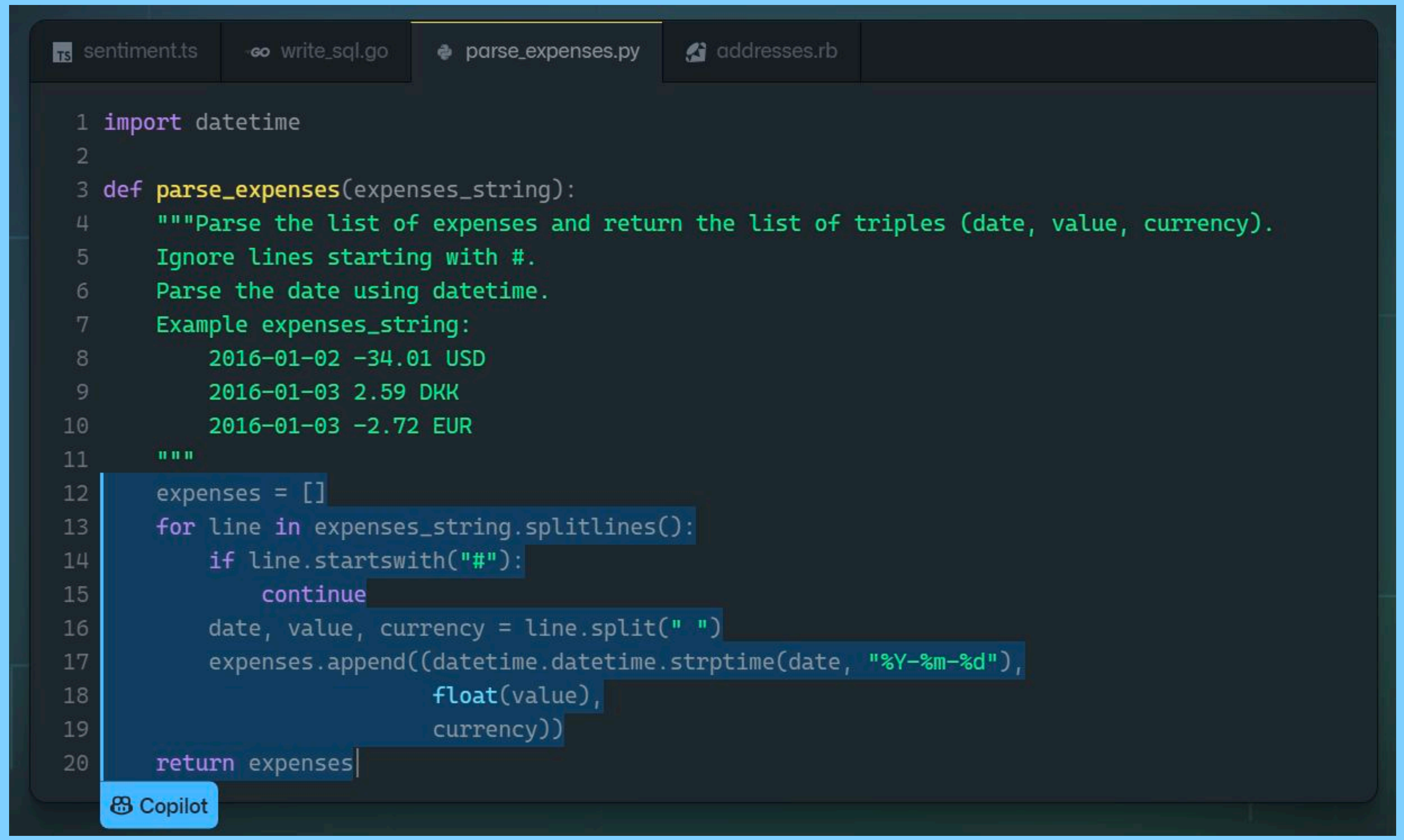
Using Natural Language for Task Specification in Sequential Decision Making Problems

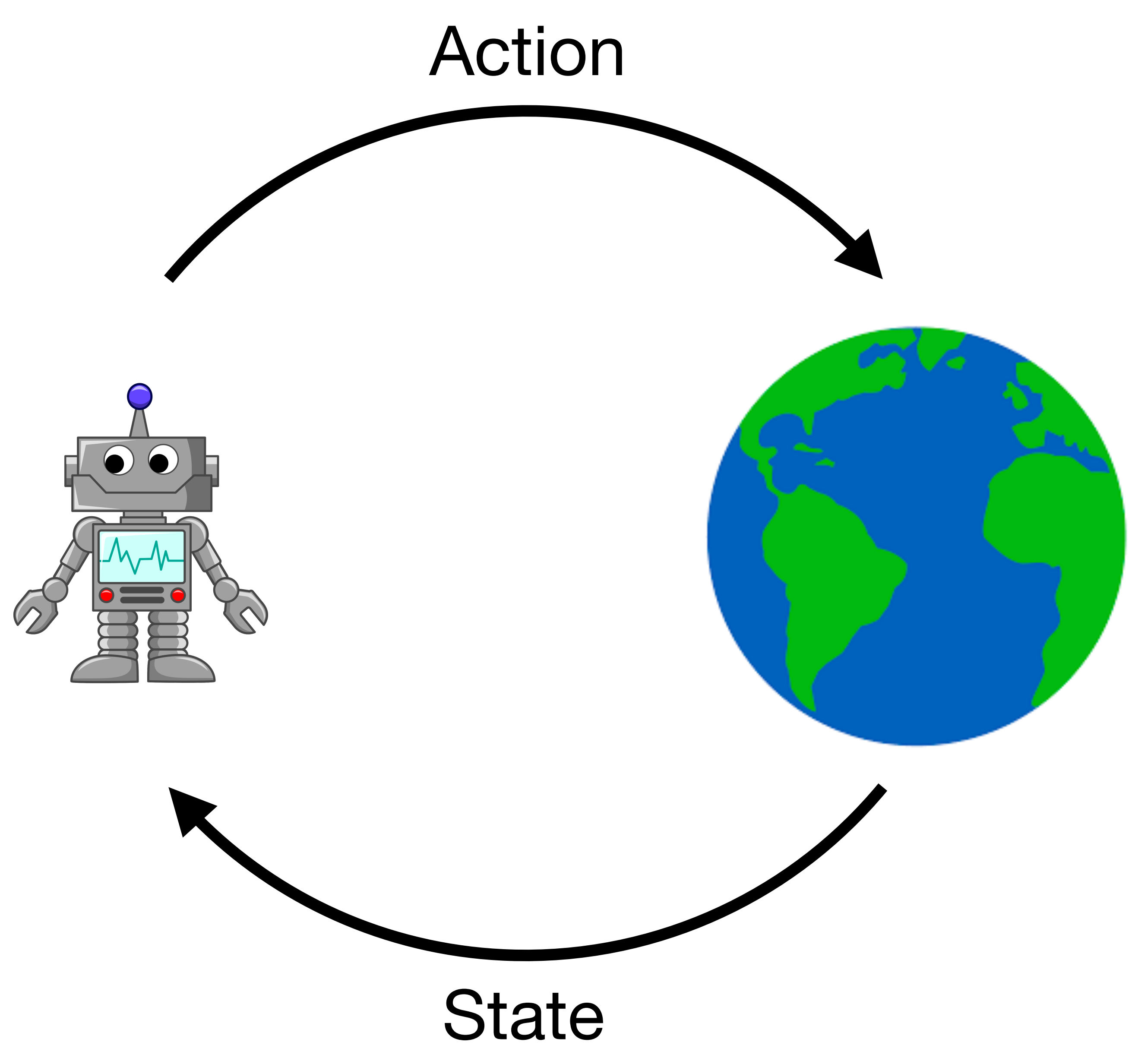
Prasoon Goyal

Dissertation Defense

A lot of progress in AI in the last several decades…

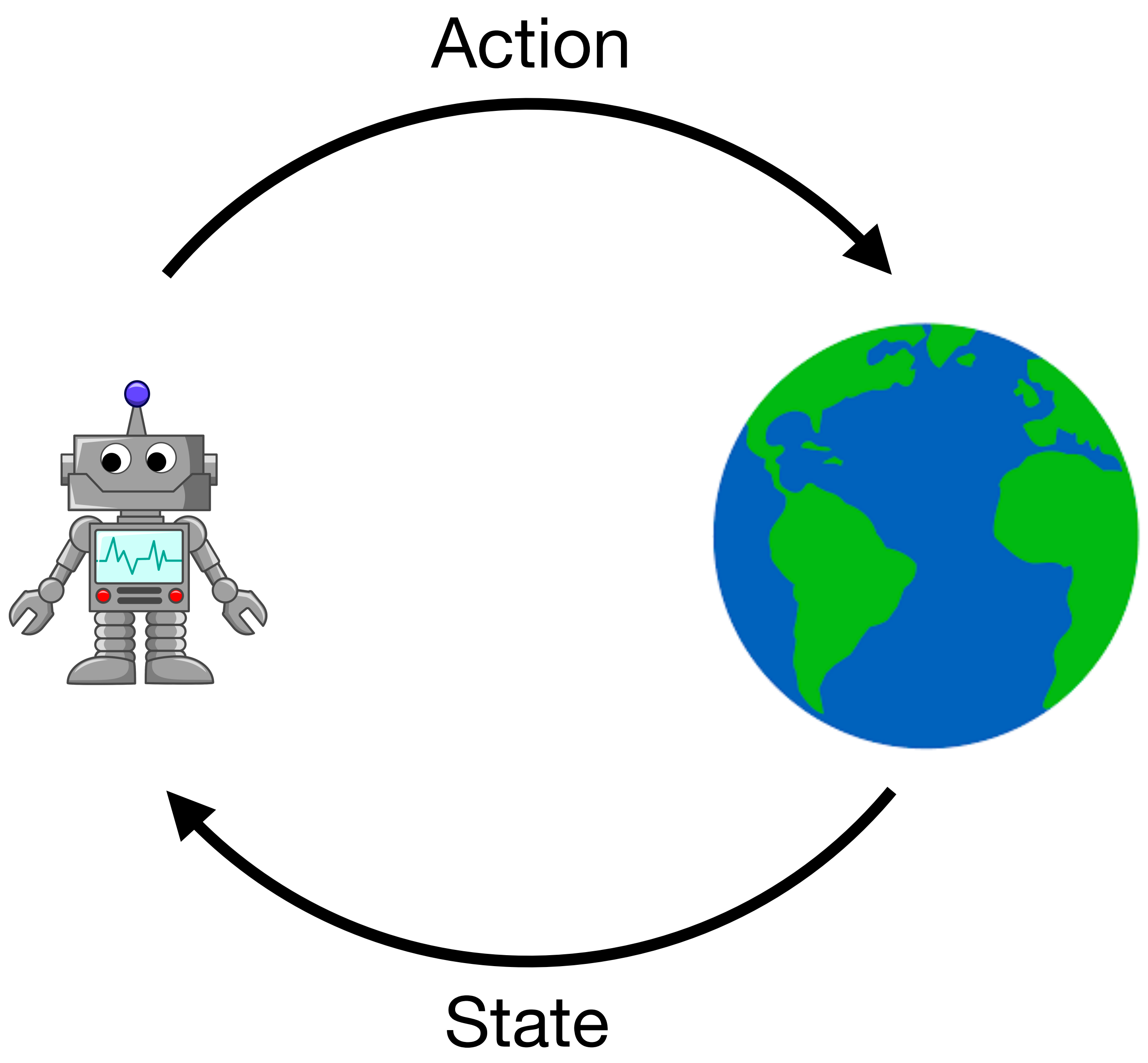


Sequential Decision Making



Can be used for a broad class of real-world tasks:
- **Turn-based games like chess, Go, etc.**
- **Video games**
- **Cooking tasks**
- **...**

Sequential Decision Making



Can be used for a broad class of real-world tasks:
- **Turn-based games like chess, Go, etc.**
- **Video games**
- **Cooking tasks**
- **...**

Describing the desired task to the agent

Rewards

Reinforcement Learning (RL)

Demonstrations

Imitation Learning (IL)

How can natural language be used as an auxiliary signal to reduce the burden of task specification on the end user?

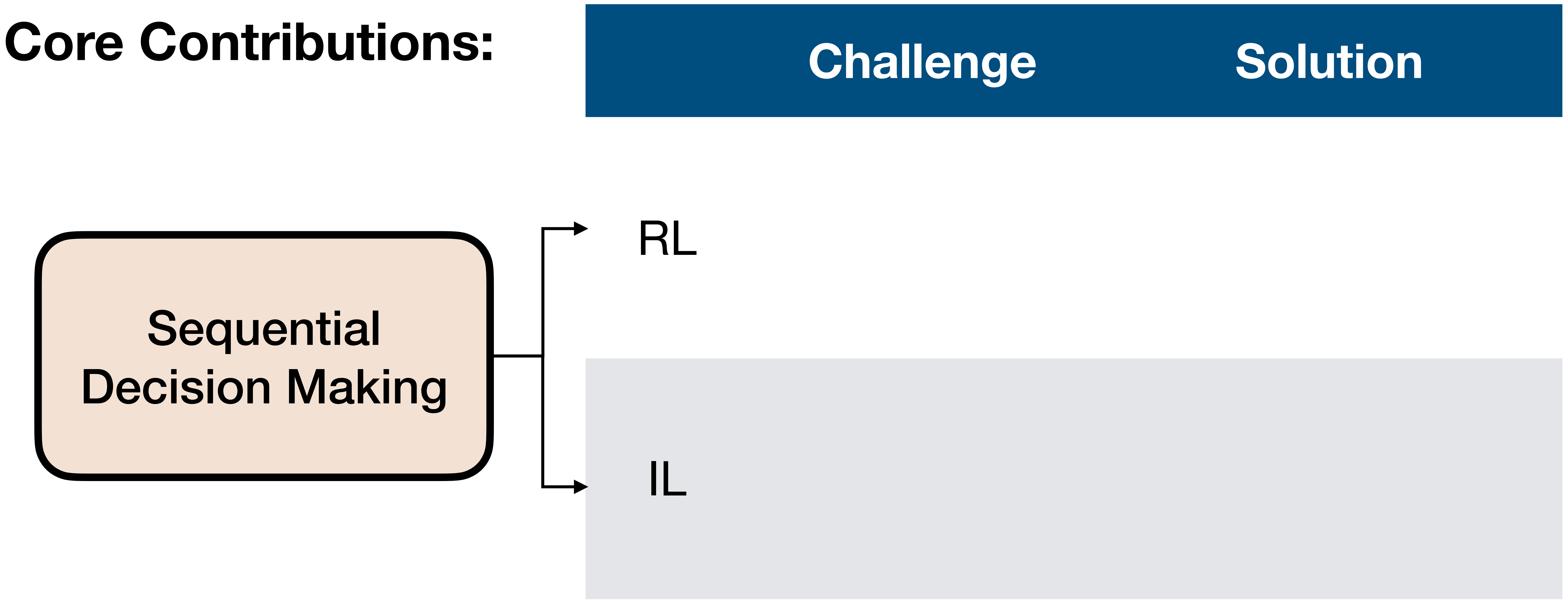
Sequential Decision Making

	Challenge	Solution
RL		
IL		

Talk Outline

Background

Core Contributions:



Future Directions

Talk Outline

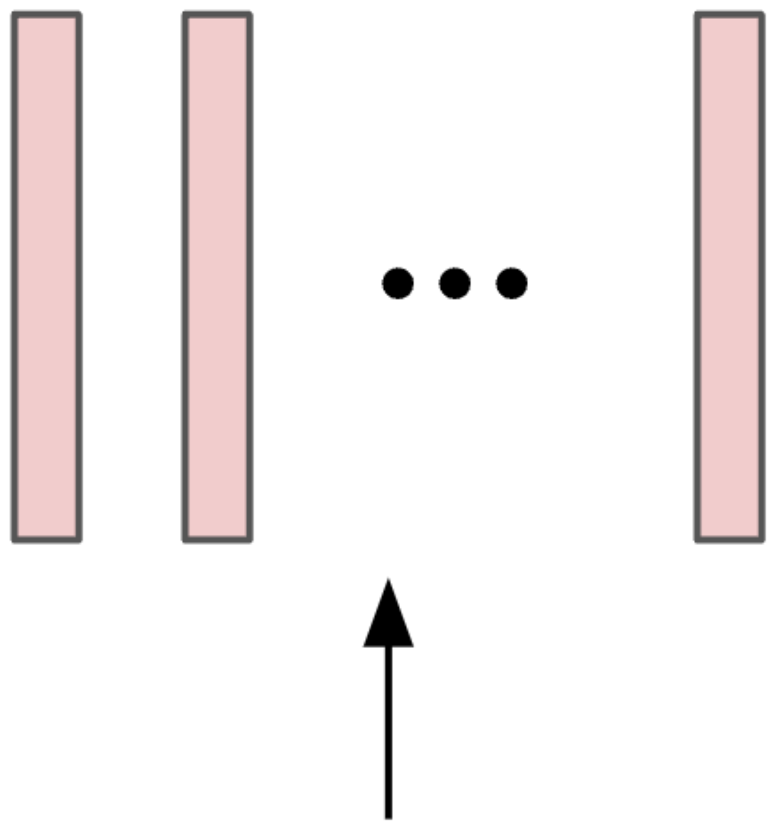
Background

Core Contributions:

	Challenge	Solution
Sequential Decision Making → RL		
Sequential Decision Making → IL		

Future Directions

Background: Language Encoders



Turn the red handle down.

- One-hot embeddings of words + RNN to encode the sequence
- Word embeddings + RNN to encode the sequence
- Transformer-based pretrained sentence encoders (e.g. BERT, CLIP)
- Transformer-based sentence encoder trained from scratch

Talk Outline

Background

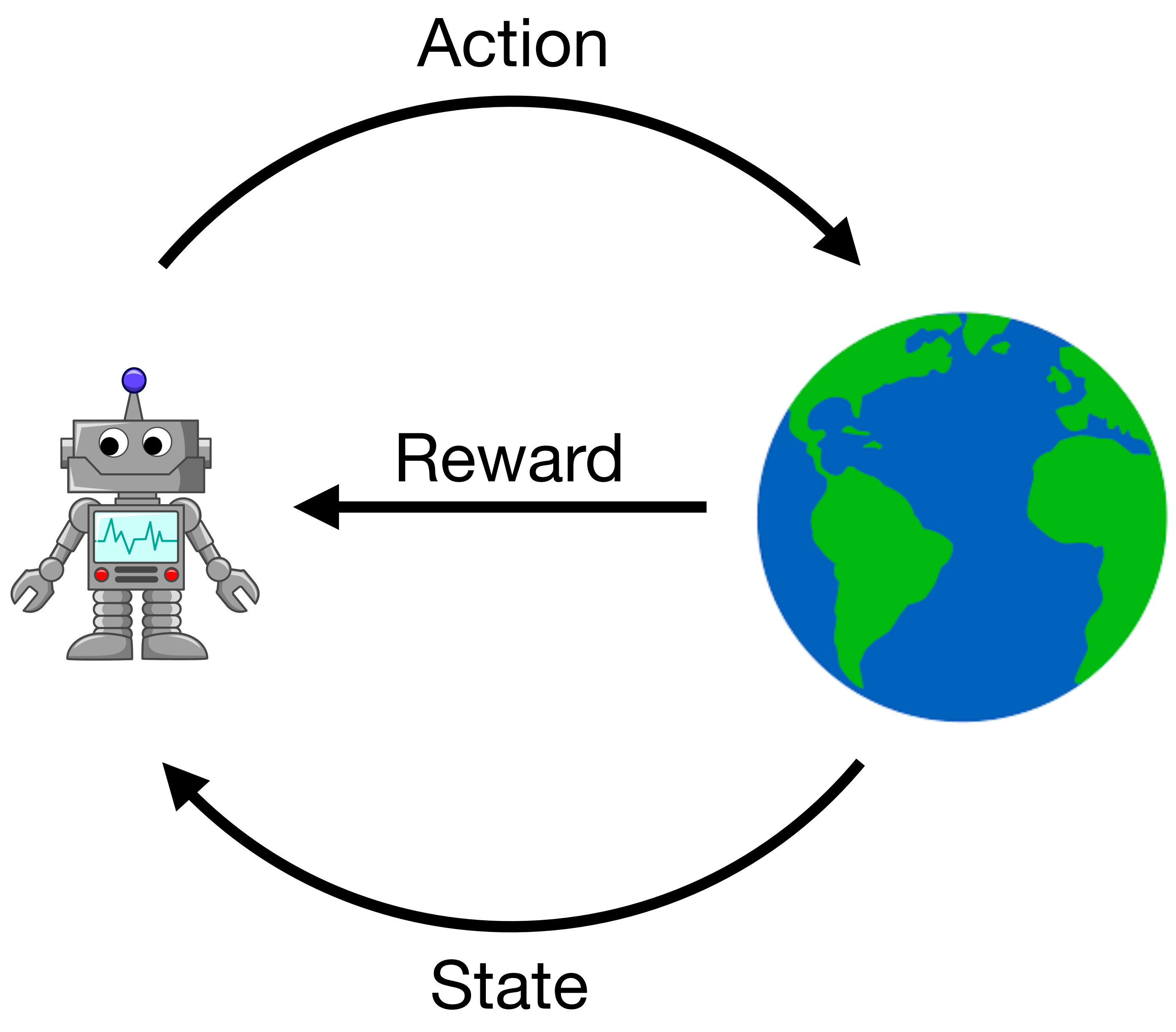
Core Contributions:

