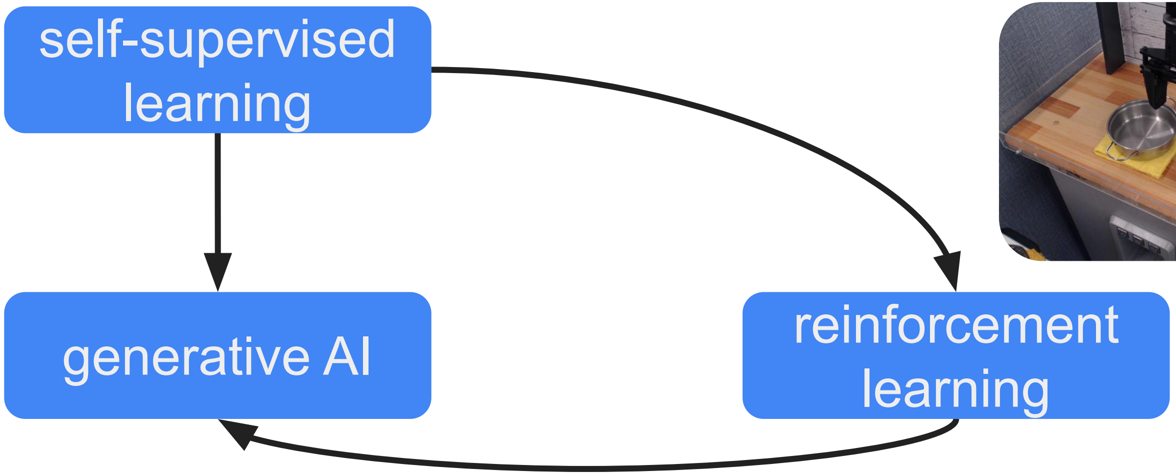
RL is generative modeling

	(Standard) Generative models	**Reinforcement Learning**
Examples		
Sampling	 	

RL is generative AI: "a castle surrounded by mountains"




This is your generative model.
"Build a castle surrounded by mountains"



Self-supervised RL is generative AI

But what about tasks beyond goal reaching?

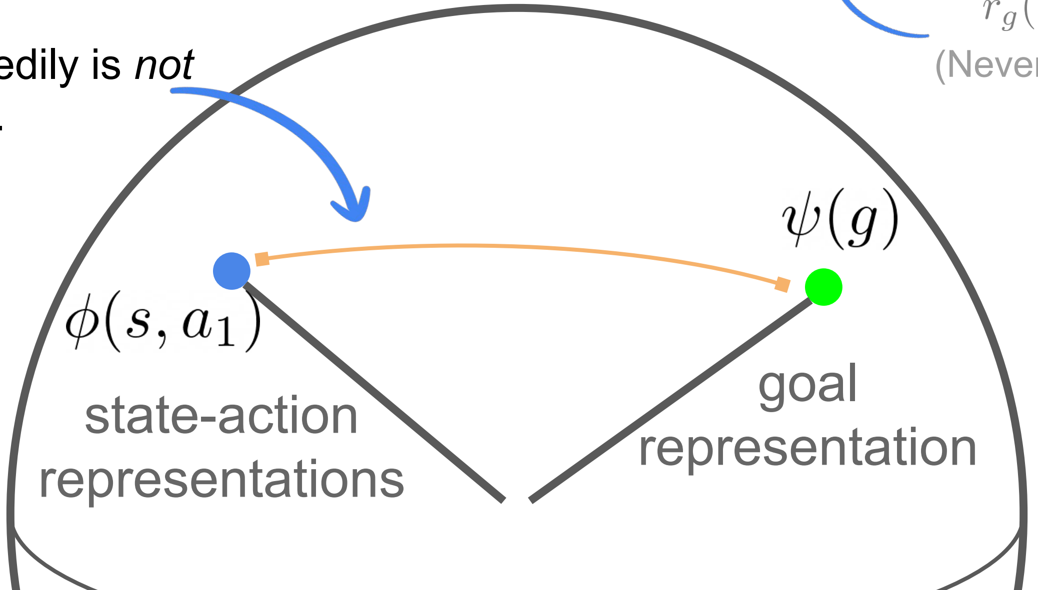
Theorem: The dot-product between the learned representations encodes the future returns, up to a constant.

$$e^{\phi(s,a)^T \psi(g)} = \frac{1-\gamma}{p(g)} \mathrm{E}_\pi \left[\sum_t \gamma^t r_g(s_t, a_t) \right] = Q_g^\beta(s_t, a_t)$$

$r_g(s,a) = p(s' = g \mid s, a)$

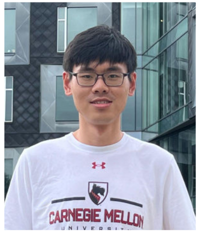
(Never need to compute this!)

Acting greedily is *not* a heuristic.

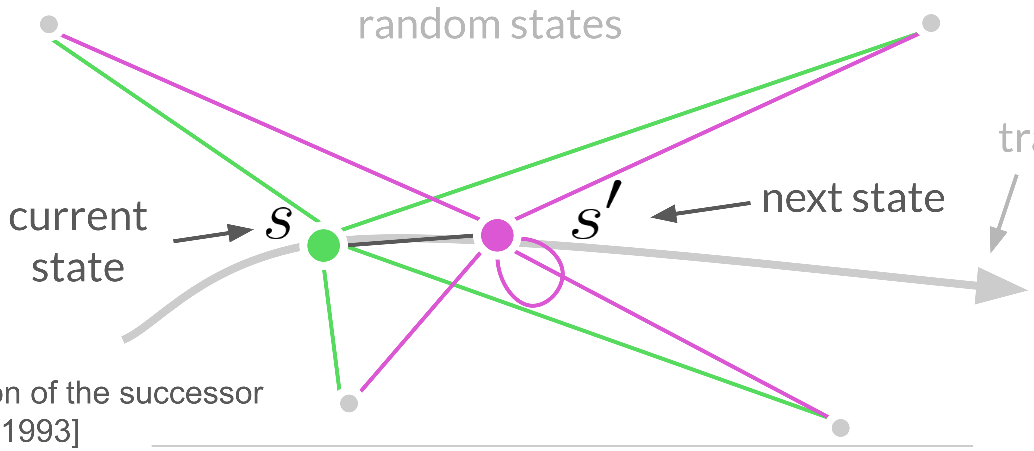


Estimating contrastive successor representations from off-policy data

Zheng, et al. "Contrastive Difference Predictive Coding." ICLR 2024.



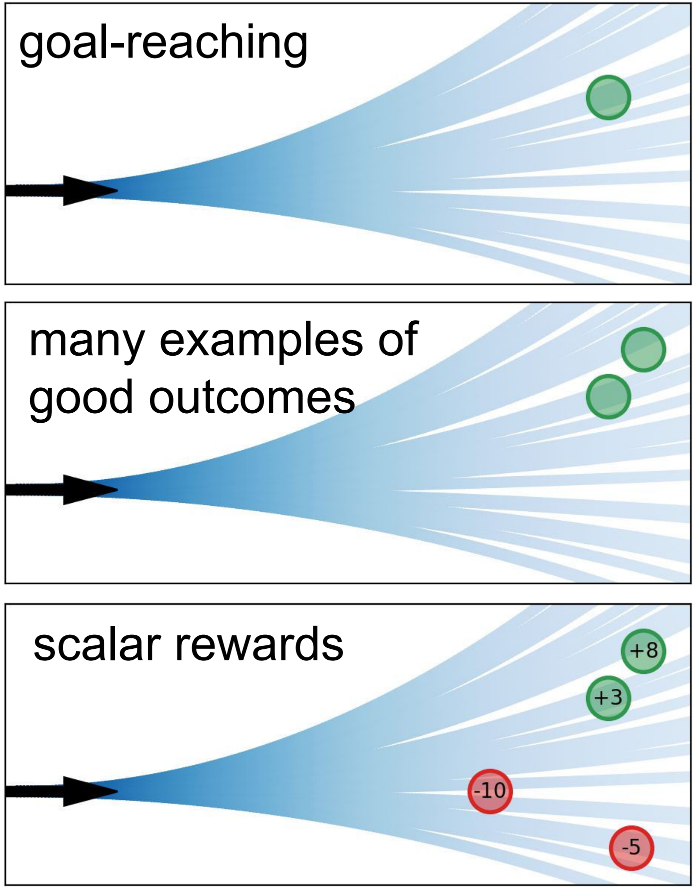
Chongyi Zheng



A nonparametric version of the successor representation [Dayan 1993]

Contrastive approaches to different problems


Kyle Hatch, Bogdan Mazoure



Once you can estimate probability (ratios), you can solve many different types of problems

Mazoure, Bogdan, et al. *Contrastive value learning: Implicit models for simple offline RL*. CoRL, 2023.
Hatch, Kyle Beltran, et al. *Contrastive Example-Based Control*. L4DC, 2023

Steps towards RL agents that can do anything.

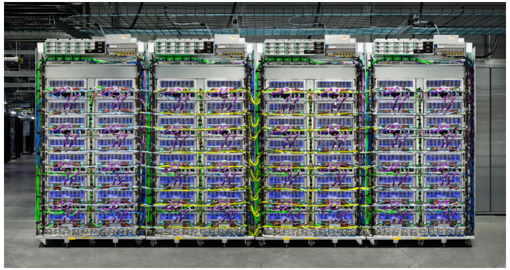
1. Fast simulators
2. Generalization

Lessons from generative AI in other domains

```
def compute_reward(observation):
    objPos = obs[3:6]
    (rightFinger, leftFinger) = (self._get_site_pos('rightEndEffector'),
                                 self._get_site_pos('leftEndEffector'))
    fingerCOM = (rightFinger + leftFinger) / 2
    heightTarget = self.heightTarget
    placingGoal = self._target_pos
    reachDist = np.linalg.norm(objPos - fingerCOM)
    placingDist = np.linalg.norm(objPos[:2] - placingGoal[:-1])
    def reachReward():
        reachRew = -reachDist
        reachDistxy = np.linalg.norm(objPos[:-1] - fingerCOM[:-1])
        zRew = np.linalg.norm(fingerCOM[-1] - self.init_fingerCOM[-1])
        if reachDistxy < 0.06:
            reachRew = -reachDist
        else:
            reachRew = -reachDistxy - zRew
        if reachDist < 0.05:
            reachRew = -reachDist + max(actions[-1], 0) / 50
        return (reachRew, reachDist)
    def pickCompletionCriteria():
        tolerance = 0.01
        if objPos[2] >= heightTarget - tolerance:
            return True
        else:
            return False
    if pickCompletionCriteria():
        self.pickCompleted = True
    def objDropped():
        return objPos[2] < self.objHeight + 0
            > 0.02 and reachDist > 0.02
    def placeCompletionCriteria():
        if abs(objPos[0] - placingGoal[0]) <
                - placingGoal[1]) < 0.05 and
            return True
        else:
            return False
// Yu et al. "Meta-world: A benchmark and eva
reinforcement learning." CoRL, 2020
```