	Challenge	Solution
RL		
IL		

Sequential Decision Making

Future Directions

Reinforcement Learning



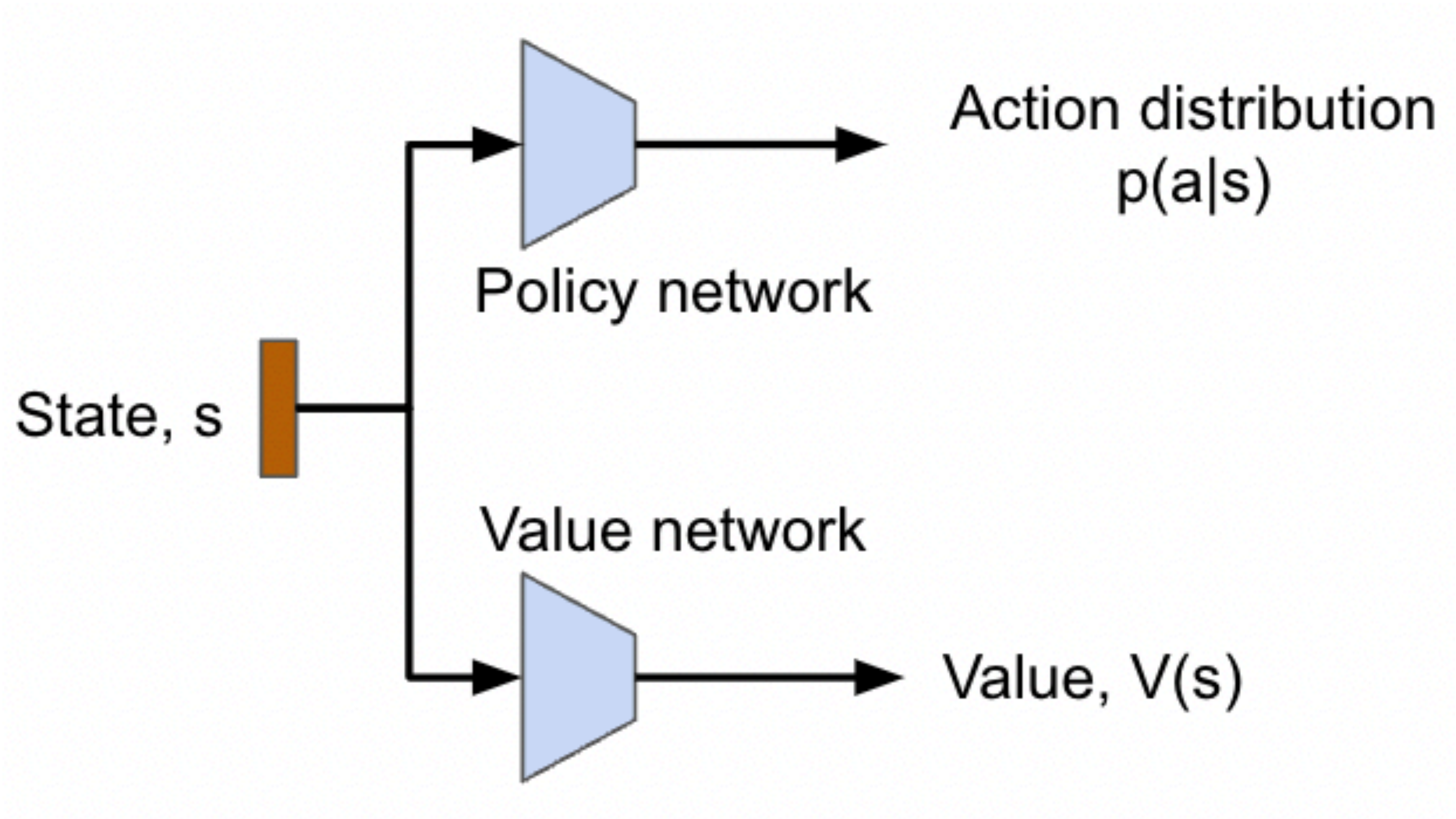
Markov Decision Process (MDP), $M = \langle S, A, T, R, \gamma \rangle$

- S : State space
- A : Action space
- T : Transition function
- R : Reward function
- γ : Discount factor

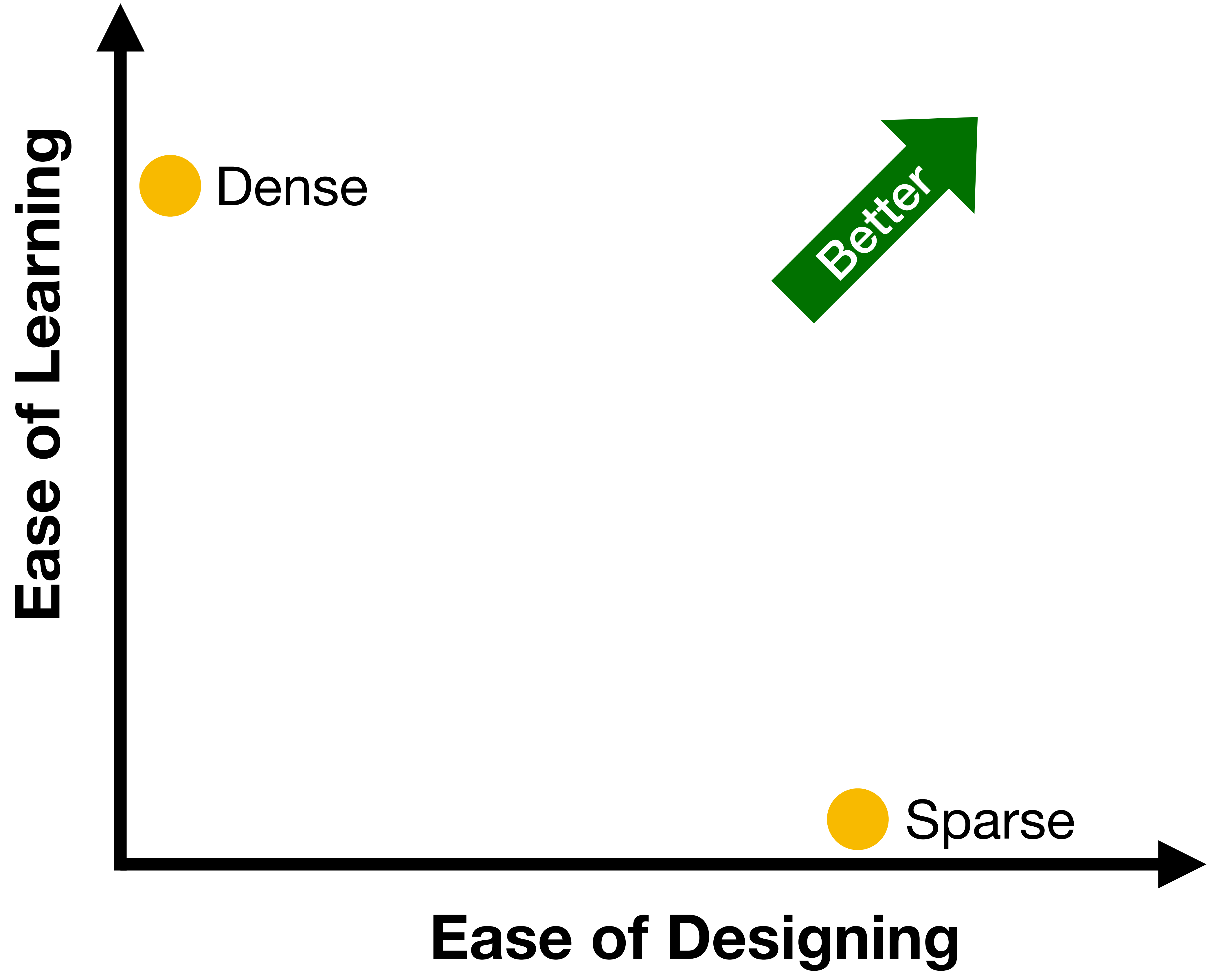
Policy: Maps states to action probabilities

Objective: Learn a policy that maximizes the discounted sum of future rewards.

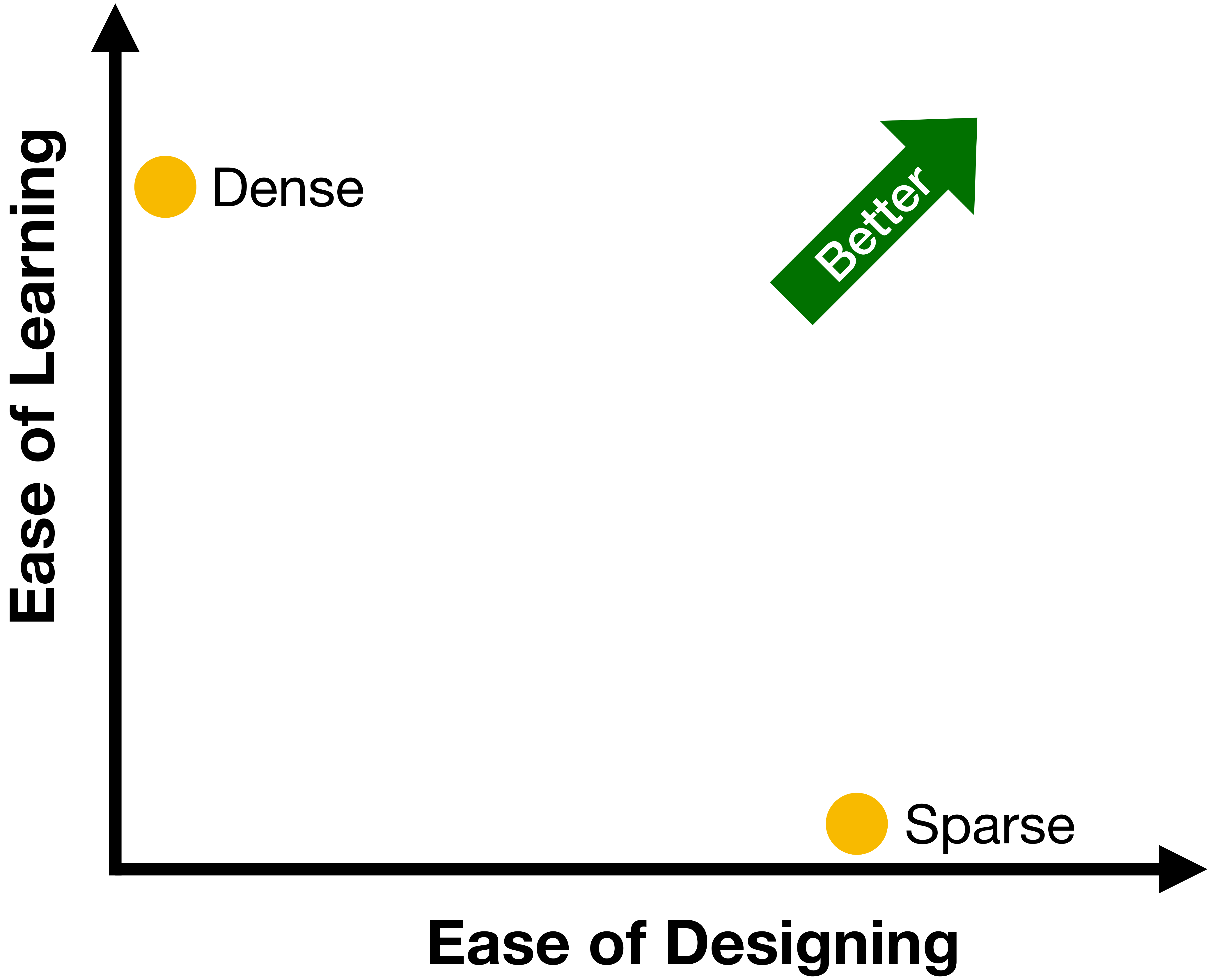
Reinforcement Learning: Proximal Policy Optimization (PPO)



Challenge in RL: Designing rewards is Hard



Challenge in RL: Designing rewards is Hard



Use natural language to generate auxiliary rewards

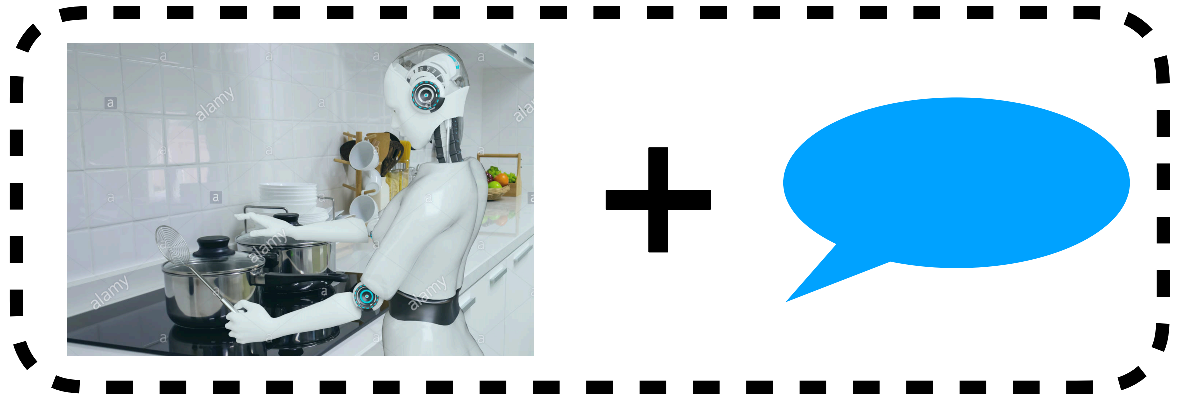
Talk Outline

Background

Core Contributions:

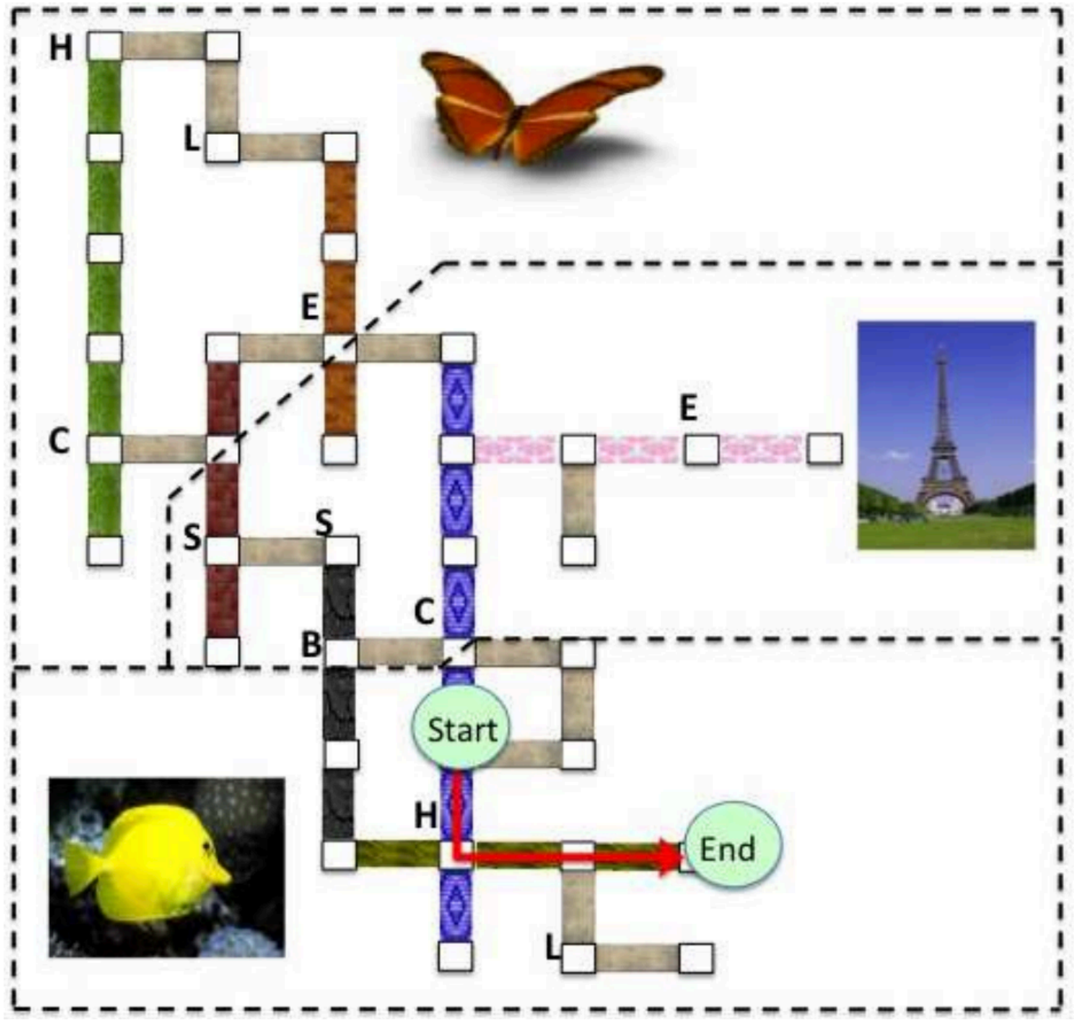
Sequential Decision Making

	Challenge	Solution
RL	Reward design	Language-based Rewards
IL		

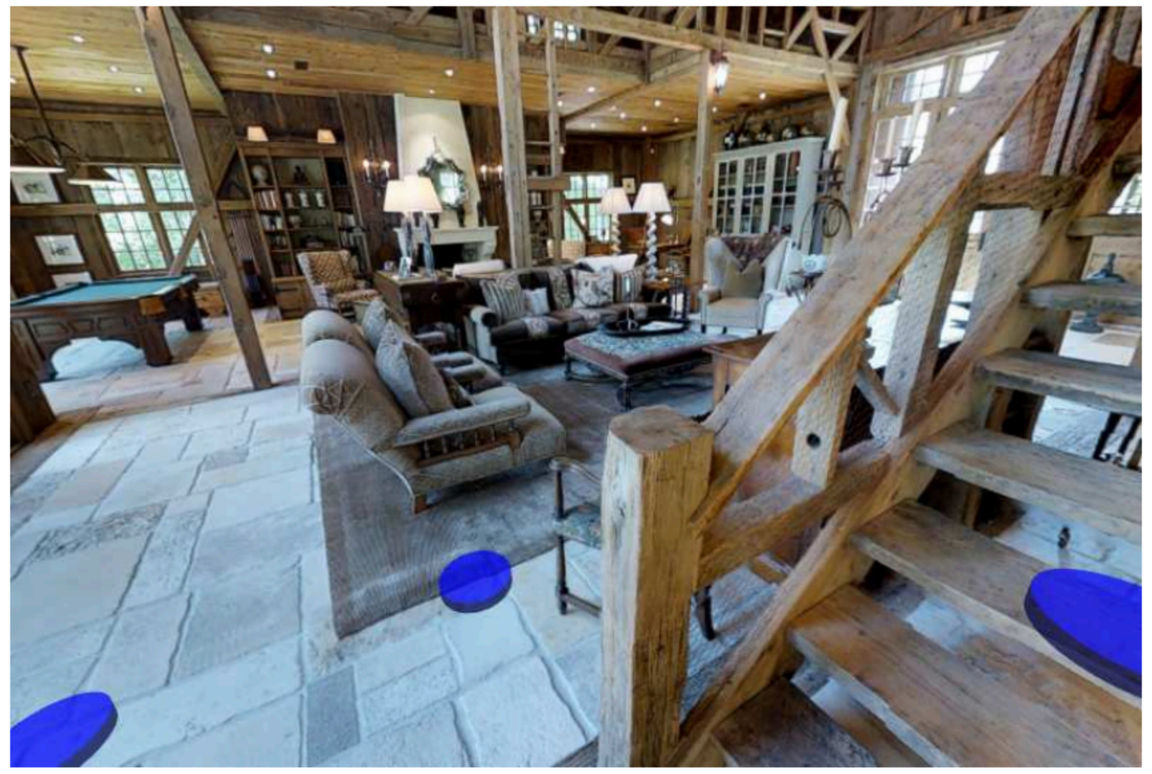


Future Directions

Related Work: Instruction-following



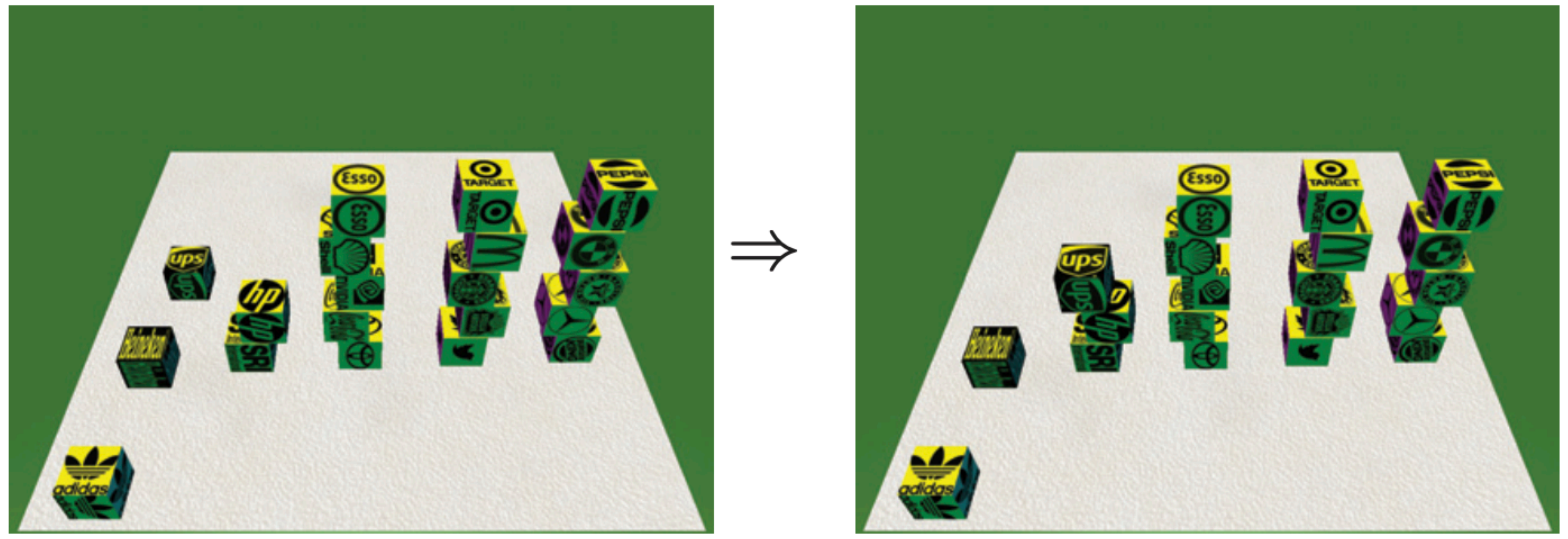
Chen and Mooney, 2011



Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

Anderson et al., 2018

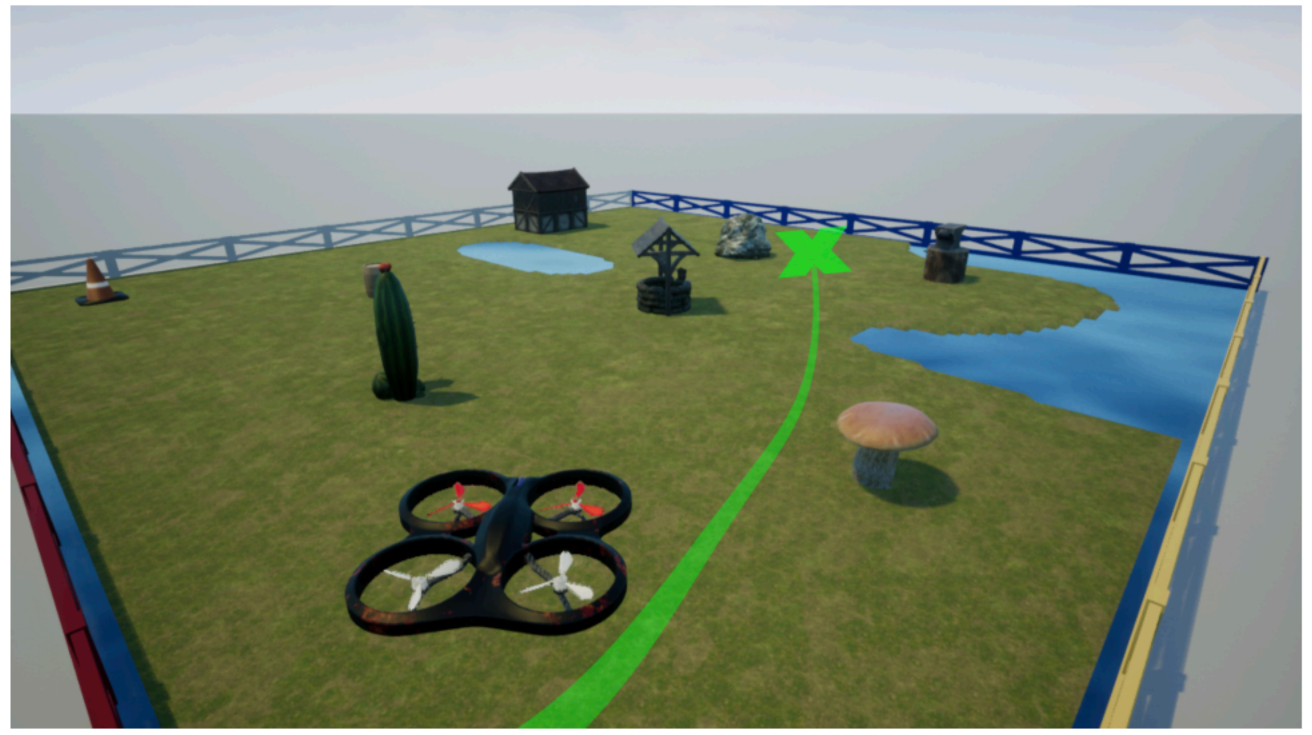
"On the (new) fourth tower, mirror Nvidia with UPS."



Bisk et al., 2018

Goal: "Rinse off a mug and place it in the coffee maker"

Shridhar et al., 2020



Go to the right side of the rock

Blukis et al., 2018

Commands from the corpus
- Go to the first crate on the left and pick it up. - Pick up the pallet of boxes in the middle and place them on the trailer to the left. - Go forward and drop the pallets to the right of the first set of tires. - Pick up the tire pallet off the truck and set it down

Tellex et al., 2011

Talk Outline

Background

Core Contributions:

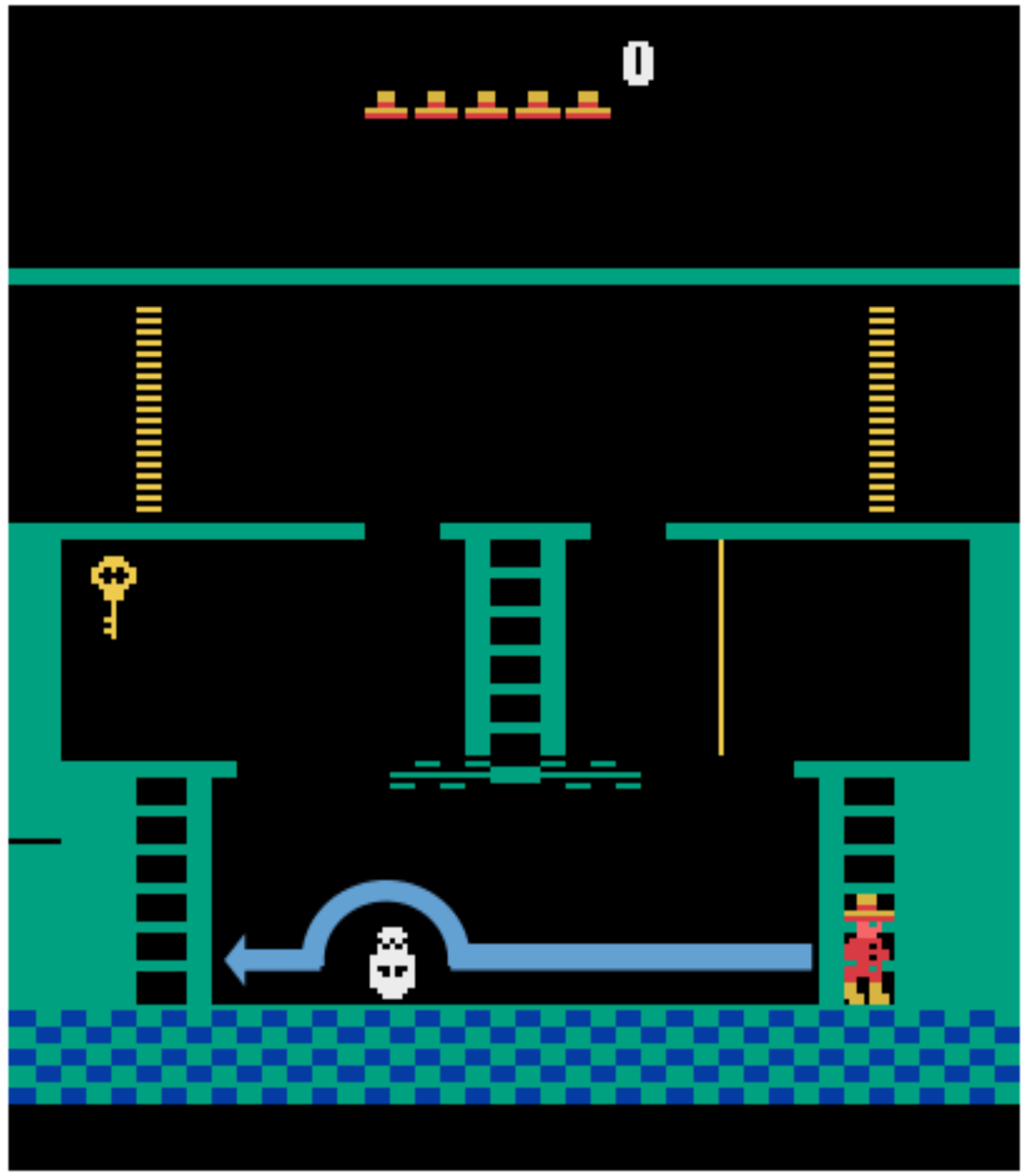
Sequential Decision Making

	Challenge	Solution
RL	Reward design	Language-based Rewards
IL		

Future Directions

[Bellemare et al., 2013]

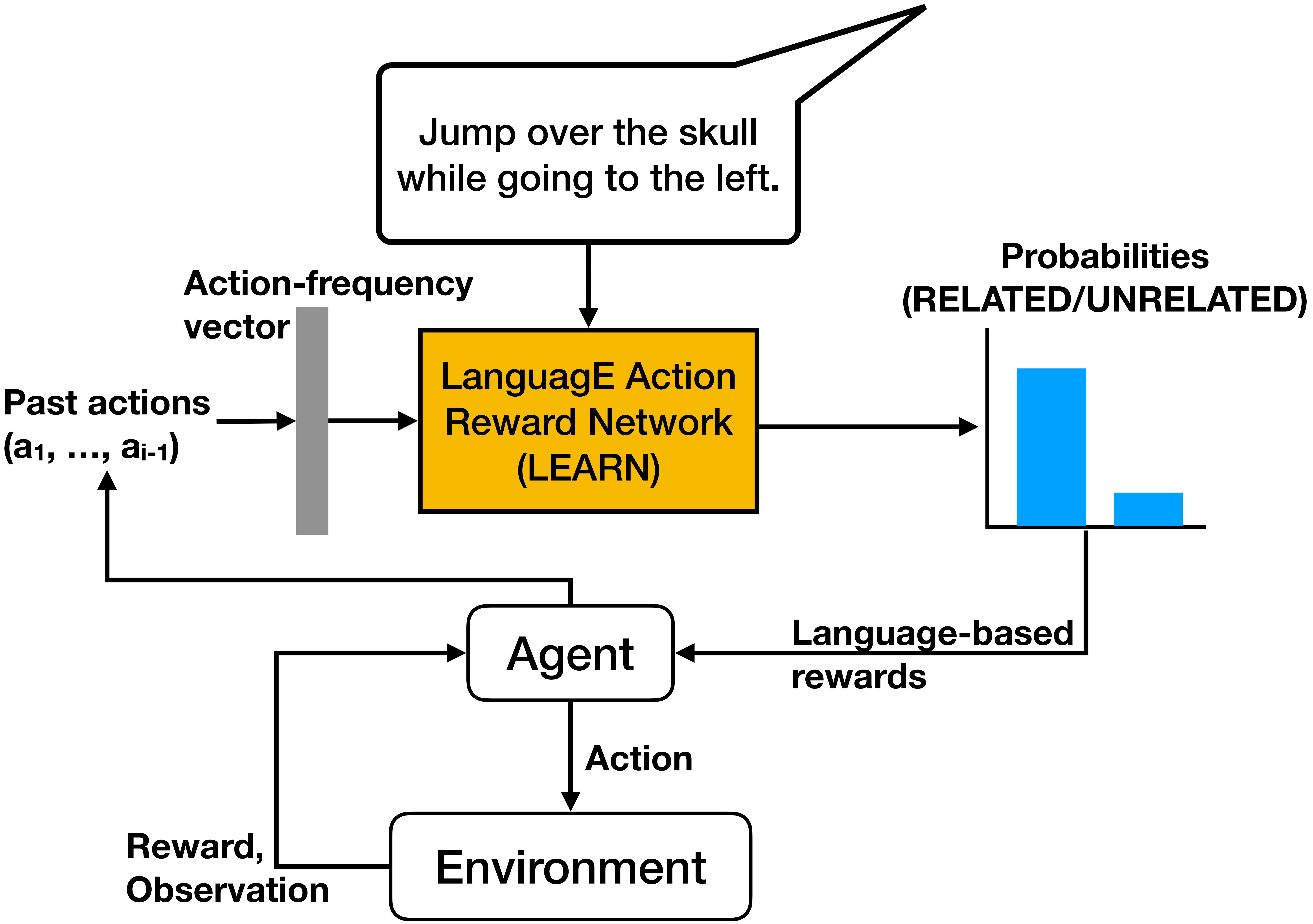
LanguagE Action Reward Network (LEARN): Motivation



Jump over the skull while going to the left.

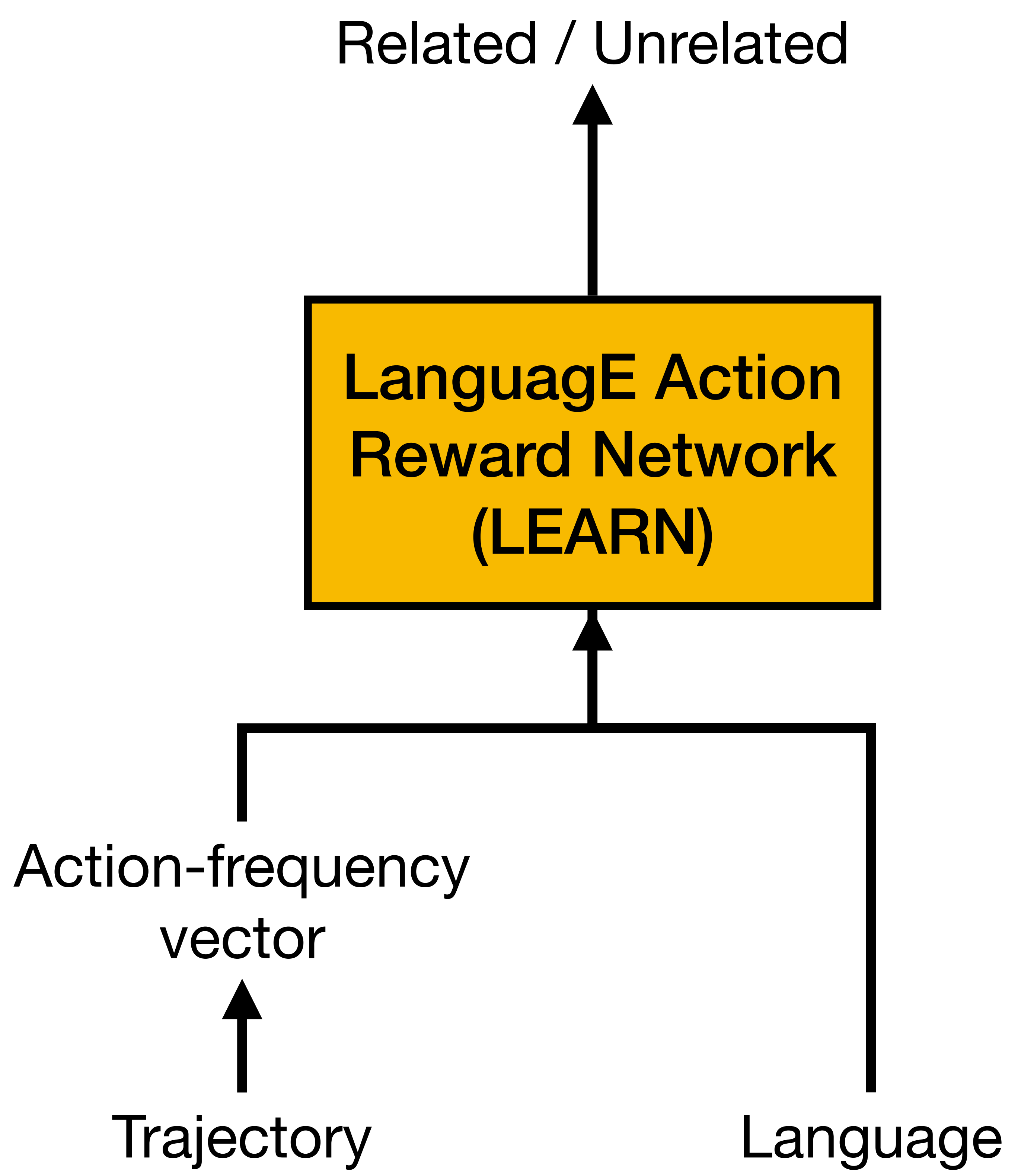
Can we use natural language to provide intermediate rewards to the agent?

LanguagE Action Reward Network (LEARN): Approach

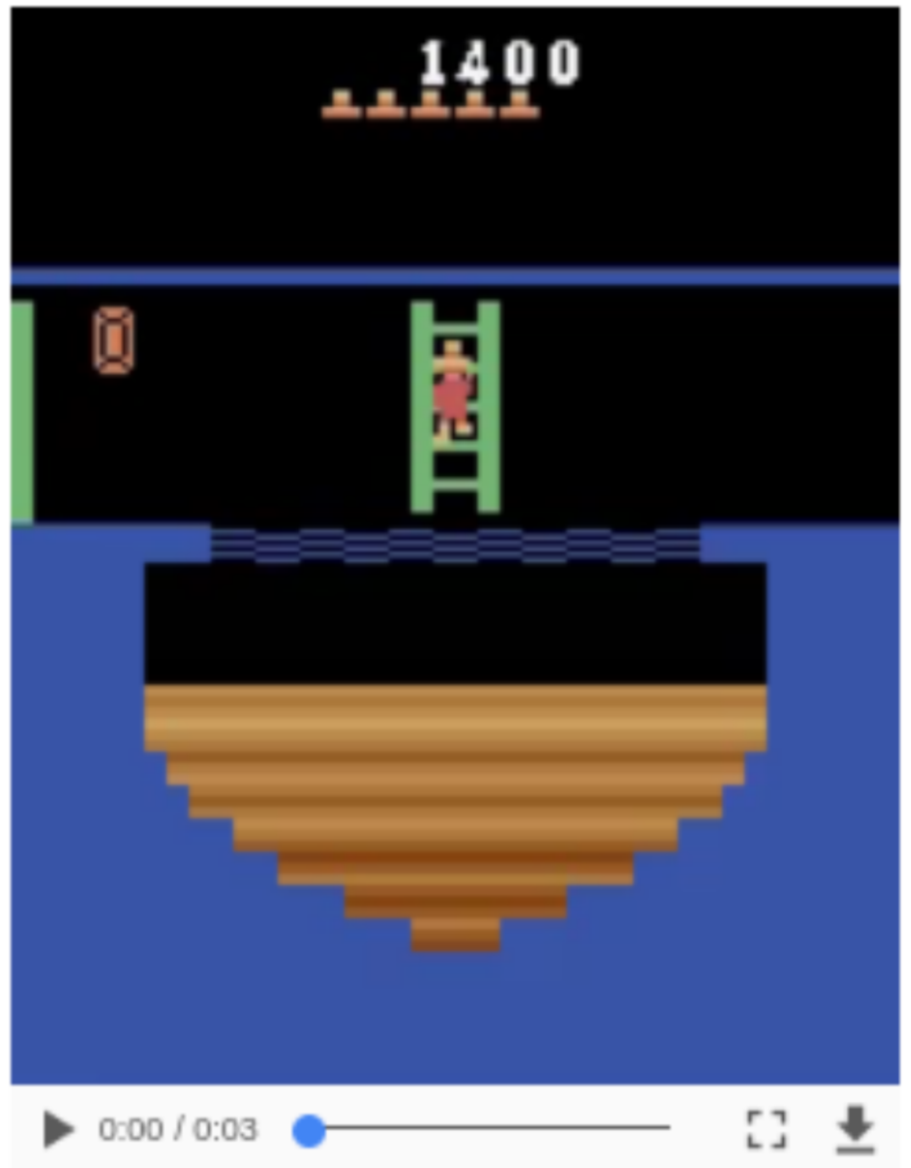


- Using the agent's past actions, generate an action-frequency vector.
- LEARN: Scores the relatedness between the action-frequency vector and the language command.
- Use the relatedness scores as language-based rewards.