✅ Self-supervised feedback

✅ Lots of data



✅ Lots of compute

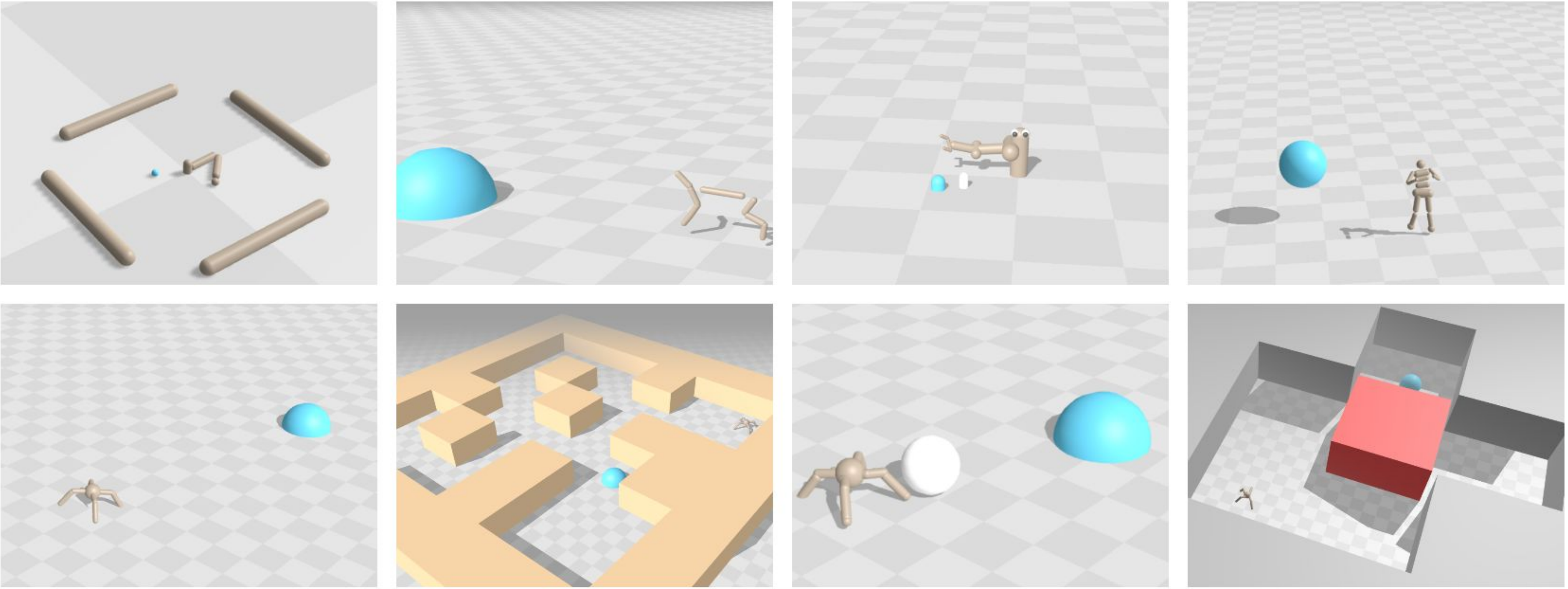
IMAGENET

COCO Common Objects in Context

✅ Standardized benchmarks

Fast Simulators: Jax GCRL

Michał Bortkiewicz,
Władysław Pałucki

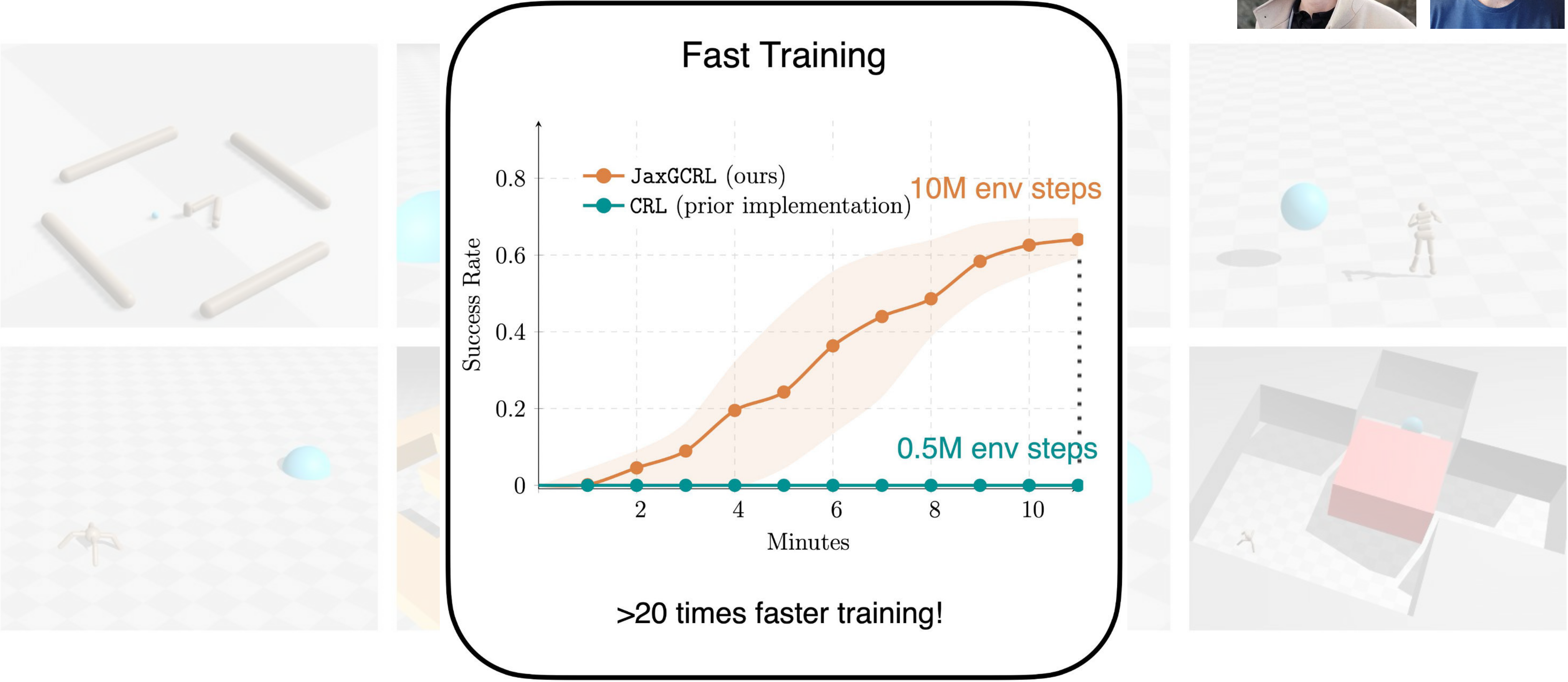


Bortkiewicz, Michał, et al. "Accelerating Goal-Conditioned RL Algorithms and Research." ICLR, 2025

https://github.com/MichalBortkiewicz/JaxGCRL/

Fast Simulators: Jax GCRL

Michał Bortkiewicz,
Władysław Pałucki

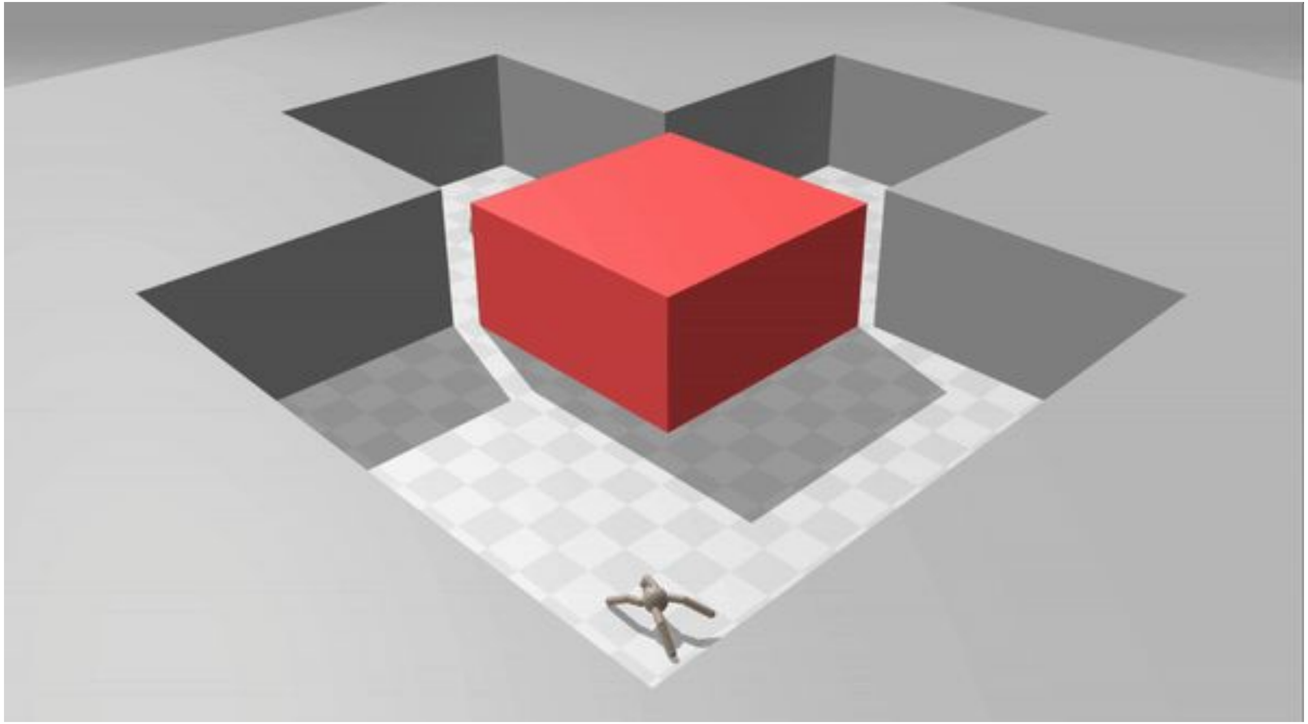


Bortkiewicz, Michał, et al. "Accelerating Goal-Conditioned RL Algorithms and Research." ICLR, 2025

https://github.com/MichalBortkiewicz/JaxGCRL/

Fast Simulators: Jax GCRL

Michał Bortkiewicz,
Władysław Pałucki



Bortkiewicz, Michał, et al. "Accelerating Goal-Conditioned RL Algorithms and Research." ICLR, 2025.