LanguagE Action Reward Network (LEARN): Approach



Clip 1:



1.	wait
2.	using the ladder on standing
3.	going slow and climb down the ladder
4.	move down the ladder and walk left
5.	go left watch the trap and move on
6.	climbing down the ladder
7.	ladder dwon and running this away
8.	stay in place on the ladder.
9.	go down the ladder
10.	go right and climb up the ladder
11.	just jump and little move to right side
12.	run all the way to the left.
13.	go left jumping once
14.	go left
15.	move right and jump over green creature then go down the ladder
16.	hop over to the middle ledge
17.	wait for the two skulls and dodge them in the middle
18.	walk to the left and then jump down
19.	jump to collected gold coin and little move
20.	wait for the platform to materialize then walk and leap to your right to collect the coins.

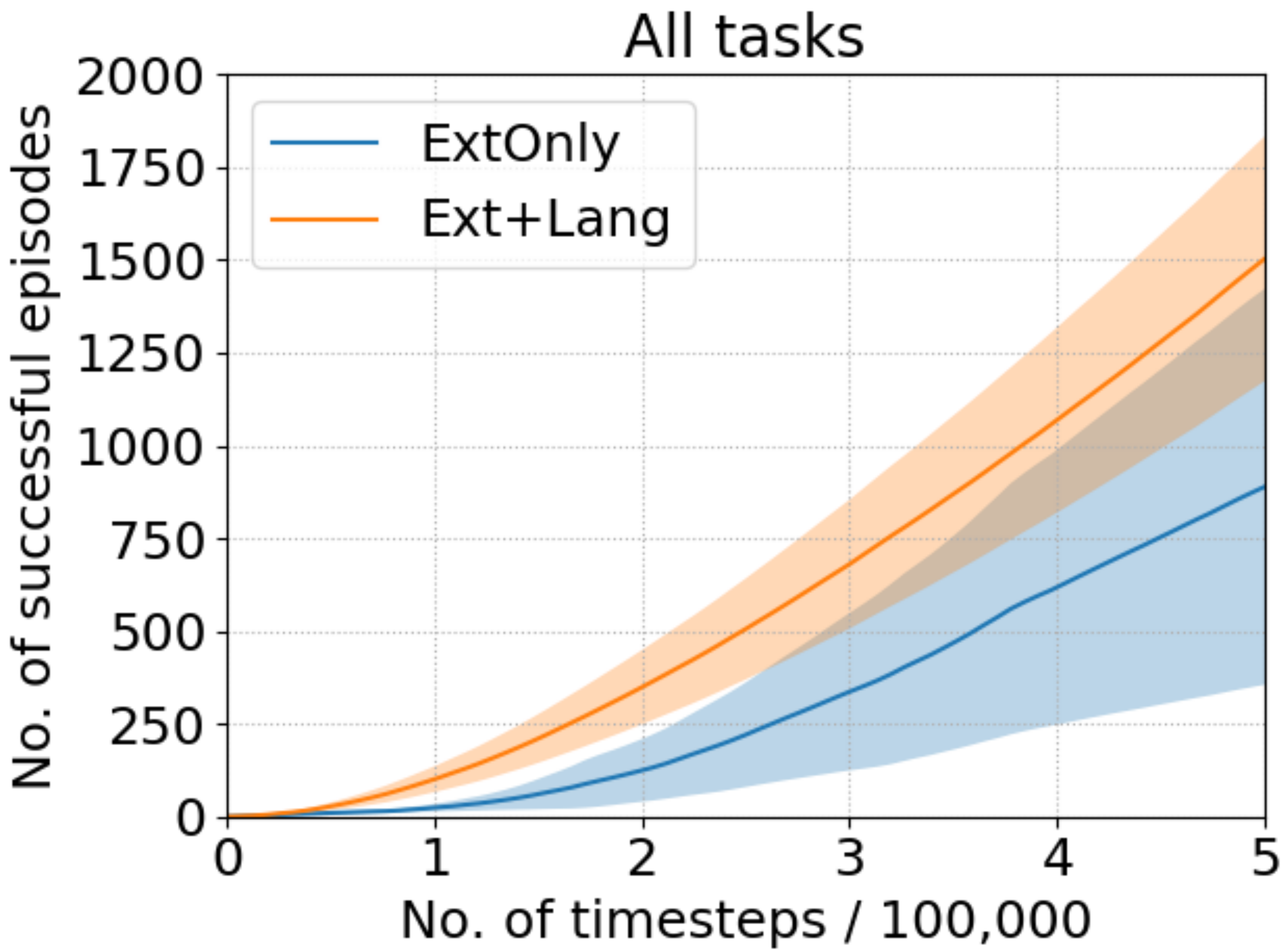
Please enter the description below:

LanguagE Action Reward Network (LEARN): Results

- Compared RL training using PPO algorithm with and without language-based reward.

- ExtOnly: Reward of 1 for reaching the goal, reward of 0 in all other cases.

- Ext+Lang: Extrinsic reward plus language-based intermediate rewards.



Talk Outline

Background

Core Contributions:

Sequential Decision Making

	Challenge	Solution
RL	Reward design	Language-based Rewards
IL		

+

LEARN

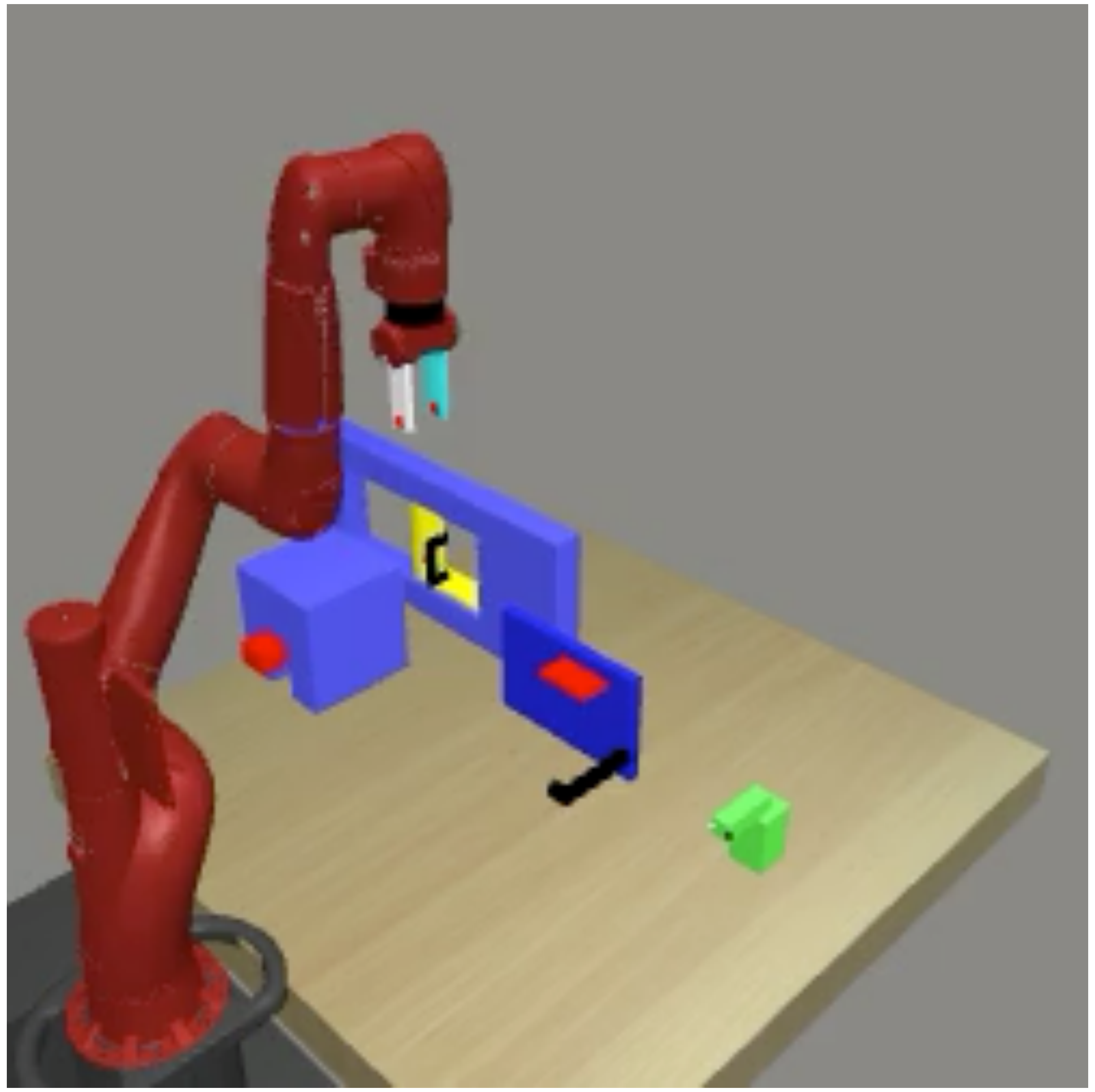
PixL2R

Future Directions

[Yu et al., 2020]

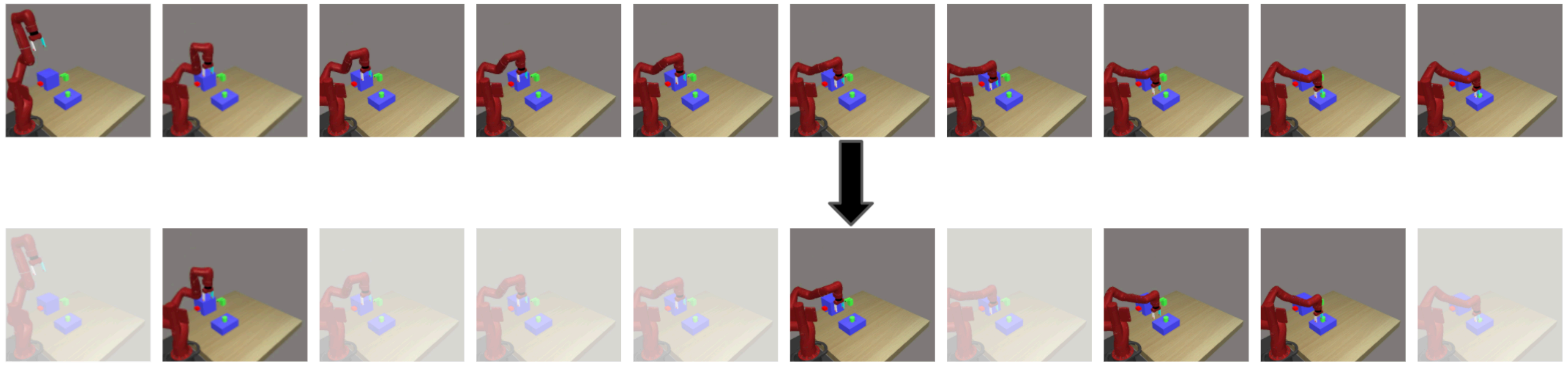
Pixels and Language to Rewards (PixL2R): Motivation

LEARN results in efficient policy learning, but
- the action-frequency vector is undefined for continuous action spaces
- discards temporal information in action sequences
- does not use state information

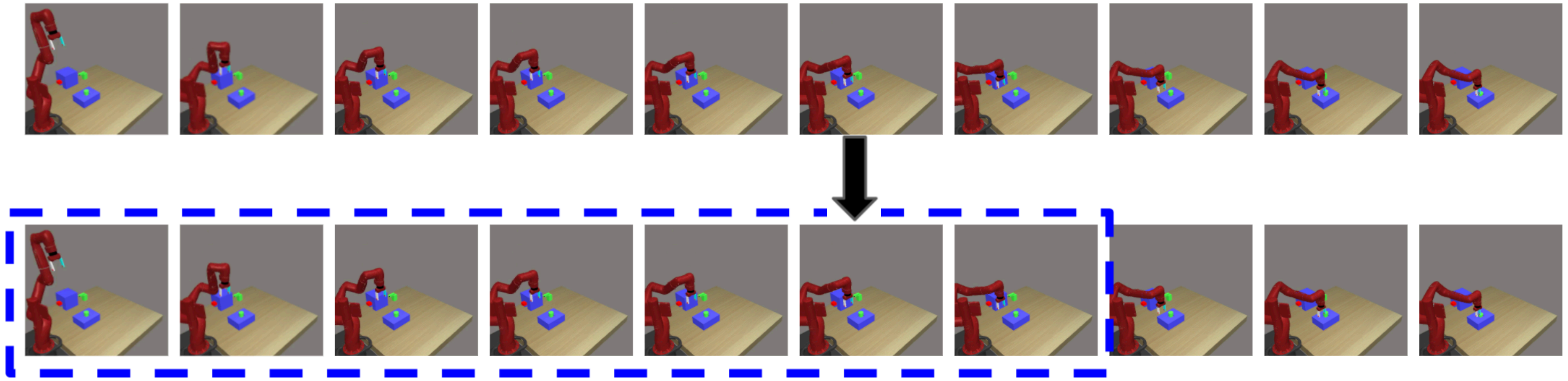


Pixels and Language to Rewards (PixL2R): Data Augmentation

- Frame dropping



- Partial trajectories



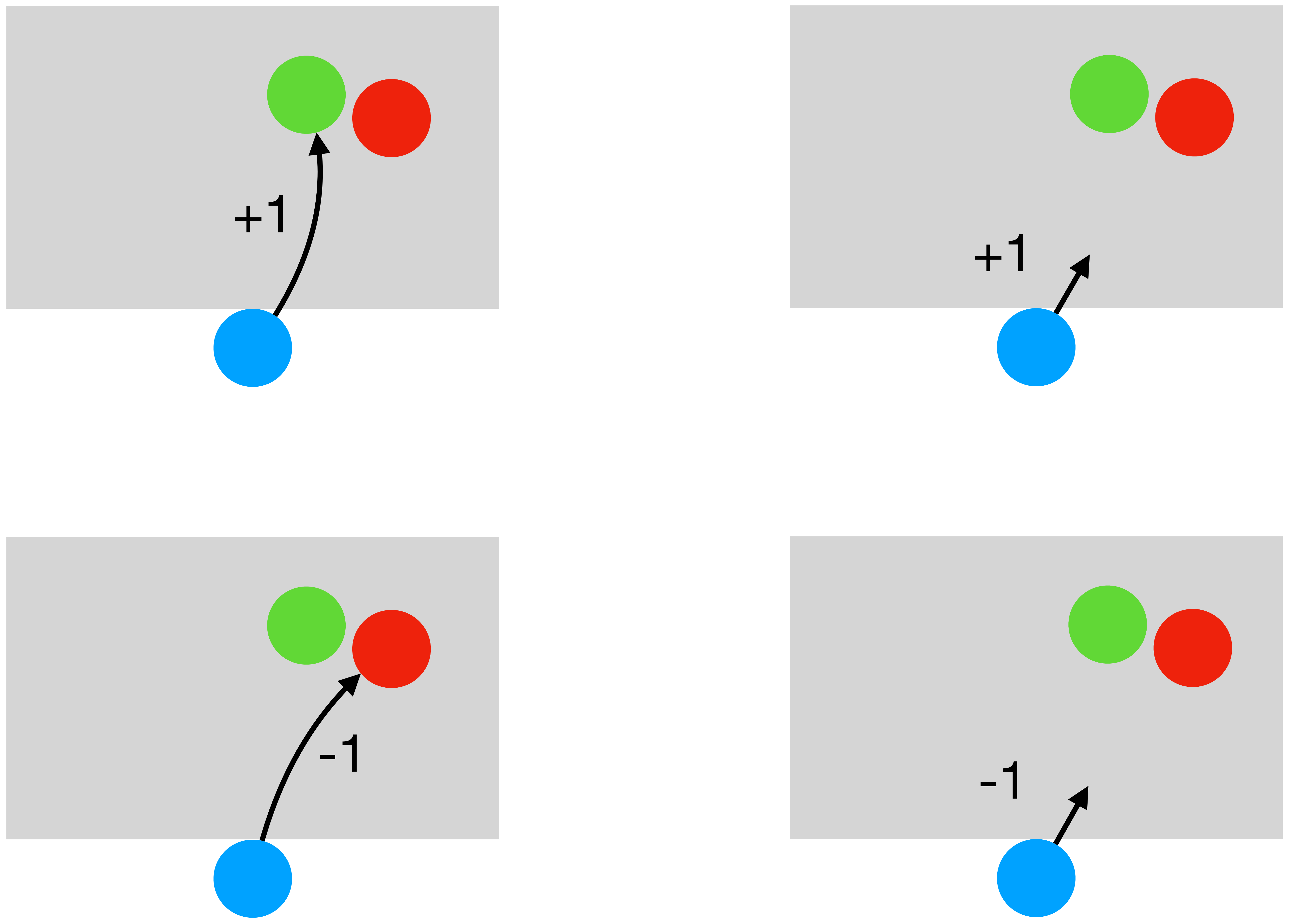
Pixels and Language to Rewards (PixL2R): Training Objective

Classification:

Starting position

Correct object

Incorrect object



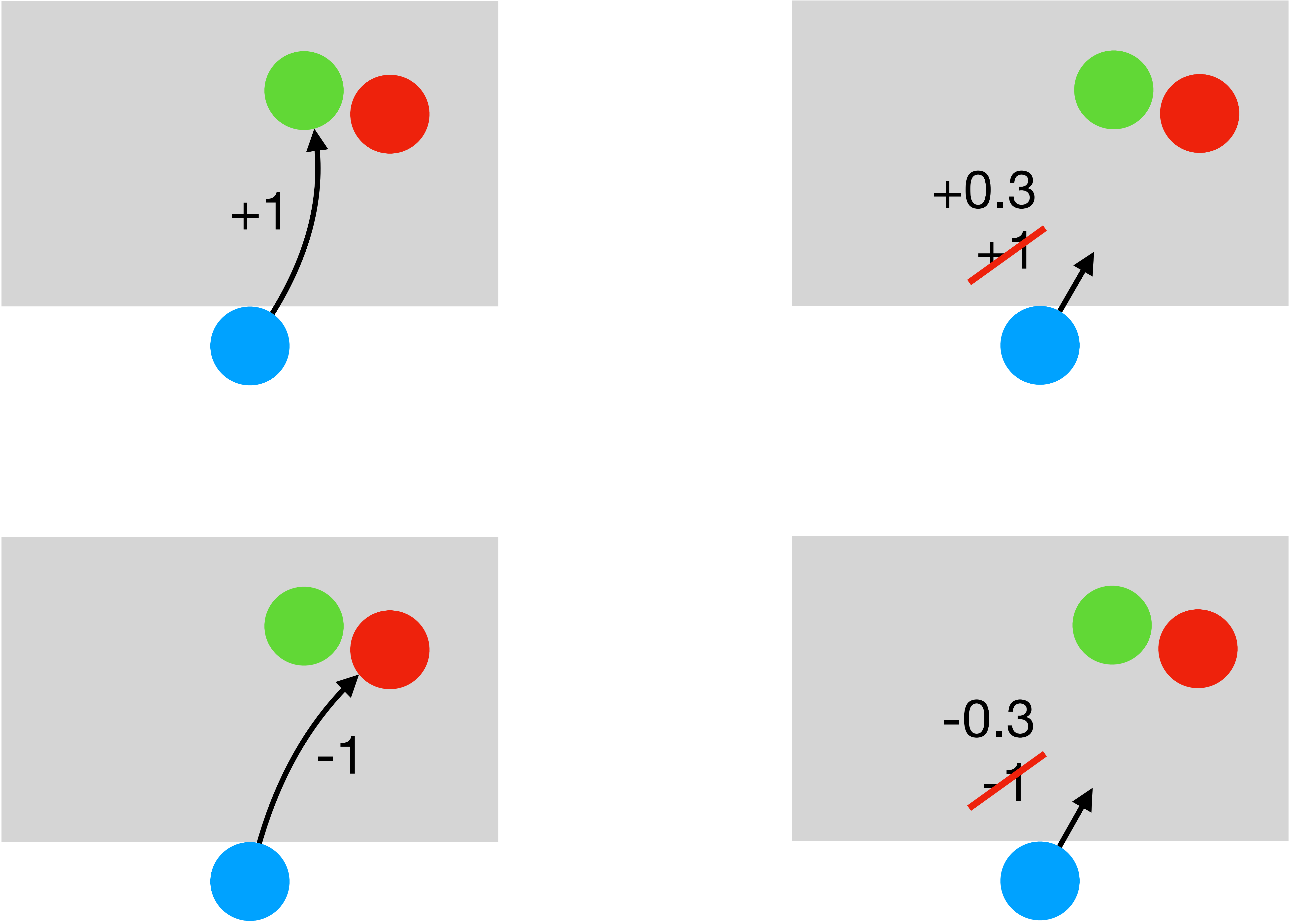
Pixels and Language to Rewards (PixL2R): Training Objective

~~Classification:~~
Regression:

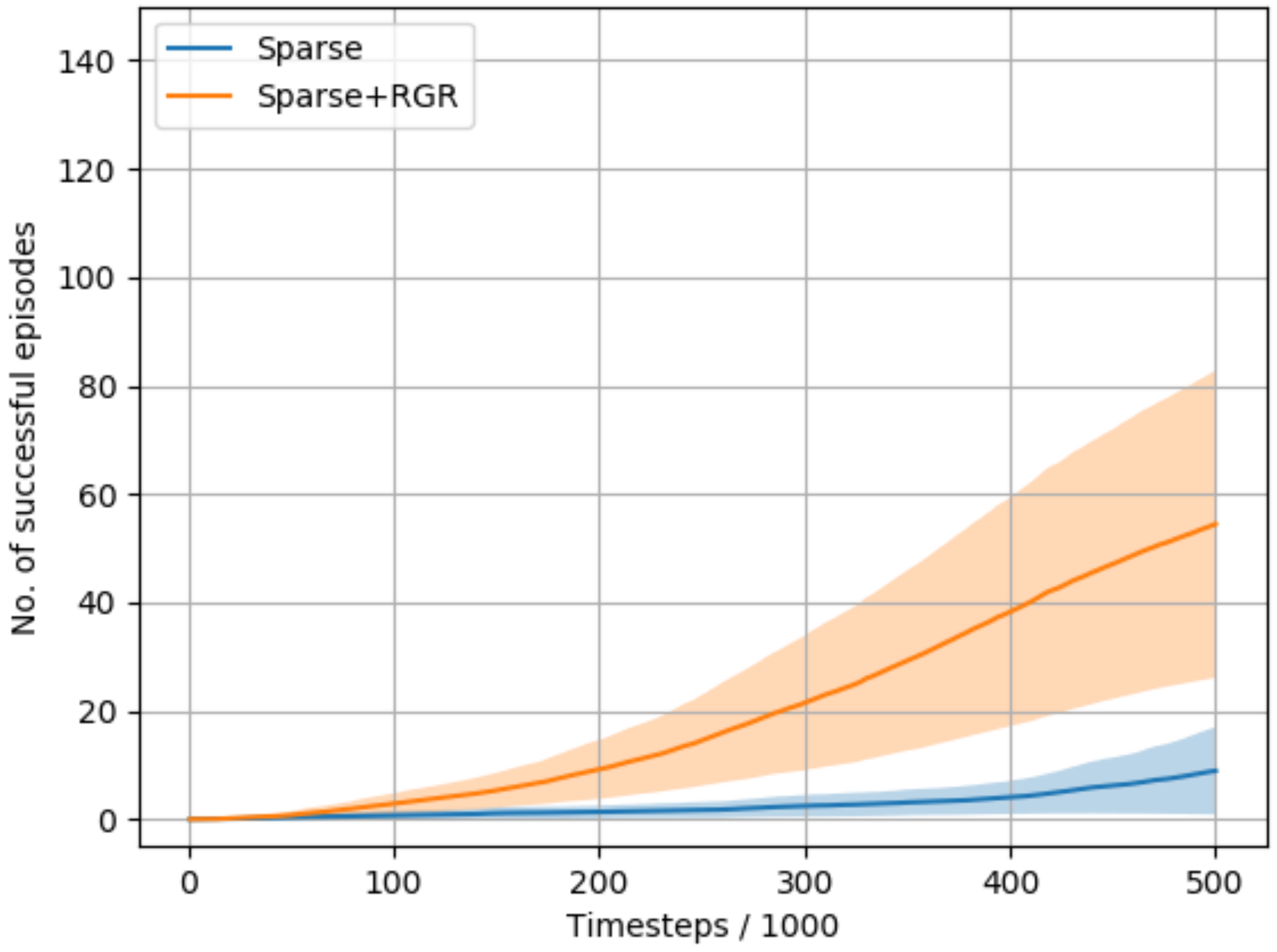
Starting position

Correct object

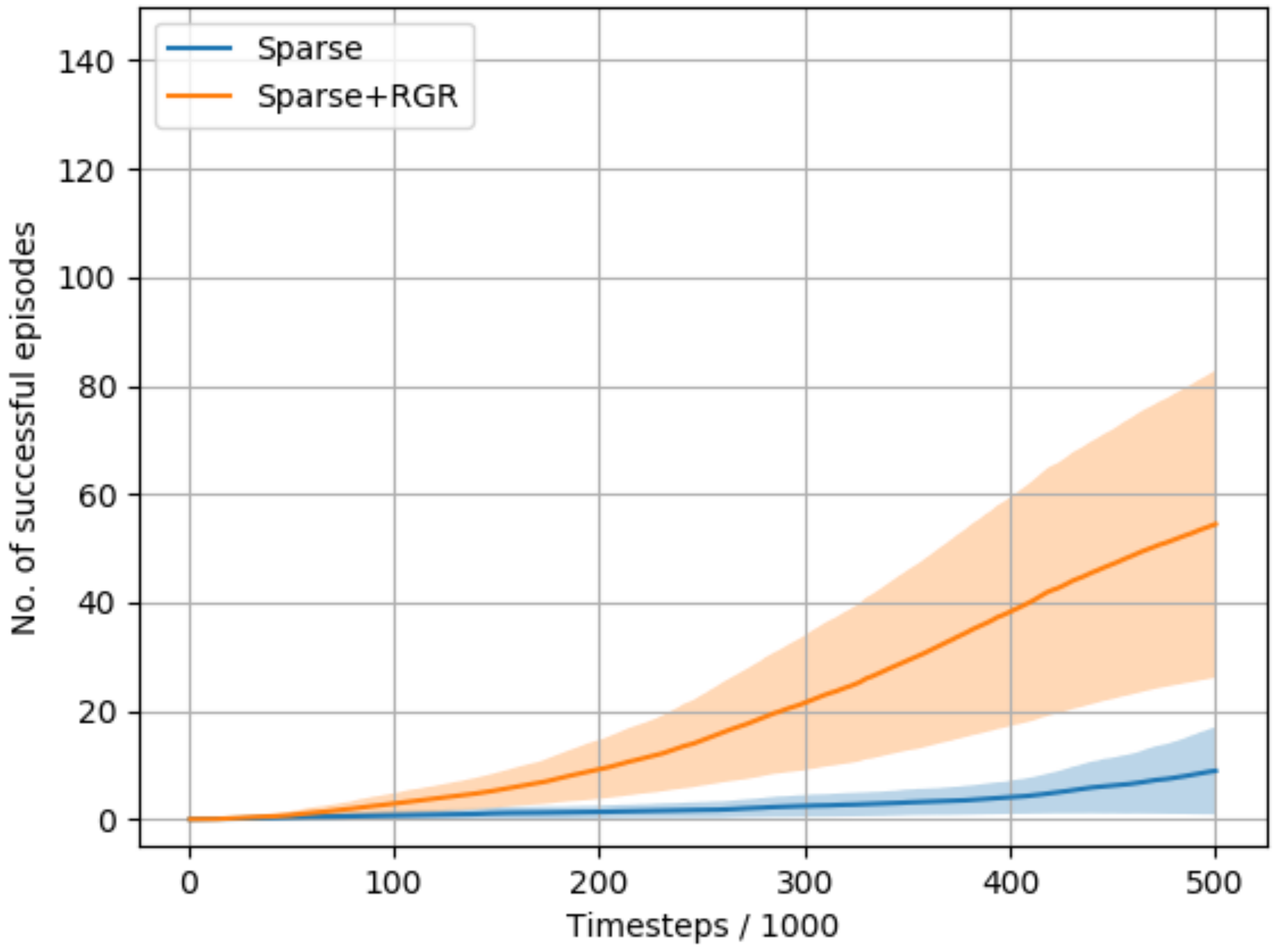
Incorrect object

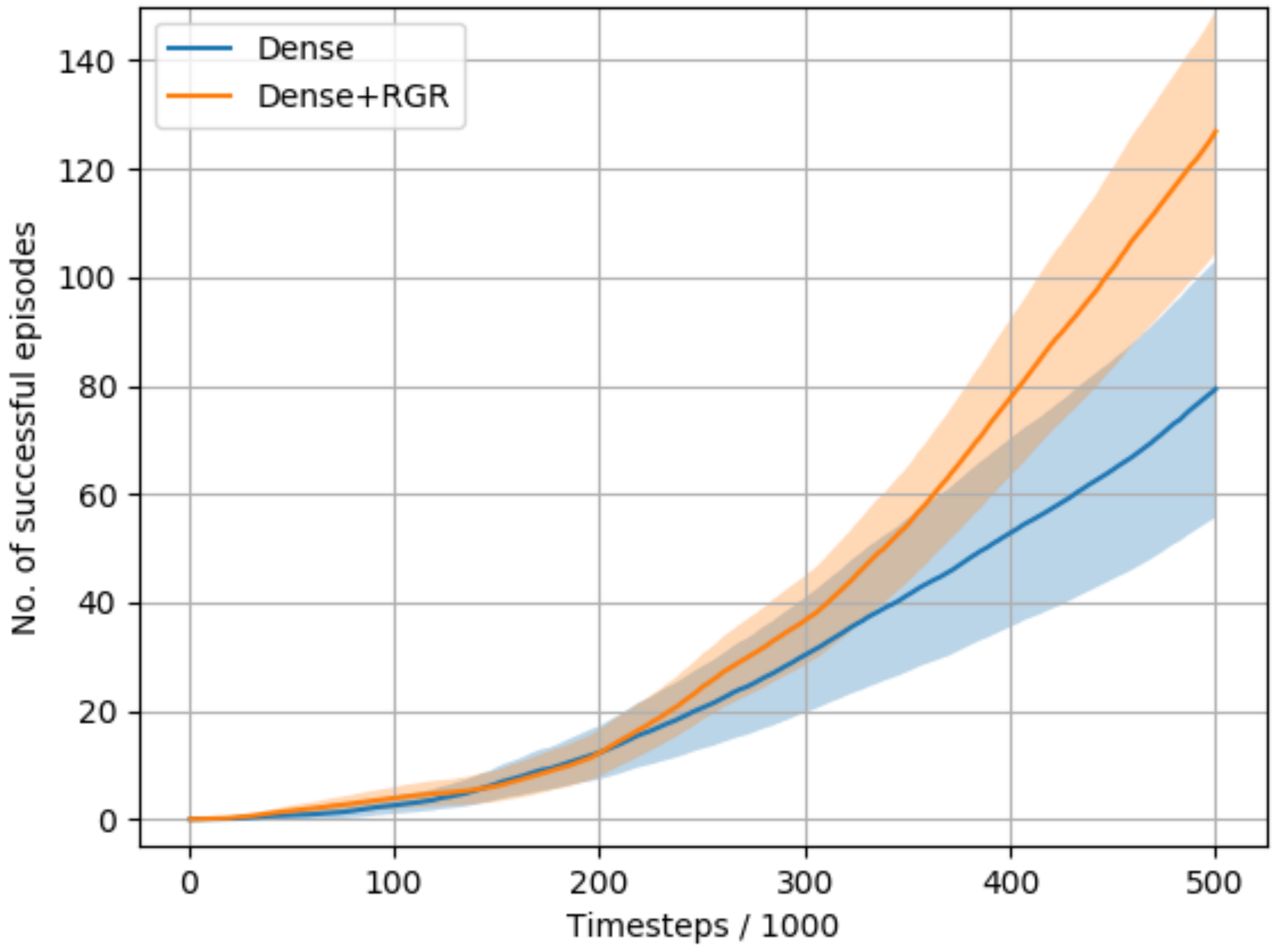


Pixels and Language to Rewards (PixL2R): Results

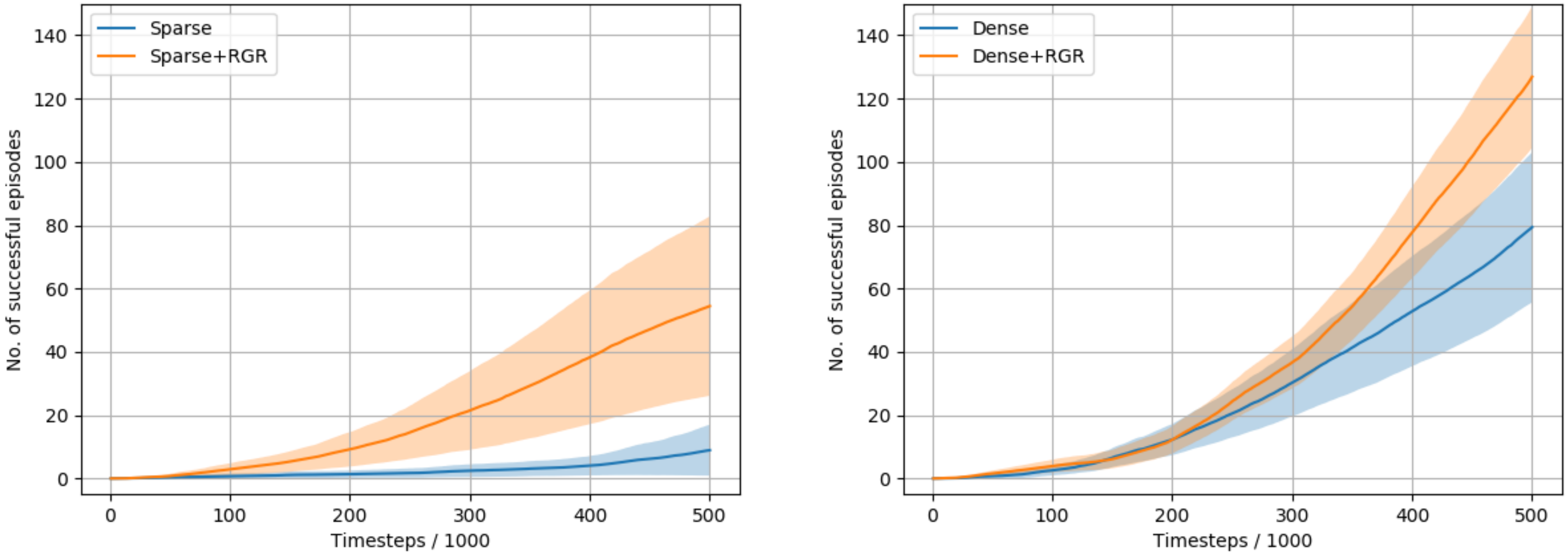


Pixels and Language to Rewards (PixL2R): Results





Pixels and Language to Rewards (PixL2R): Results



New RL training regime: Coarse hand-designed dense rewards + Language-based rewards

Talk Outline

Background

Core Contributions:

Sequential Decision Making

	Challenge	Solution
RL	Reward design	Language-based Rewards
IL		

LEARN

PixL2R

Future Directions

Imitation Learning

3 broad classes of approaches:

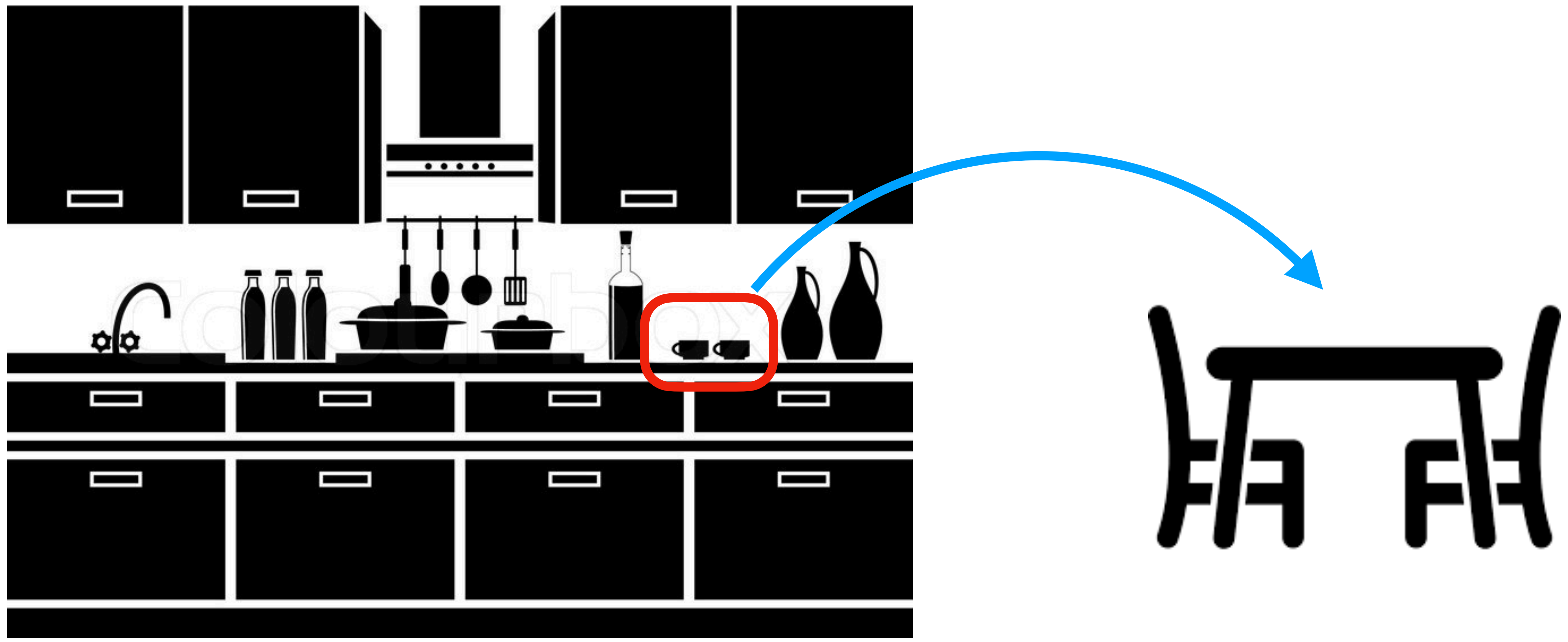
1. Behavior cloning:
   - Supervised Learning approach
   - Use state-action pairs in the demonstrations

2. Inverse Reinforcement Learning (IRL):
   - Infer a reward function from demonstrations
   - Use RL to learn a policy

3. Adversarial Imitation Learning (AIL):
   - Generator-discriminator-based approach

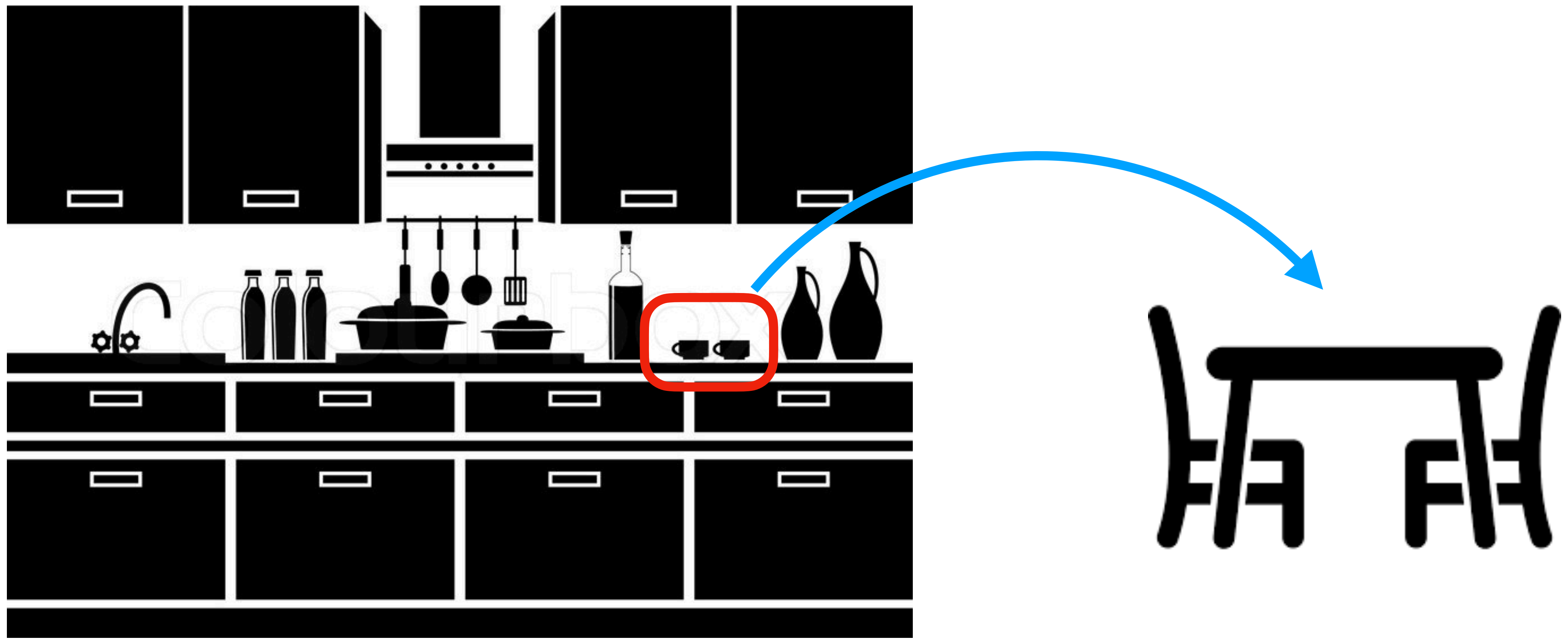
Challenge in IL: A lot of demonstrations are needed

- Multiple demonstrations are often needed to specify a task.