https://github.com/MichalBortkiewicz/JaxGCRL/

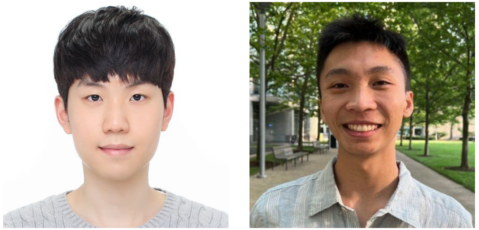
Michał Bortkiewicz,
Władysław Pałucki

Fast Simulators: Jax GCRL

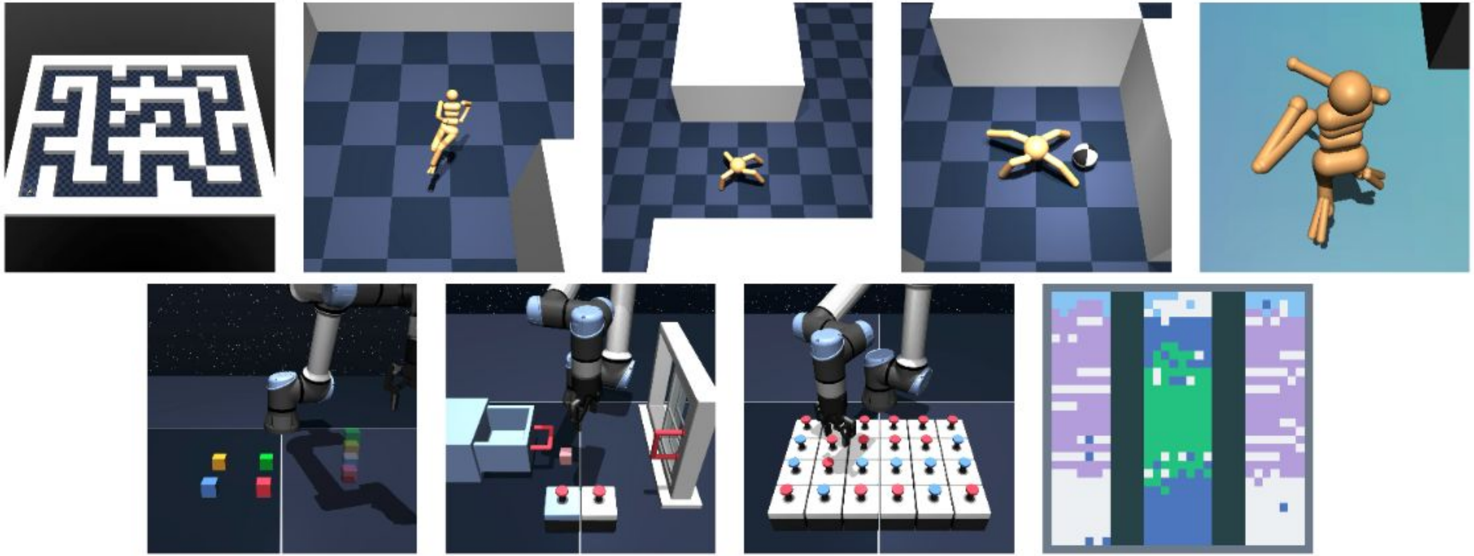
Bortkiewicz, Michał, et al. "Accelerating Goal-Conditioned RL Algorithms and Research." ICLR, 2025.

https://github.com/MichalBortkiewicz/JaxGCRL/

Fast Simulators: Jax GCRL



Seohong Park, Kevin Frans

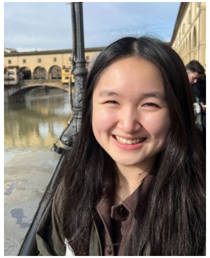


Park, et al. *OGBench: Benchmarking offline goal-conditioned RL*. ICLR, 2025.

Steps towards RL agents that can do anything.

1. Fast simulators
2. **Generalization**

What is the right notion of generalization?

Vivek Myers, Cathy Ji

neural network $\approx$ neural network


action $\approx$ action

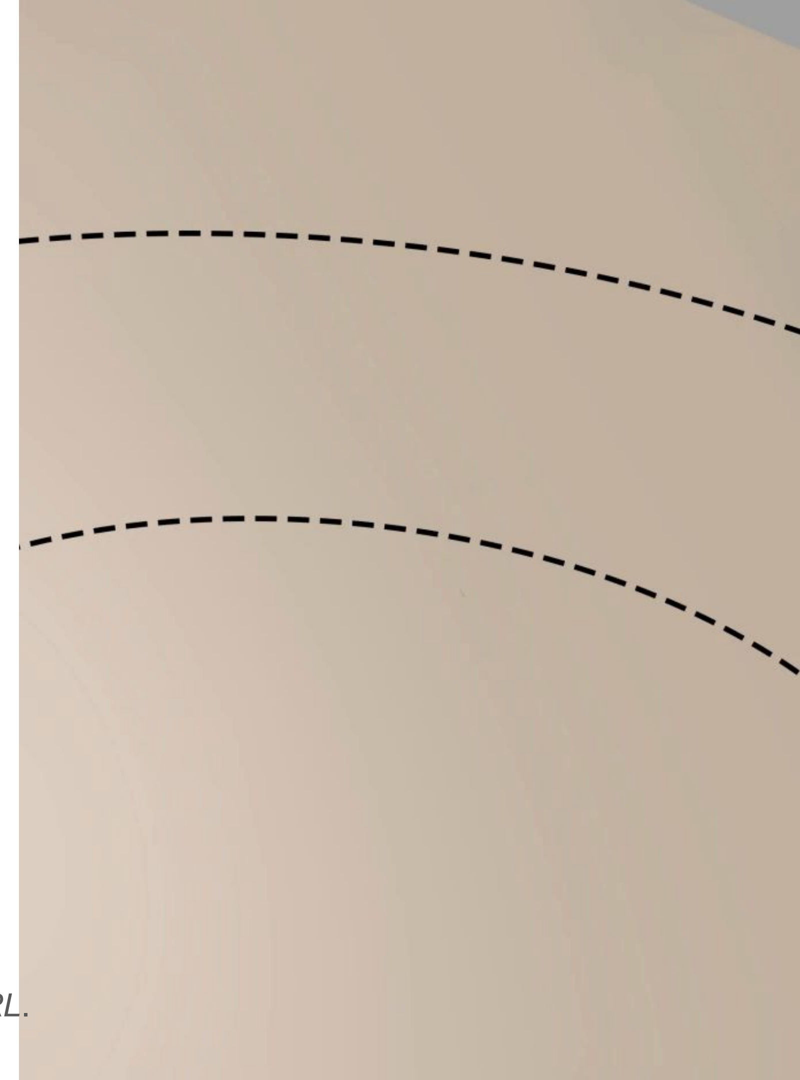
policy policy

$$(s, g) \quad (s, w)$$

Vivek Myers, Catherine Ji, BE. *Invariance to Planning in Goal-Conditioned RL*. ICLR 2025.

What is the right notion of generalization?

Horizon Generalization

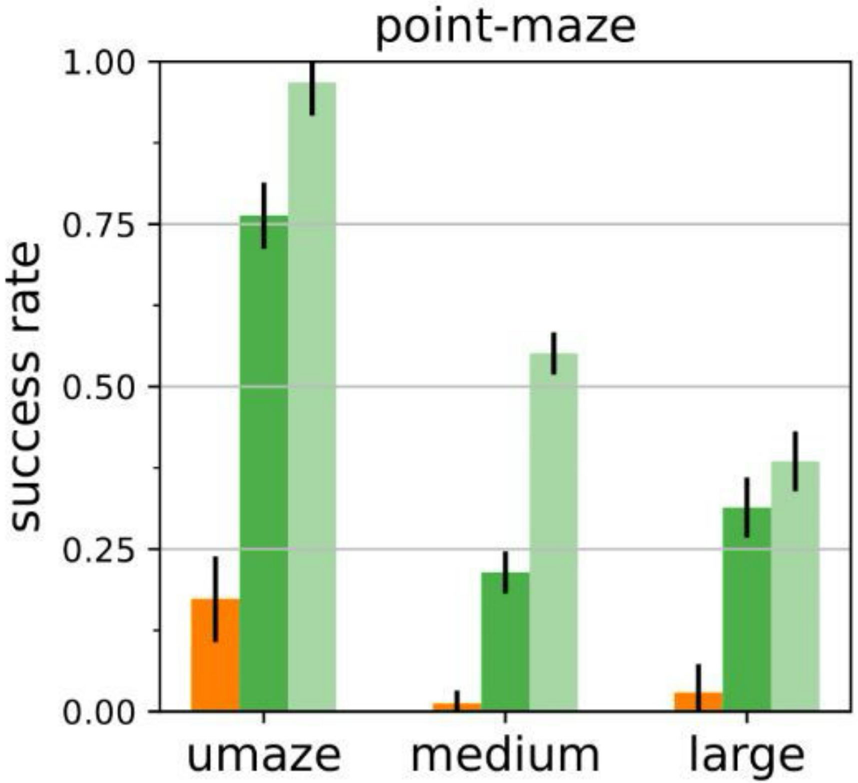


Vivek Myers, Catherine Ji, BE. *Invariance to Planning in Goal-Conditioned RL*. ICLR 2025.