Challenge in IL: A lot of demonstrations are needed

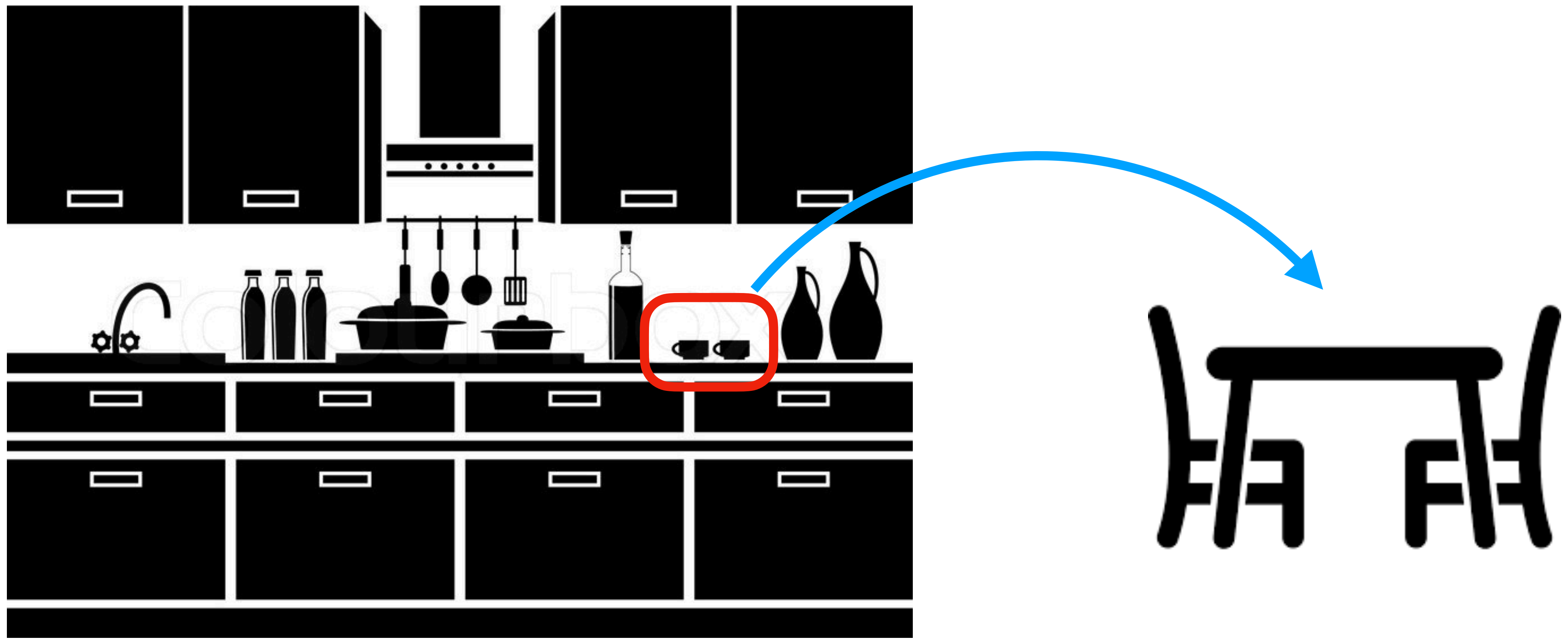
- Multiple demonstrations are often needed to specify a task.



- A new set of demonstrations is needed for each new task.

Challenge in IL: A lot of demonstrations are needed

- Multiple demonstrations are often needed to specify a task.

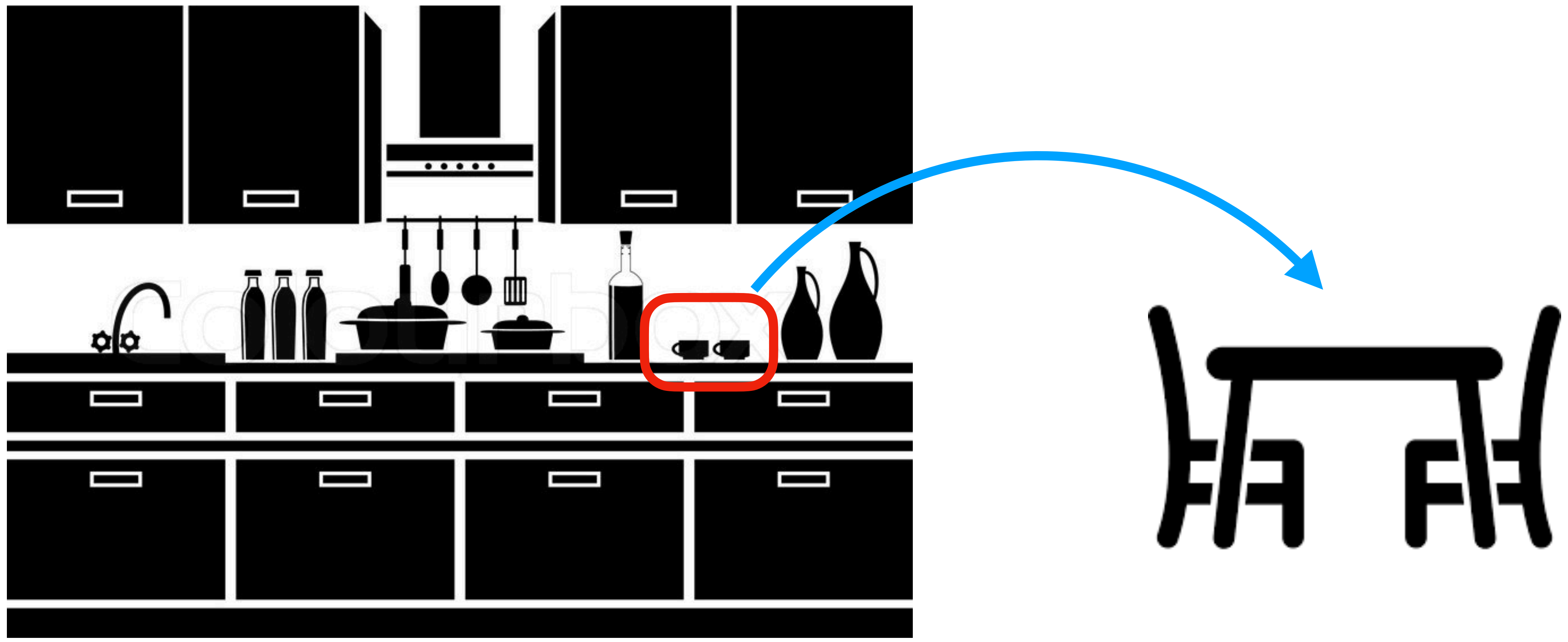


A lot of demonstrations may need to be given to teach multiple tasks.

- A new set of demonstrations is needed for each new task.

Challenge in IL: A lot of demonstrations are needed

- Multiple demonstrations are often needed to specify a task.



→ A lot of demonstrations may need to be given to teach multiple tasks.

→ **Use natural language to reuse demos from related tasks**

- A new set of demonstrations is needed for each new task.

Talk Outline

Background

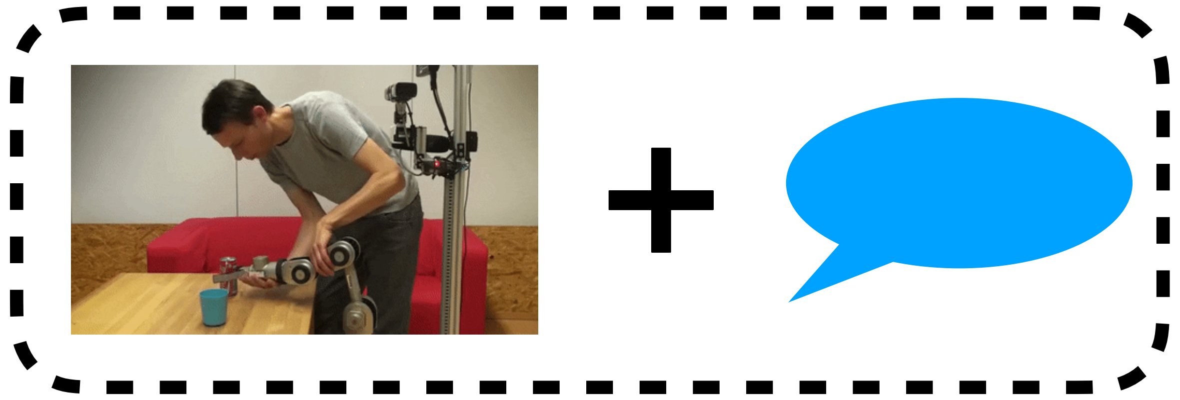
Core Contributions:

Sequential Decision Making

	Challenge	Solution
RL	Reward design	Language-based Rewards
IL	Many demos needed	Language-guided Task Adaptation

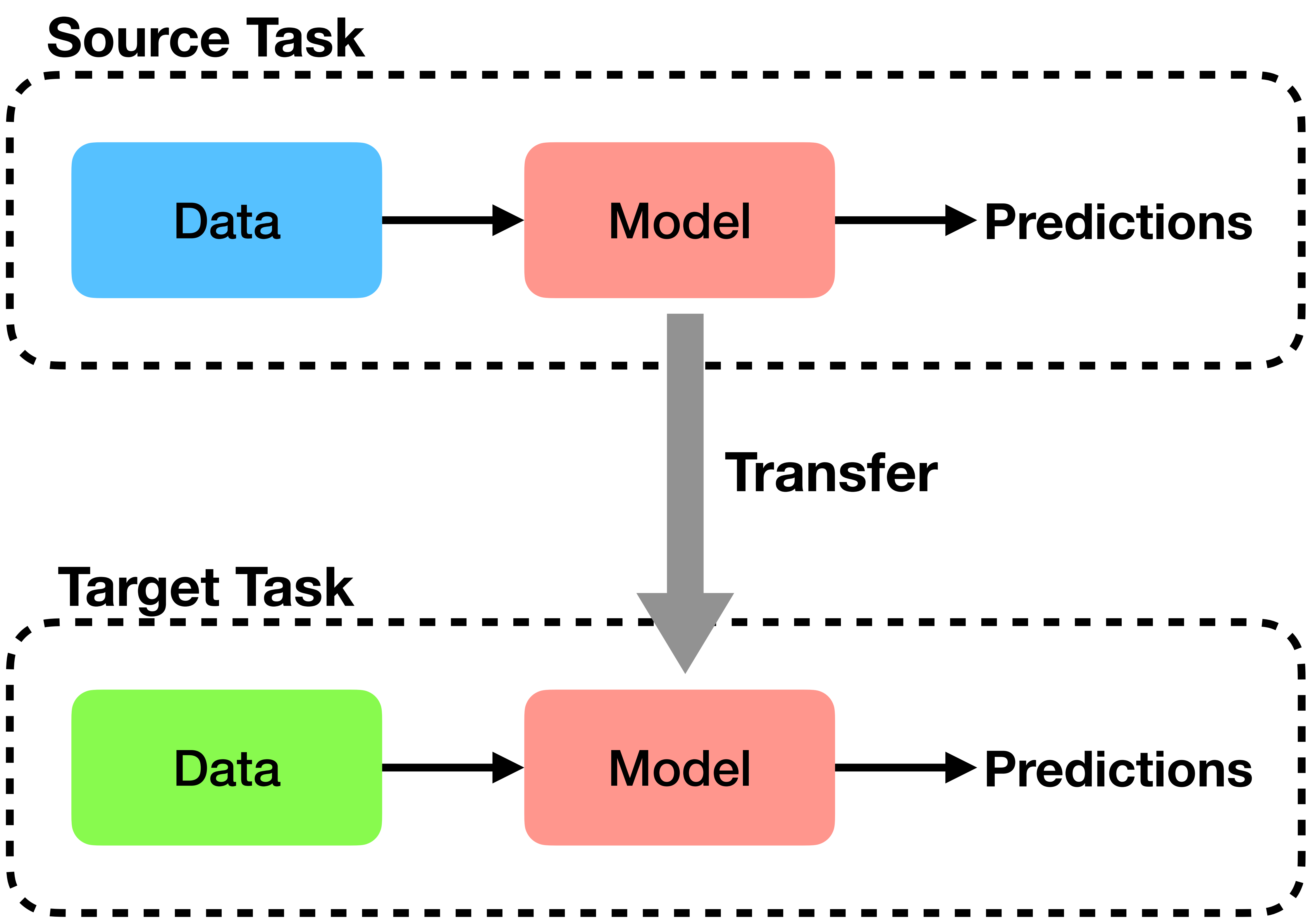
LEARN

PixL2R



Future Directions

Related Work: Transfer Learning



Talk Outline

Background

Core Contributions:

Sequential Decision Making

	Challenge	Solution
RL	Reward design	Language-based Rewards
IL	Many demos needed	Language-guided Task Adaptation

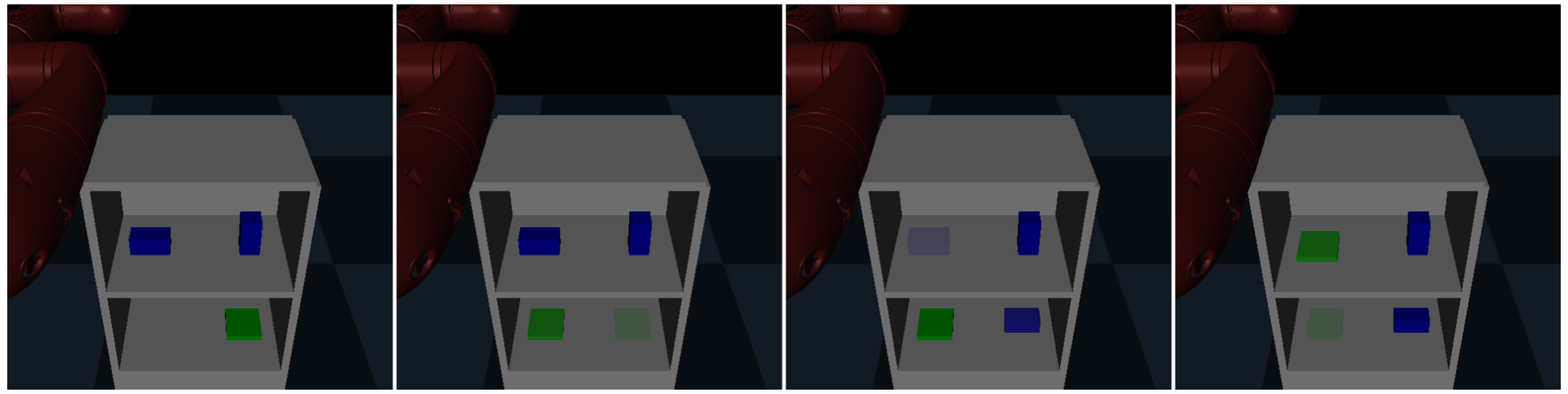
+ → LEARN, PixL2R

+ → LARVA

Future Directions

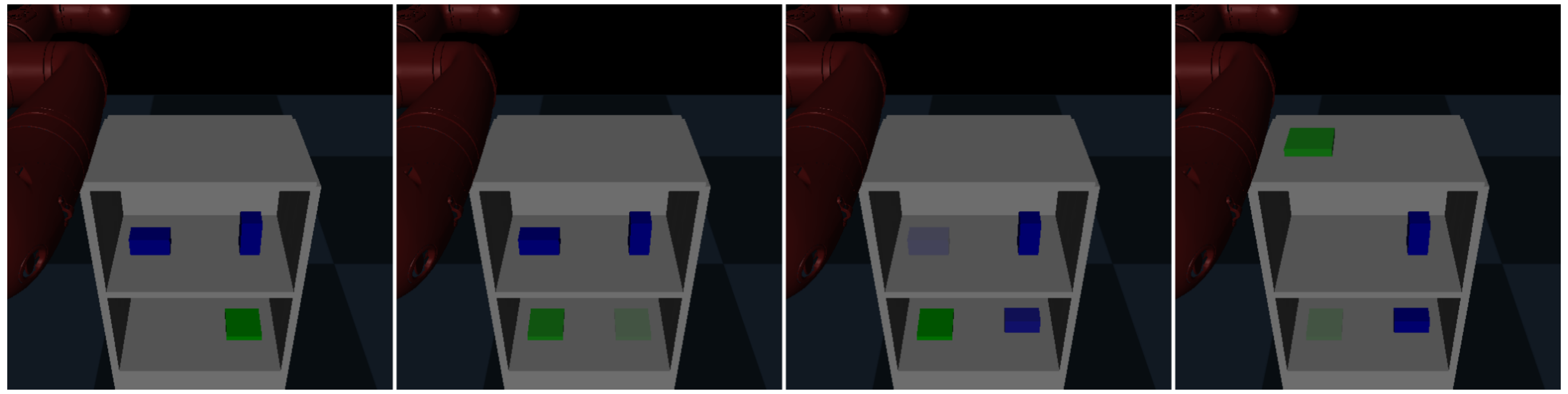
Language-Aided Reward and Value Adaptation (LARVA): Motivation

Source task



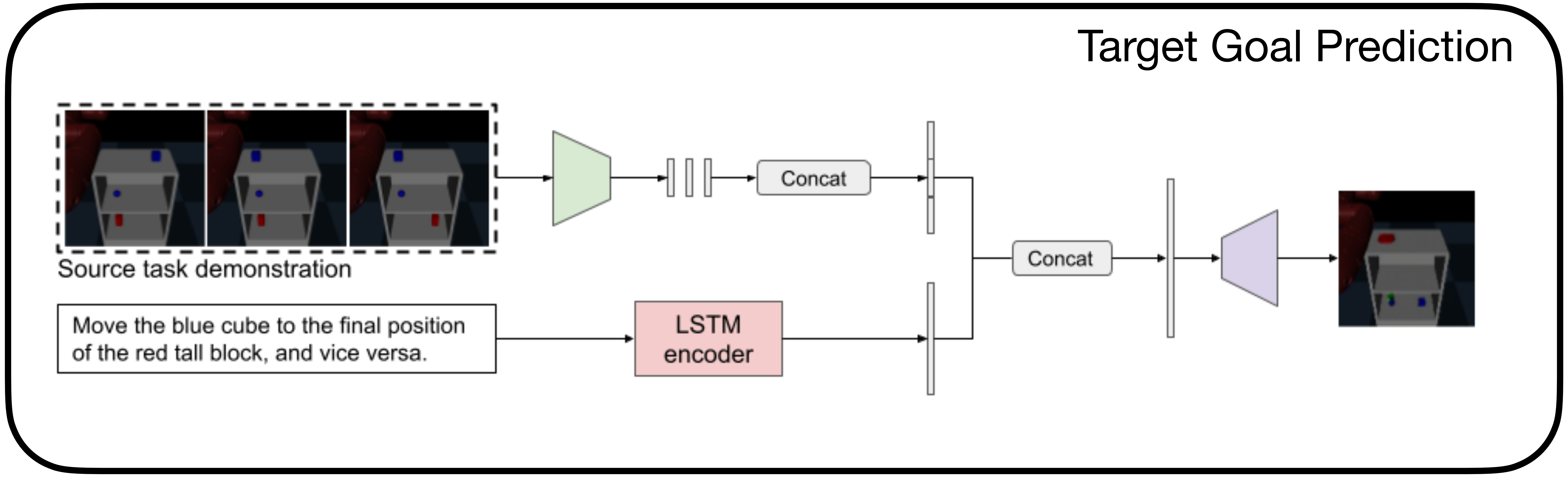
In the third step, move the green flat block from bottom left to top left.

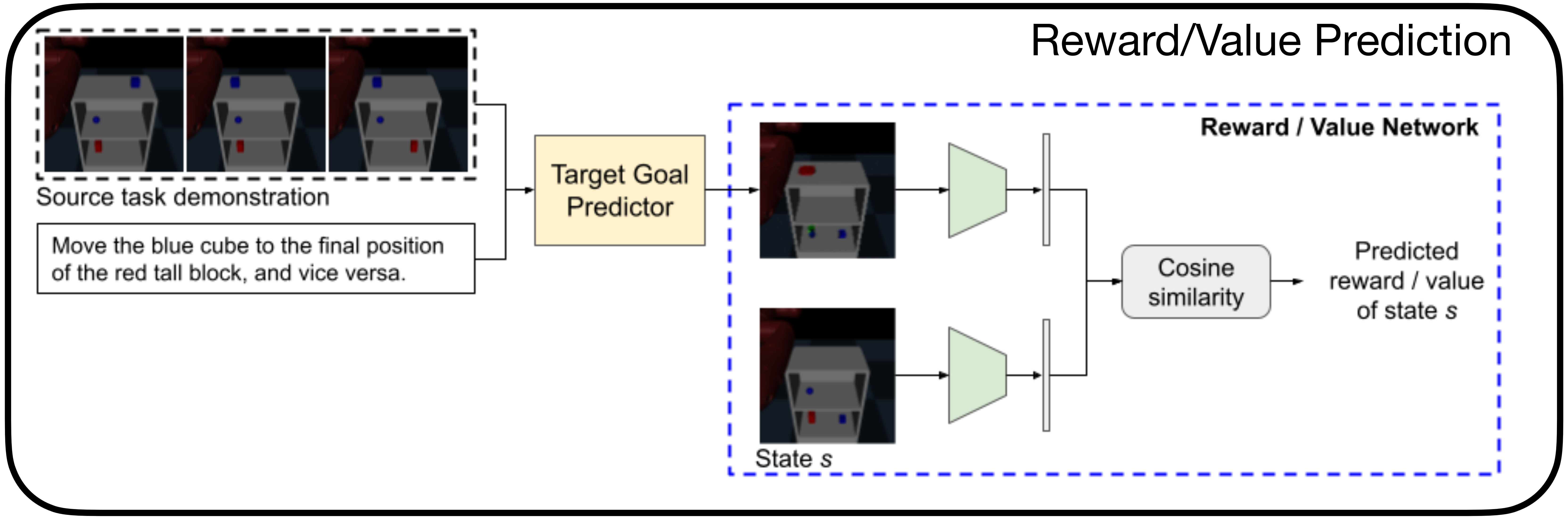
Target task



GOAL:
Learn the target task *without any demonstrations*.

Language-Aided Reward and Value Adaptation (LARVA): Approach



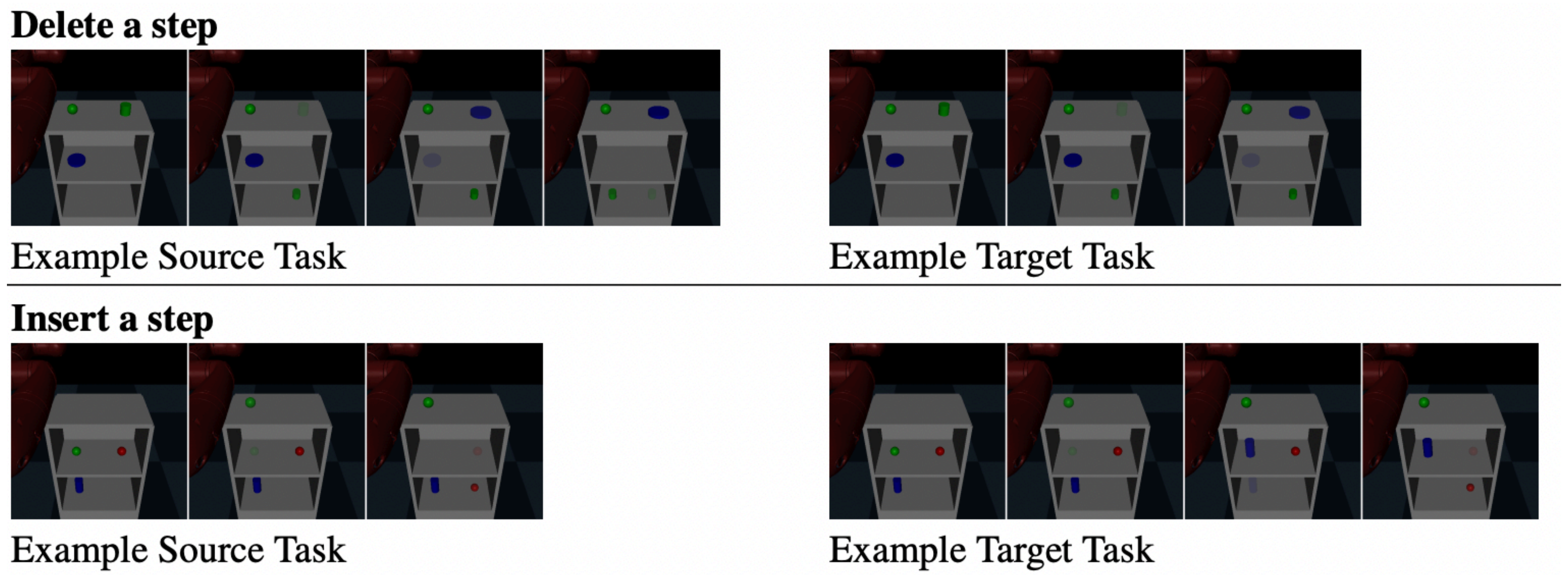


Training data:
- Source demo
- Language
- Target goal
- Target reward / value function

Loss functions:
- Reward / Value prediction: Mean-squared error
- Target goal prediction: Mean-squared error

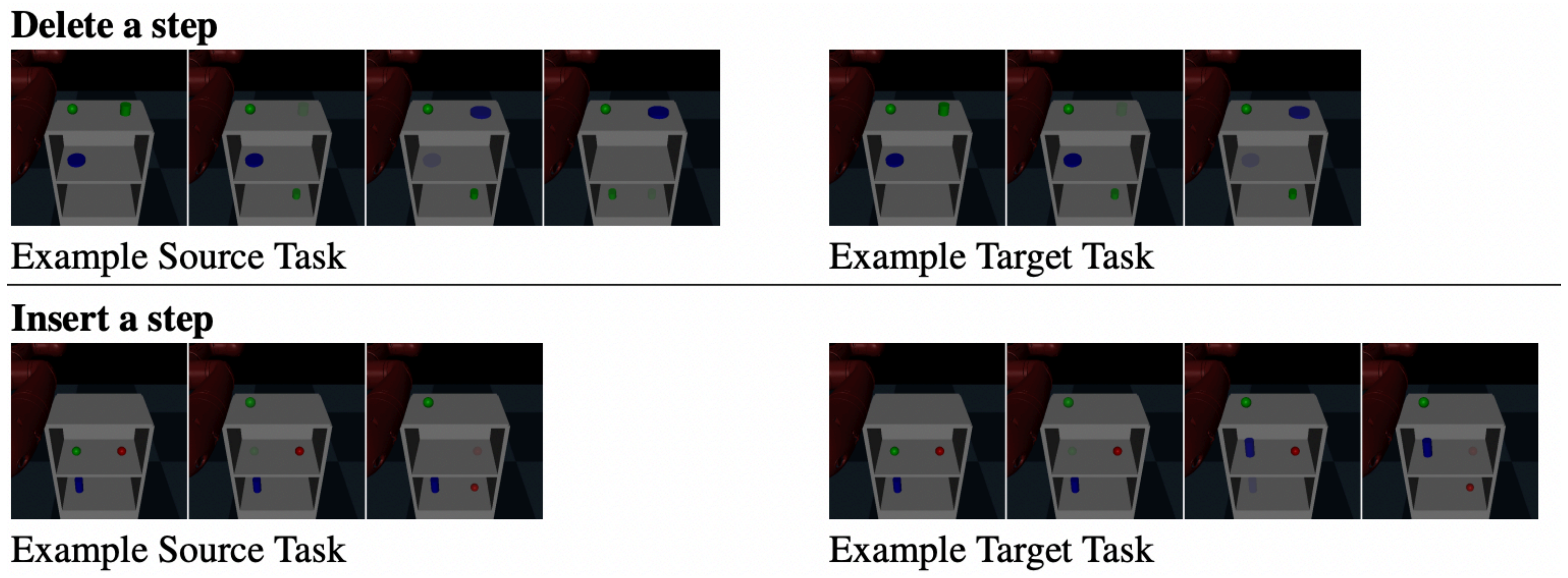
Language-Aided Reward and Value Adaptation (LARVA): Experiments

6 types of adaptations: e.g. add a step, delete a step, etc.



Language-Aided Reward and Value Adaptation (LARVA): Experiments

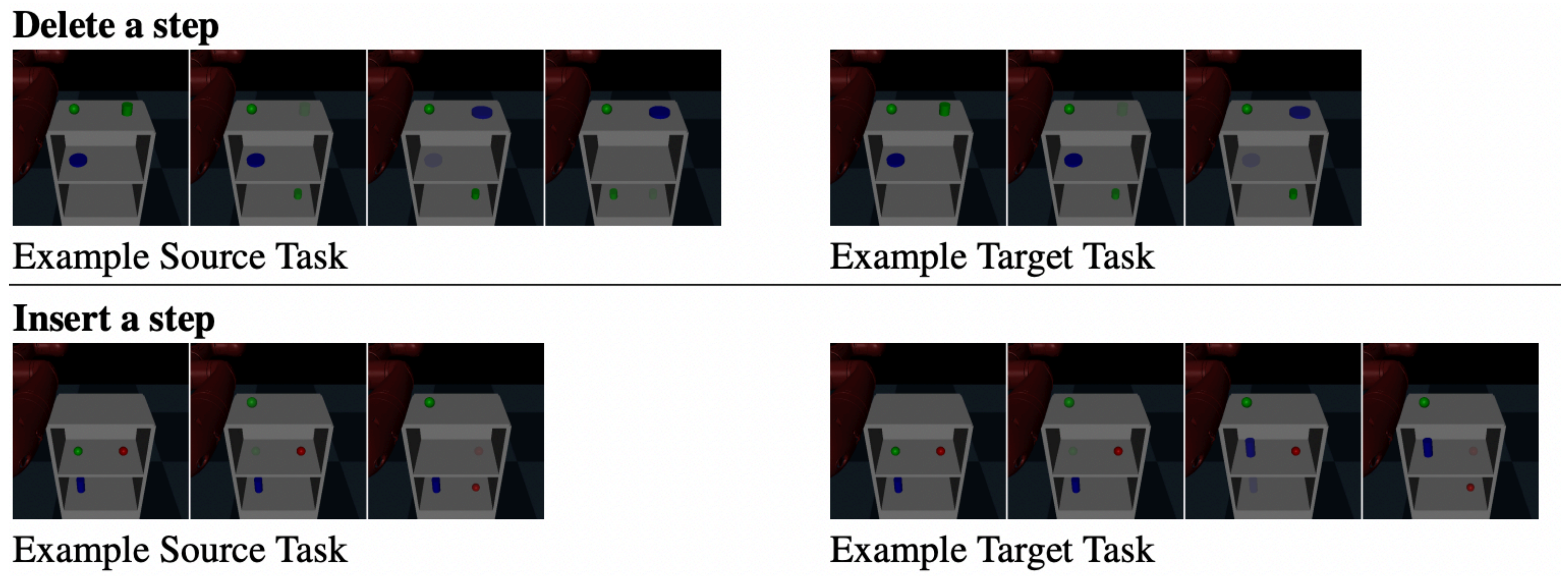
6 types of adaptations: e.g. add a step, delete a step, etc.



Language Data:
- Template-based
- Paraphrases from Amazon Mechanical Turk

Language-Aided Reward and Value Adaptation (LARVA): Experiments

6 types of adaptations: e.g. add a step, delete a step, etc.



Language Data:

- Template-based
- Paraphrases from Amazon Mechanical Turk

Evaluation:

$$s_1 \longrightarrow R_1$$

$$s_2 \longrightarrow R_2$$

$$\cdots$$

$$s_i \longrightarrow R_i$$

Max value

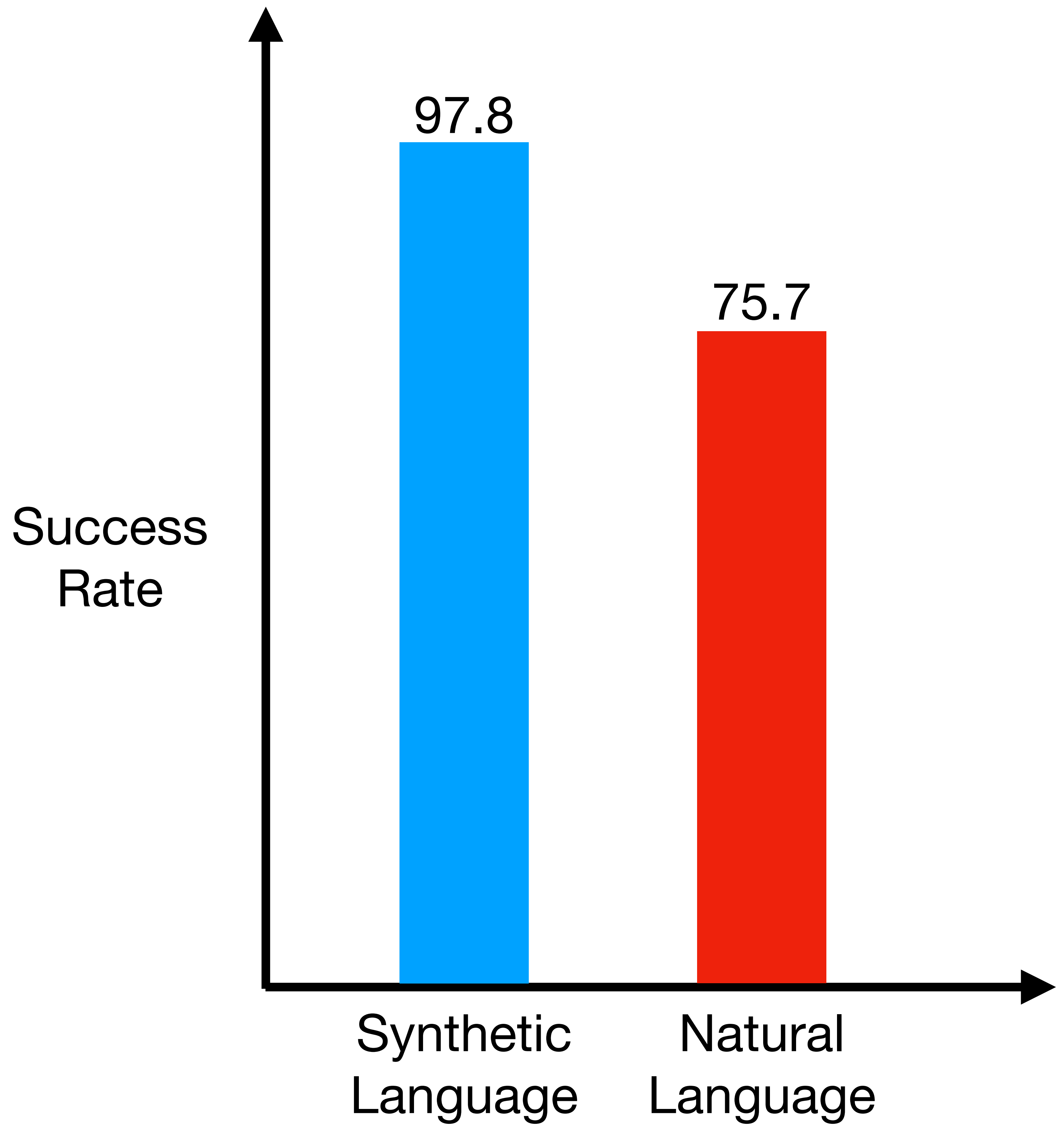
$$\cdots$$

$$s_N \longrightarrow R_N$$

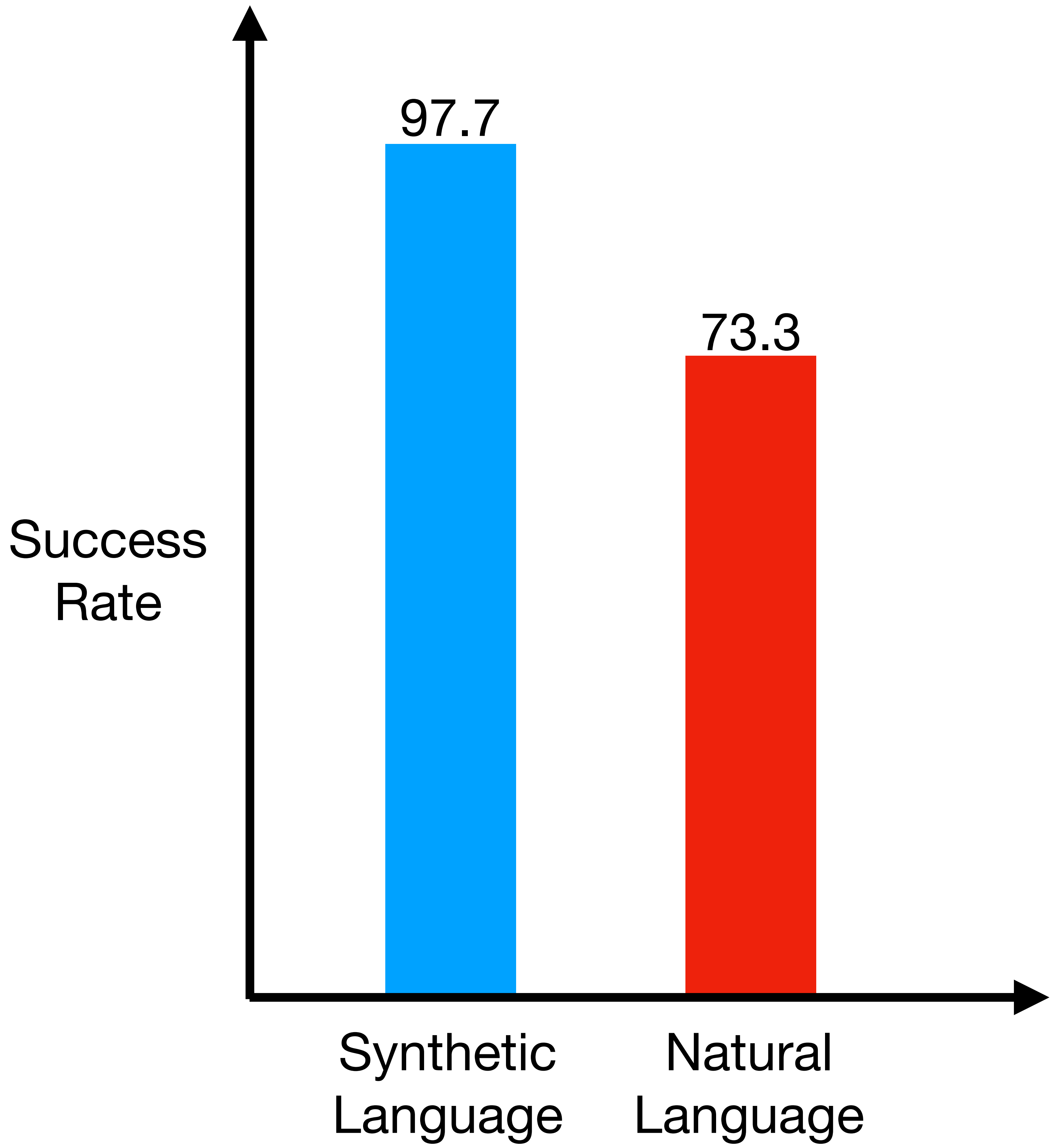
If s_i is the true goal state for the target task, then success=1 else 0.