What is the right notion of generalization?

Horizon Generalization

- Data augmentation is useful




Raj Ghugare

Ghugare, Raj, et al. *Closing the Gap between TD Learning and Supervised Learning-A Generalisation Point of View*. ICLR 2024.

Steps towards RL agents that can do anything.

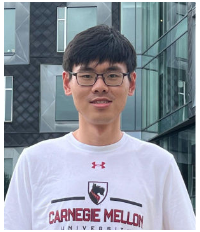
1. Fast simulators
2. Generalization

Self-supervised RL is generative AI
… so can we learn to do anything?

Three preliminary signs of life.

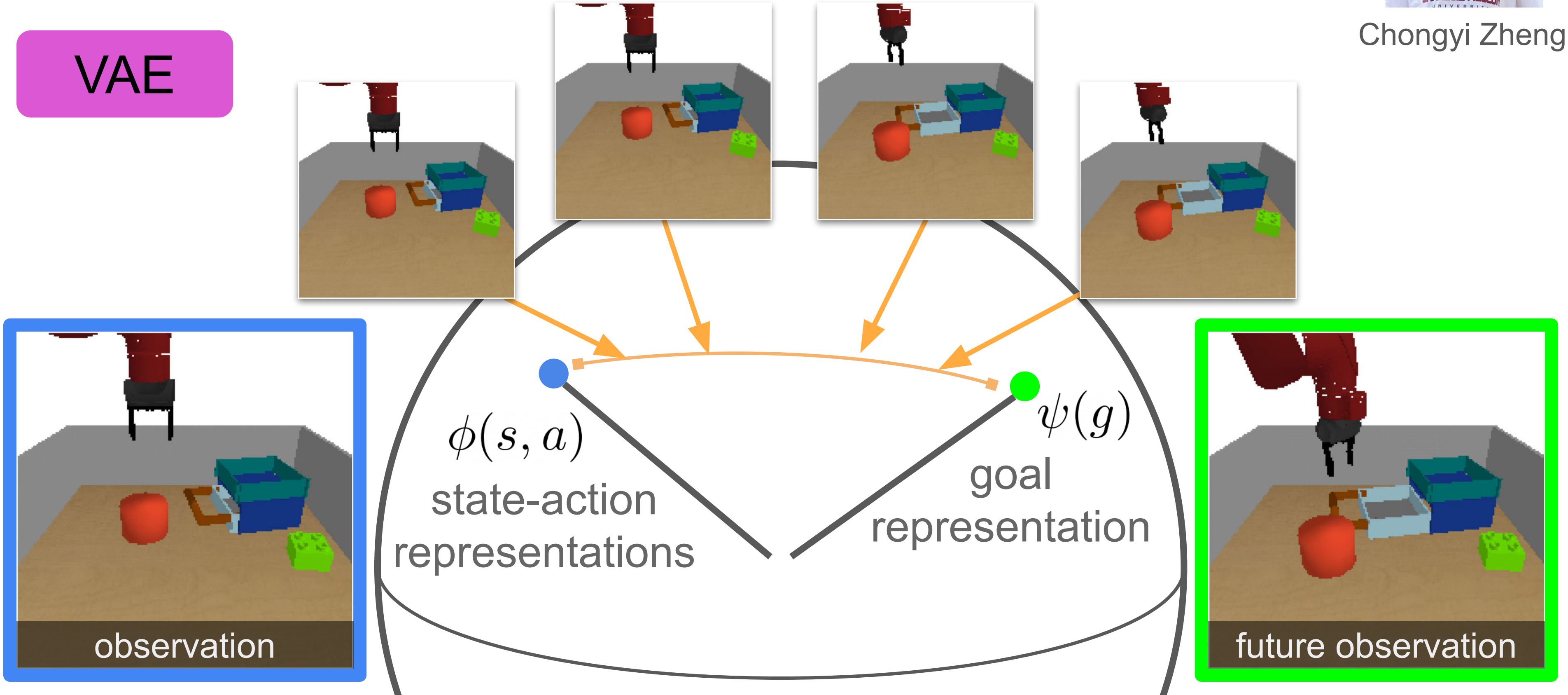
Emergent Properties in Self-Supervised RL 1/

Zheng, et al. *Stabilizing Contrastive RL: Techniques for Robotic Goal Reaching from Offline Data*. ICLR, 2024



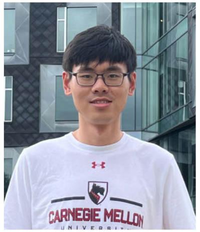
Chongyi Zheng

VAE



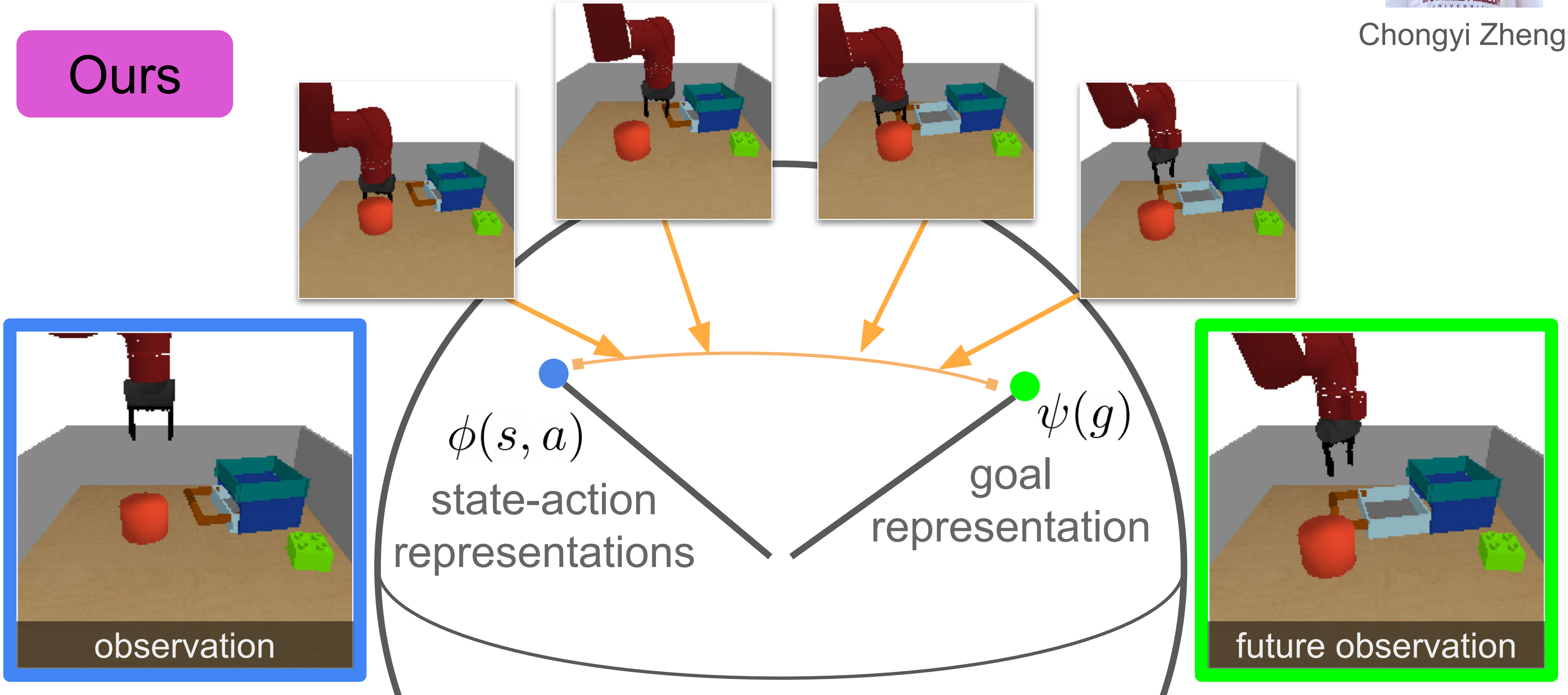
Emergent Properties in Self-Supervised RL 1/

Zheng, et al. *Stabilizing Contrastive RL: Techniques for Robotic Goal Reaching from Offline Data*. ICLR, 2024



Chongyi Zheng

Ours



Emergent Properties in Self-Supervised RL
1/ Representations that Interpolate

Vivek Myers

Theorem (informal): Under some assumptions, planning over representations corresponds to inference on a Gaussian graphical model.

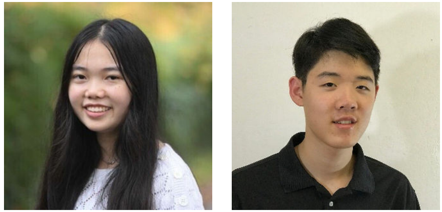
$$p(\psi_{1:n}) \propto \exp\left(-\tfrac{1}{2}\psi_{1:n}^T \Sigma^{-1} \psi_{1:n} + \eta^T \psi_{1:n}\right),$$

$$\Sigma^{-1} = \begin{pmatrix} \frac{c}{c+1}A^T A + \frac{c+1}{c} I & -A^T & \\ -A & \frac{c}{c+1}A^T A + \frac{c+1}{c} I & -A^T \\ & & \ddots \end{pmatrix}$$

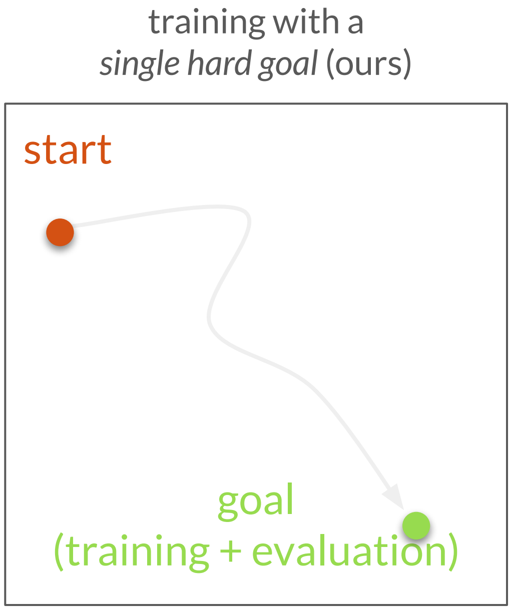
$$\text{and } \eta = \begin{pmatrix} A\psi_0 \\ 0 \\ \vdots \\ A^T \psi_{t+} \end{pmatrix}.$$

BE*, Vivek Myers, et al. *Inference via interpolation: Contrastive representations provably enable planning and inference.* NeurIPS, 2024.

Emergent Properties in Self-Supervised RL
2/



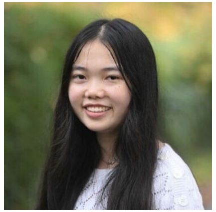
Grace Liu, Michael Tang



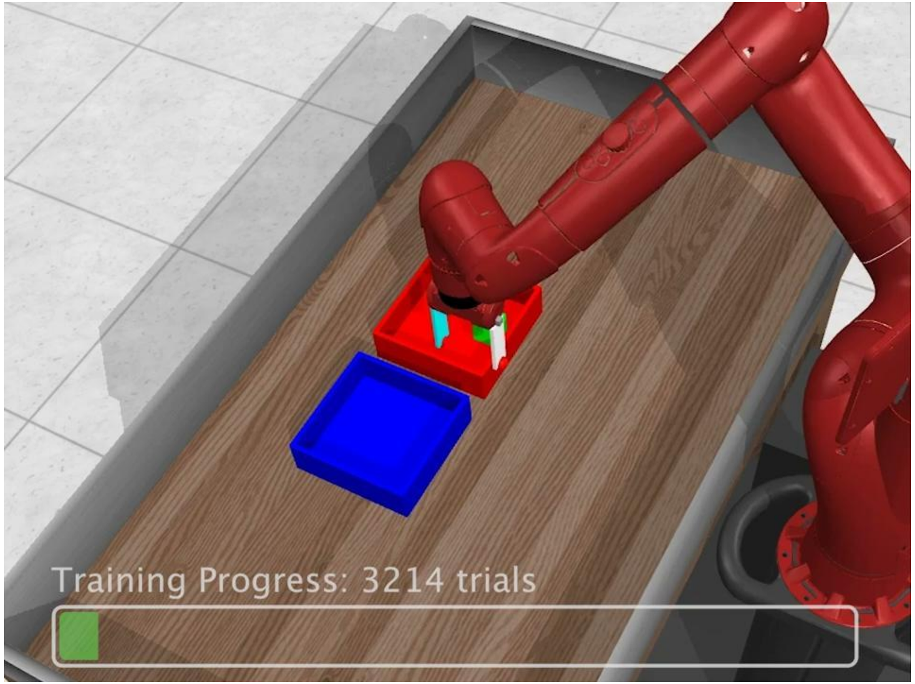
If random exploration never reaches the goal, will this learn anything?

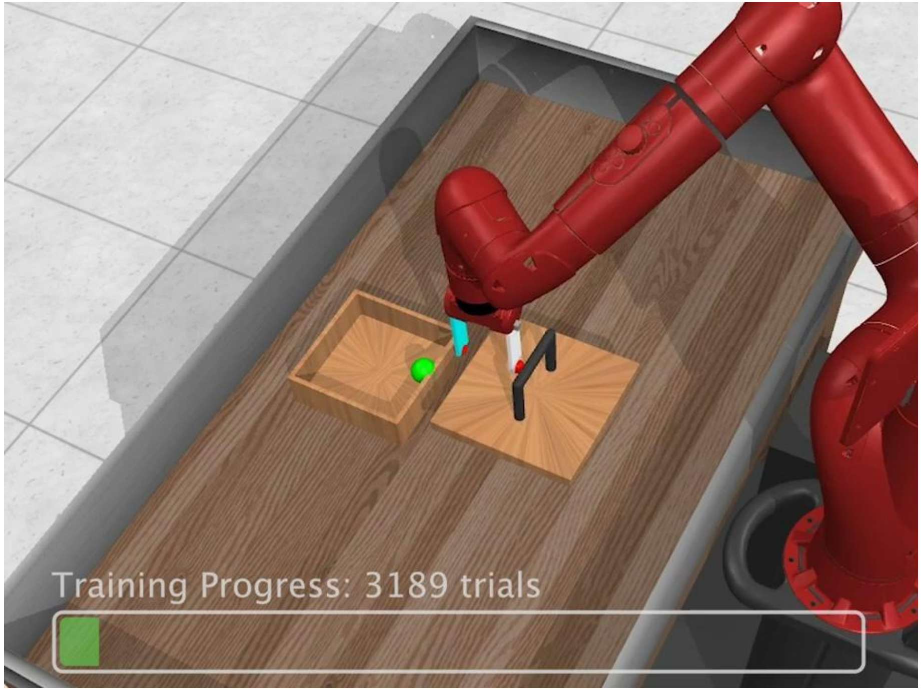
Liu, Grace, Michael Tang, and BE. *A Single Goal is All You Need: Skills and Exploration Emerge from Contrastive RL without Rewards, Demonstrations, or Subgoals*. ICLR, 2025.

Emergent Properties in Self-Supervised RL
2/




Grace Liu, Michael Tang





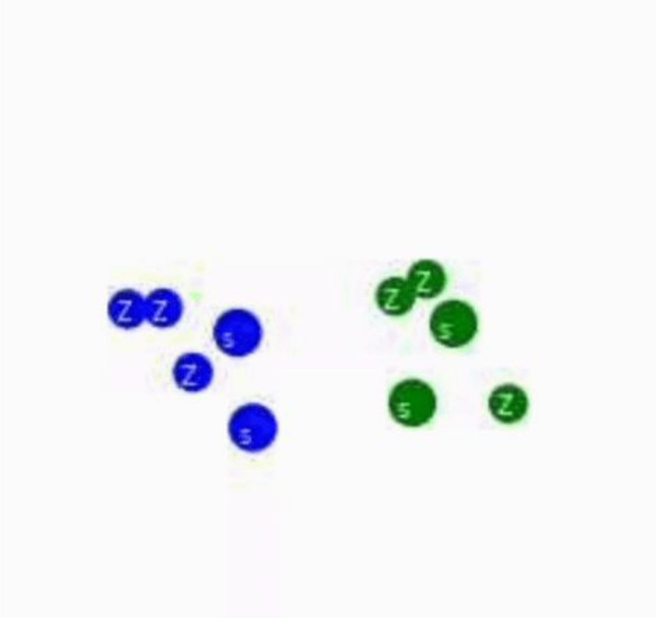
Liu, Grace, Michael Tang, and BE. *A Single Goal is All You Need: Skills and Exploration Emerge from Contrastive RL without Rewards, Demonstrations, or Subgoals*. ICLR, 2025

Emergent Properties in Self-Supervised RL
2/ Multi-Agent Exploration

Chirayu Nimonkar, Shlok Shah

Samvelyan, Mikayel, et al. "The starcraft multi-agent challenge." arXiv preprint arXiv:1902.04043 (2019).

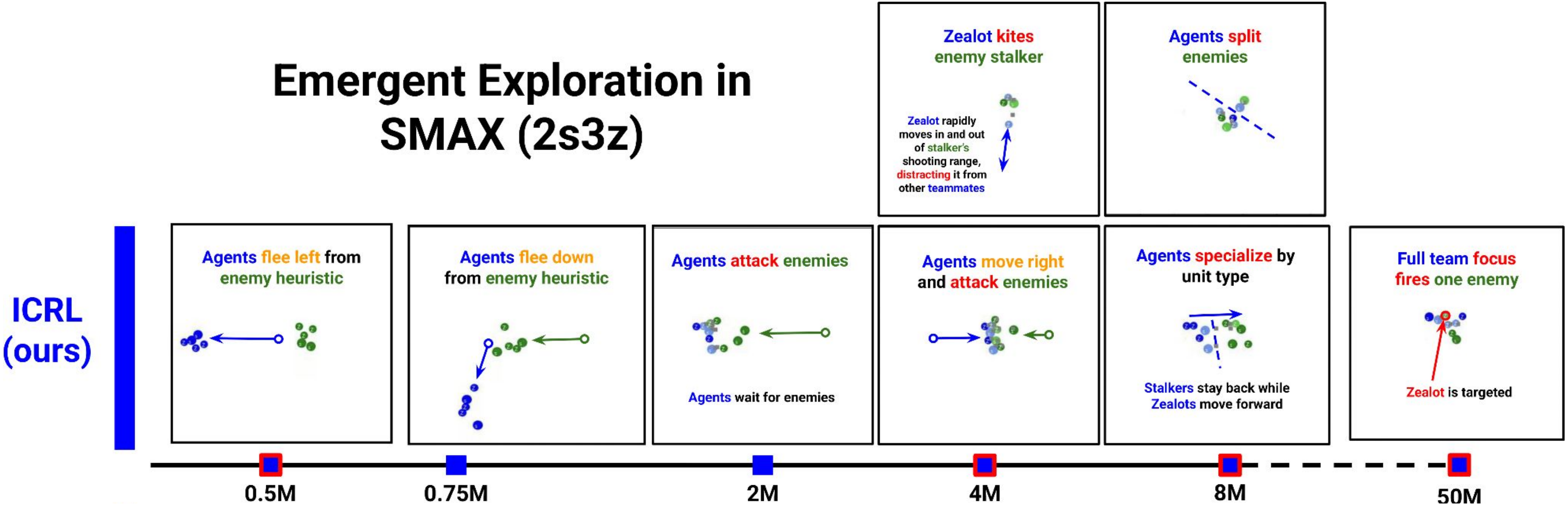


Chirayu Nimonkar, Shlok Shah, and BE. *Goal-Conditioned Multi-Agent Cooperation with Contrastive RL*. In Submission, 2025.

Emergent Properties in Self-Supervised RL

2/ Multi-Agent Exploration

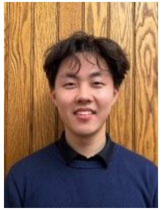
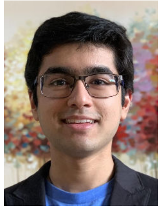
Chirayu Nimonkar, Shlok Shah



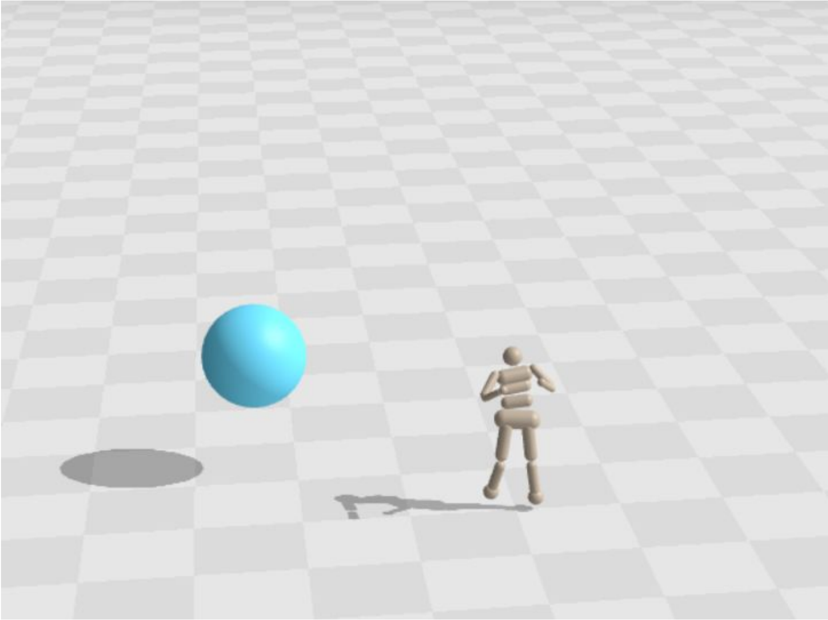
Chirayu Nimonkar, Shlok Shah, and BE. *Goal-Conditioned Multi-Agent Cooperation with Contrastive RL*. In Submission, 2025.

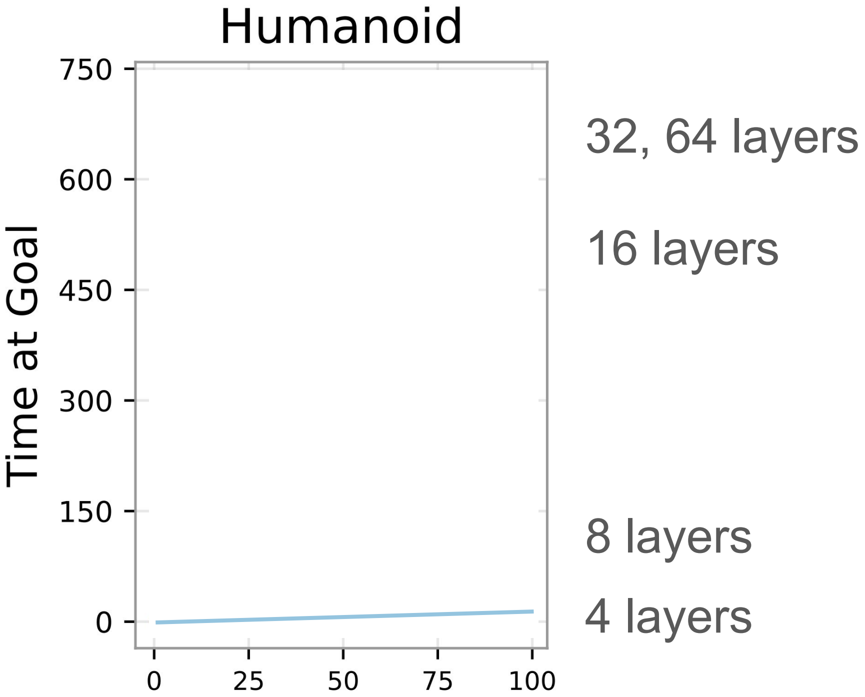
Emergent Properties in Self-Supervised RL

3/ Scale unlocks new behaviors

Kevin Wang, Ishaan Javali

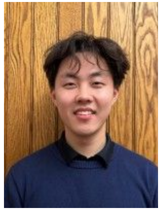
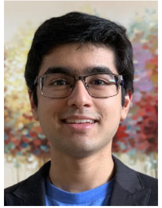




Kevin Wang, Ishaan Javali, et al. *500 Layer Networks for Self-Supervised RL: Scaling Depth Can Enable New Goal-Reaching Capabilities.* In Submission, 2025.

Emergent Properties in Self-Supervised RL

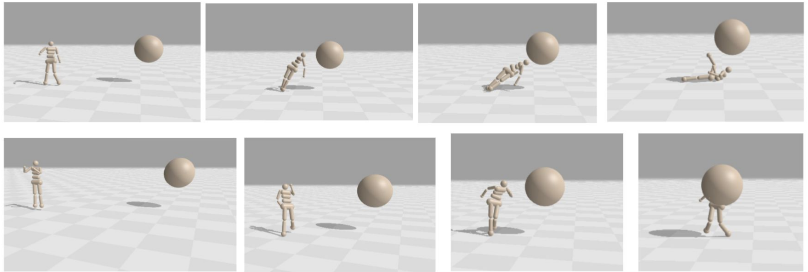
3/ Scale unlocks new behaviors

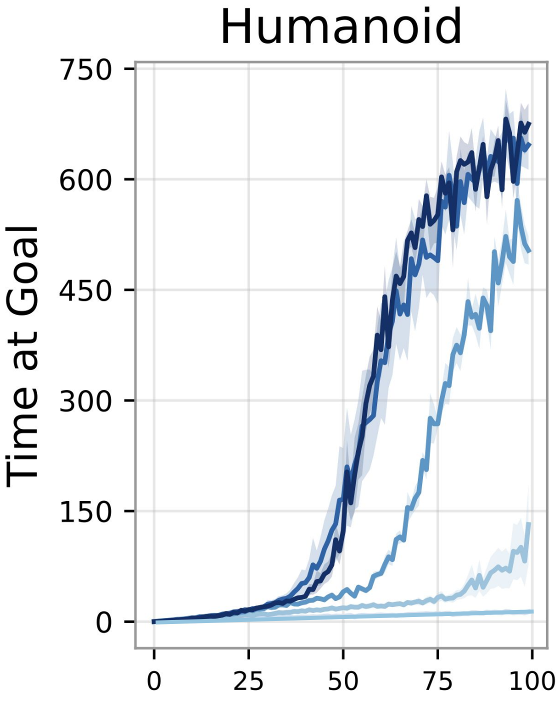



Kevin Wang, Ishaan Javali

Depth 4

Depth 16





32, 64 layers

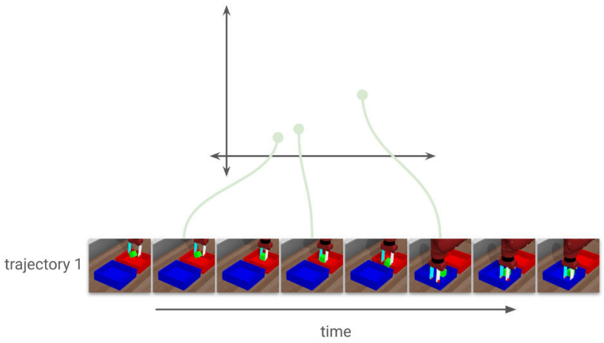
16 layers

8 layers

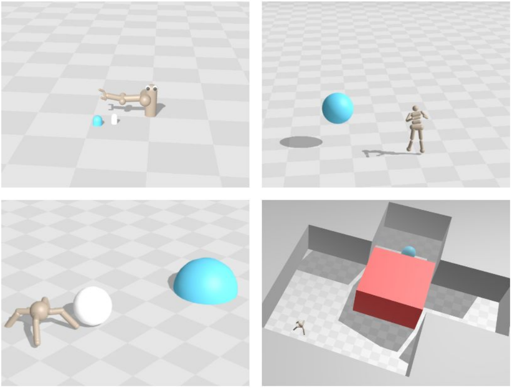
4 layers

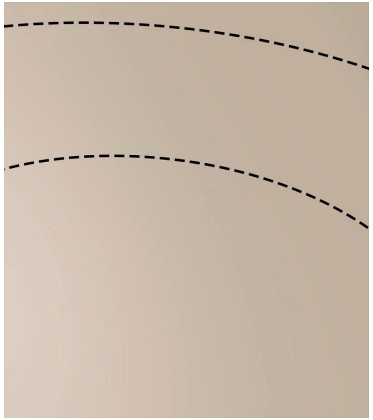
Kevin Wang, Ishaan Javali, et al. *500 Layer Networks for Self-Supervised RL: Scaling Depth Can Enable New Goal-Reaching Capabilities.* In Submission, 2025.

Key takeaways



1/ Self-supervised RL





2/ Steps towards RL agents that can learn to do anything with minimal feedback

Get started with research!

Papers highlighted today:

1. BE, et al. *Contrastive learning as goal-conditioned reinforcement learning*. NeurIPS, 2022.
2. Bortkiewicz, et al. *Accelerating Goal-Conditioned RL Algorithms and Research*. ICLR, 2025.
3. Liu, et al. *A Single Goal is All You Need: Skills and Exploration Emerge from Contrastive RL witho Rewards, Demonstrations, or Subgoals*. ICLR, 2025.
4. Myers, Ji, BE. *Horizon Generalization in Reinforcement Learning.* ICLR, 2025.