Language-Aided Reward and Value Adaptation (LARVA): Experiments

Reward Prediction



Value Prediction



Talk Outline

Background

Core Contributions:

Sequential Decision Making

	Challenge	Solution
RL	Reward design	Language-based Rewards
IL	Many demos needed	Language-guided Task Adaptation

+

LEARN

PixL2R

+

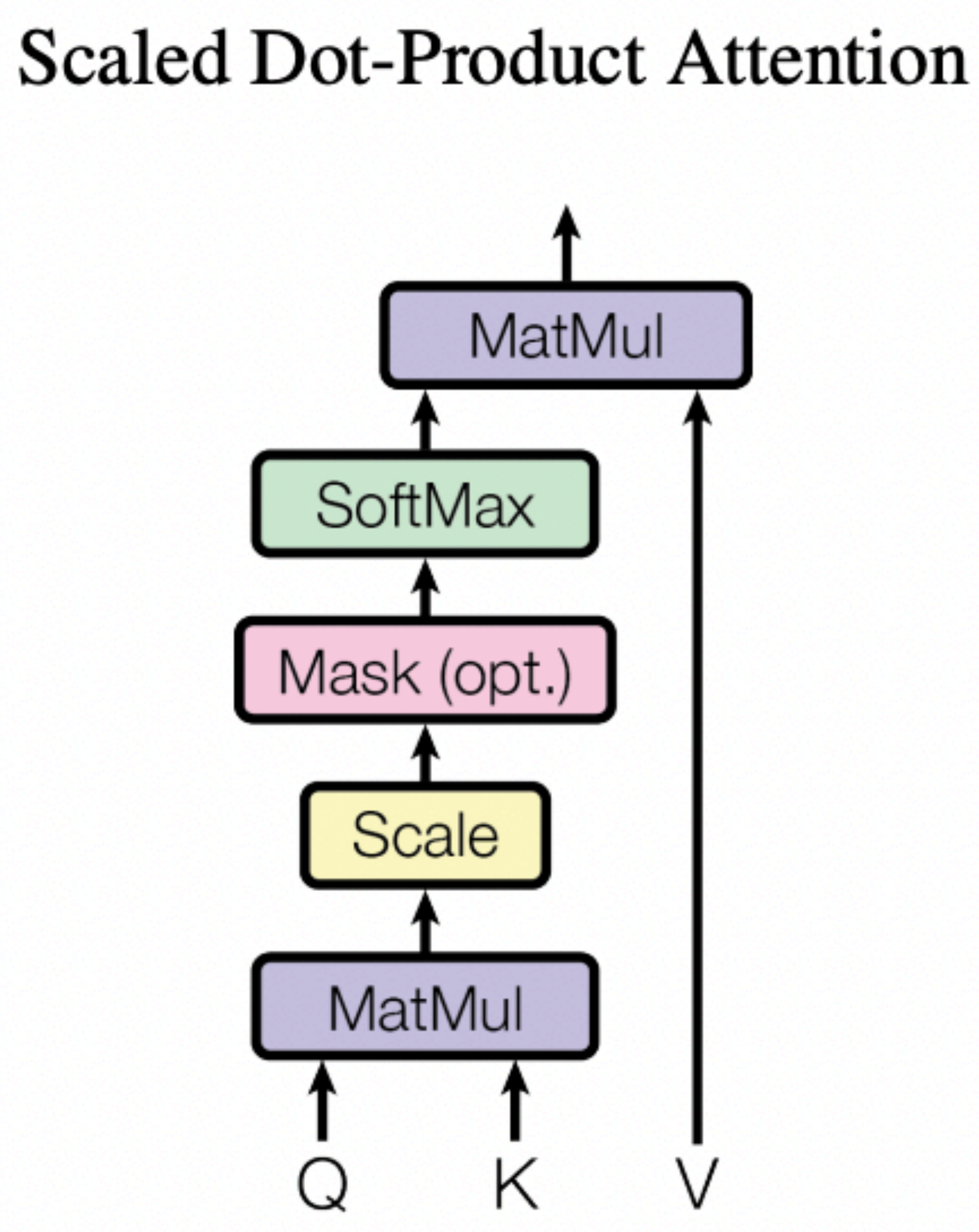
LARVA

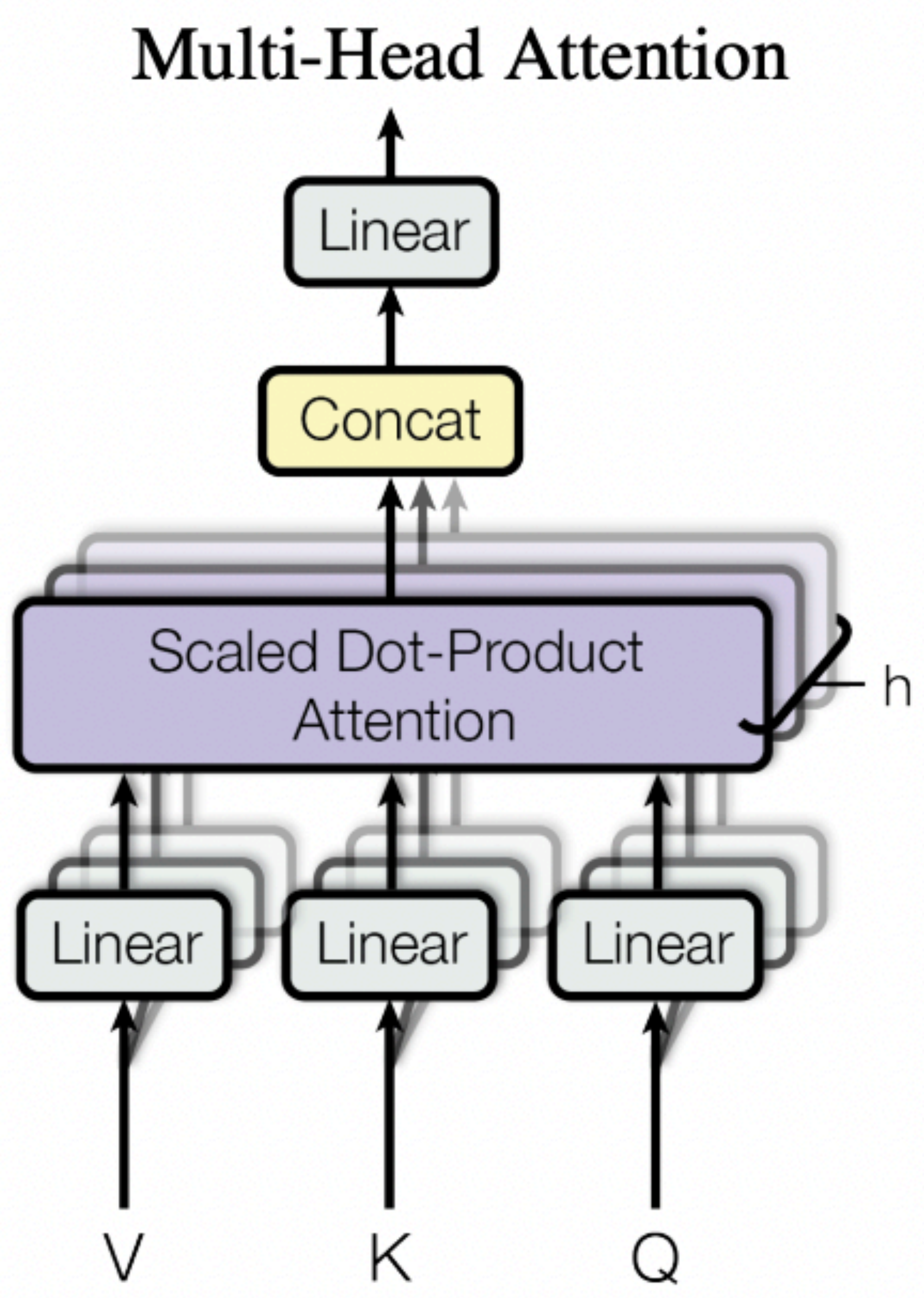
RETAIL

Future Directions

[Vaswani et al., 2017]

Background: Transformer Model



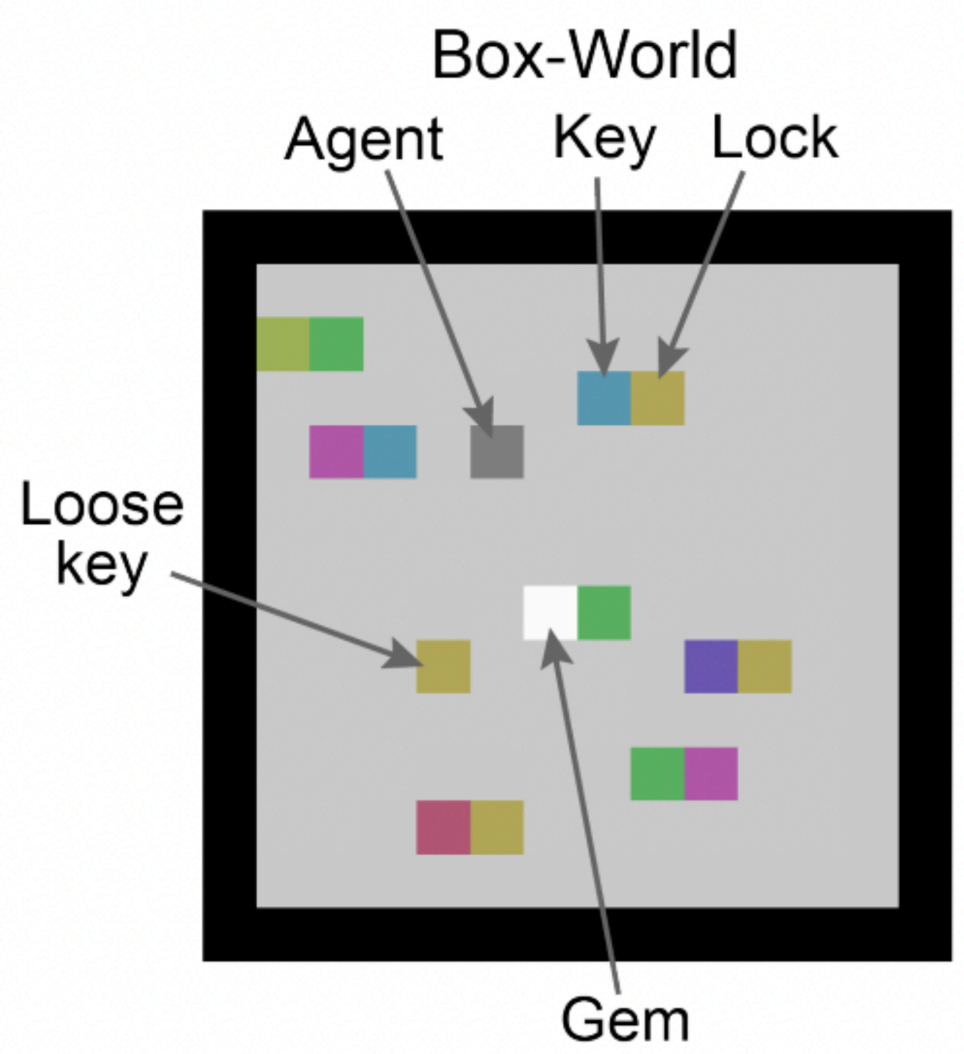


Background: Relational Reasoning

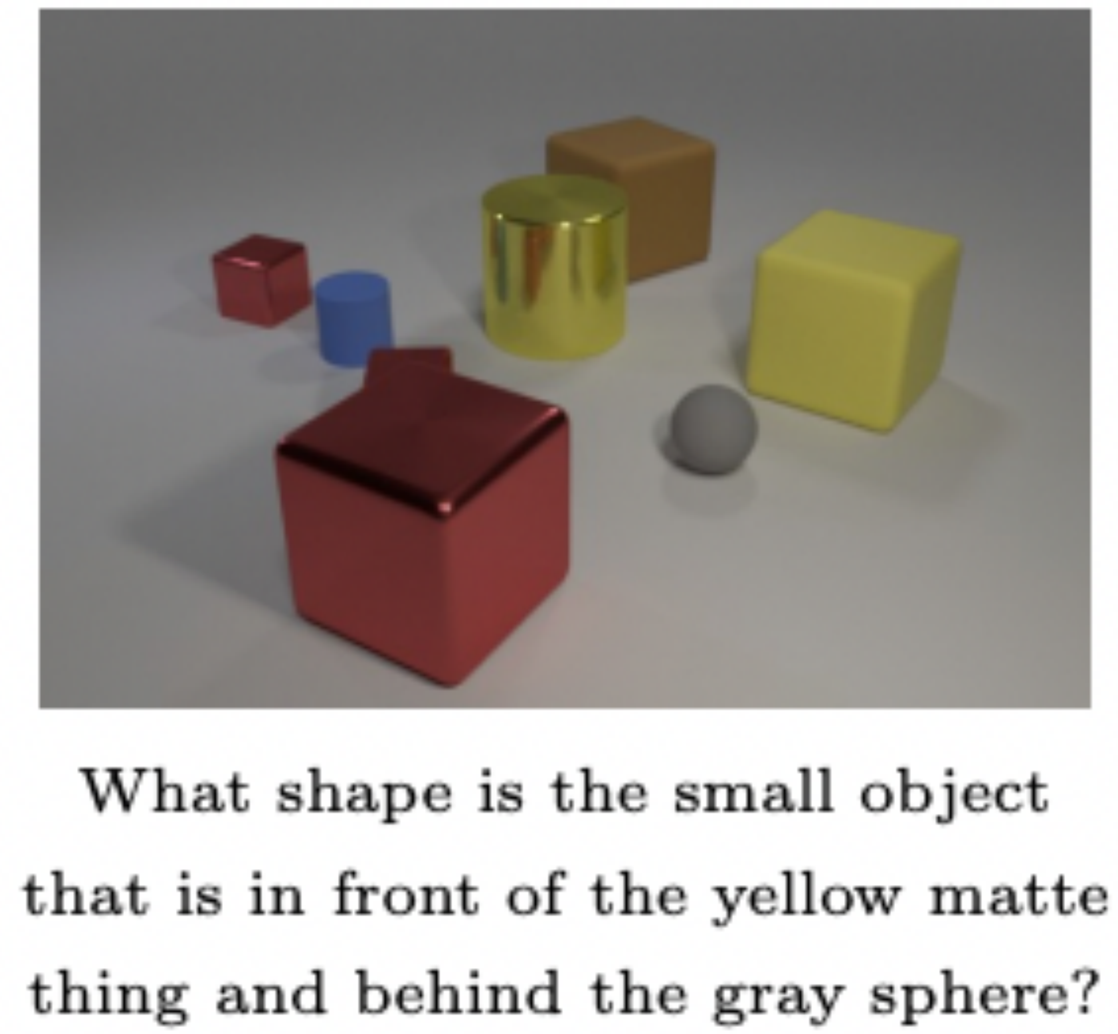
- Input: Set of entities
- Model reasons about the relationships between these entities

Background: Relational Reasoning

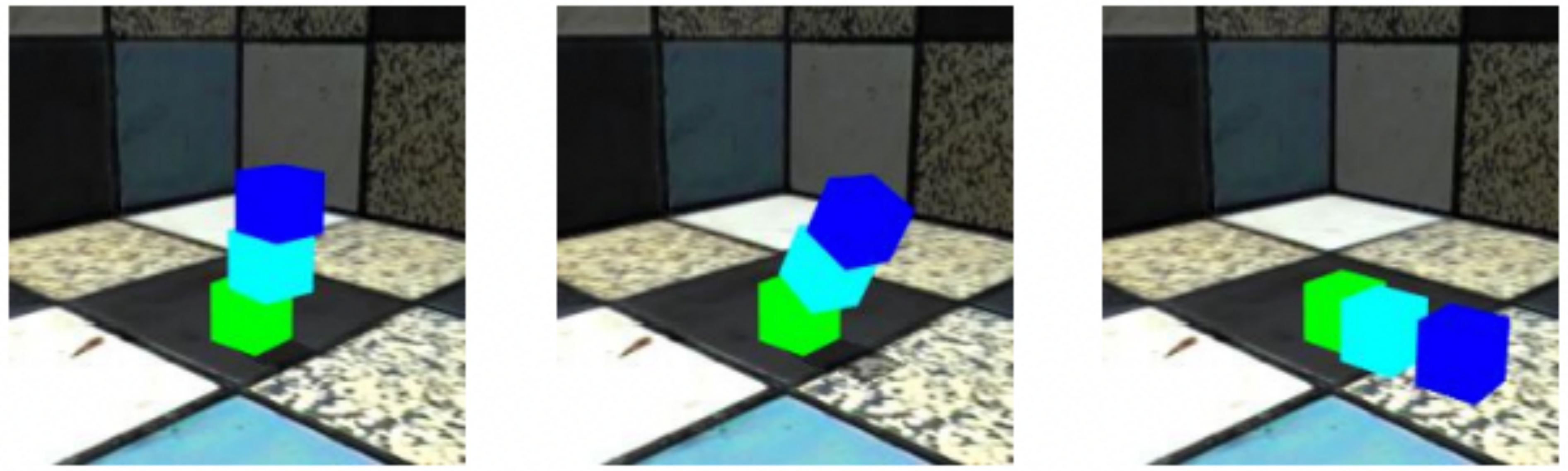
- Input: Set of entities
- Model reasons about the relationships between these entities
- Shown to be effective in
  - RL
  - Visual question answering
  - Learning passive dynamics



[Zambaldi et al., 2018]



[Santoro et al., 2017]



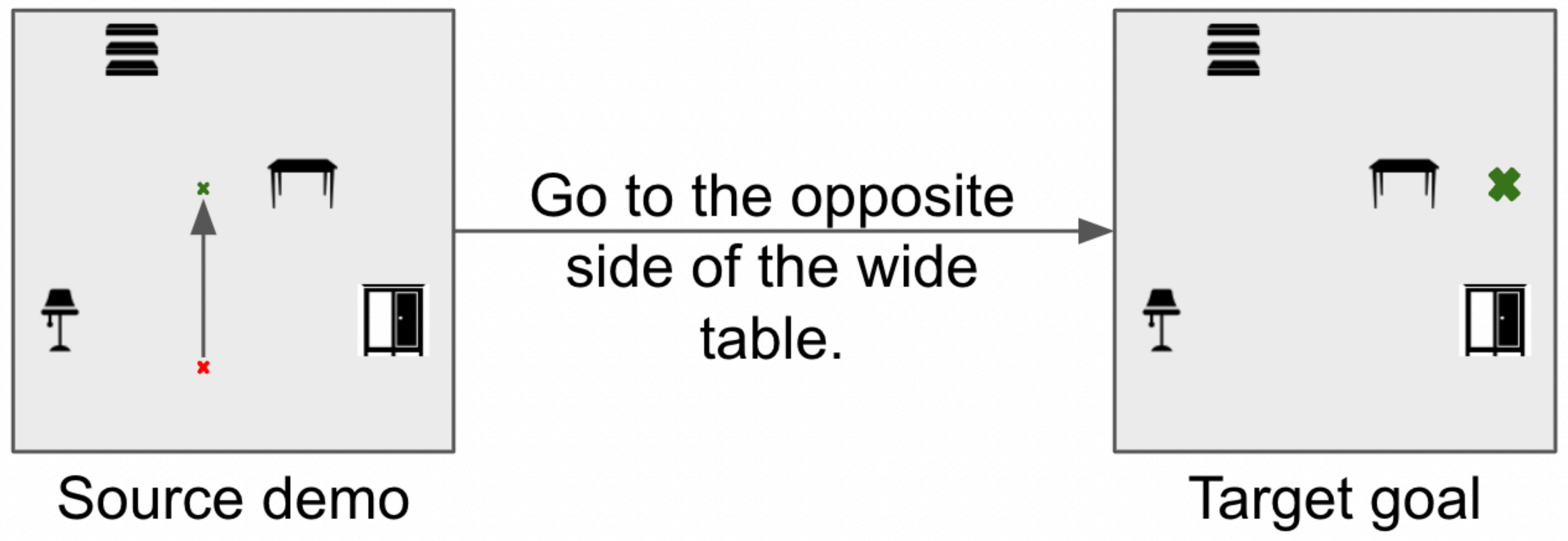
[Didolkar et al., 2021]

REIational Task Adaptation for Imitation with Language (RETAIL)

Motivation

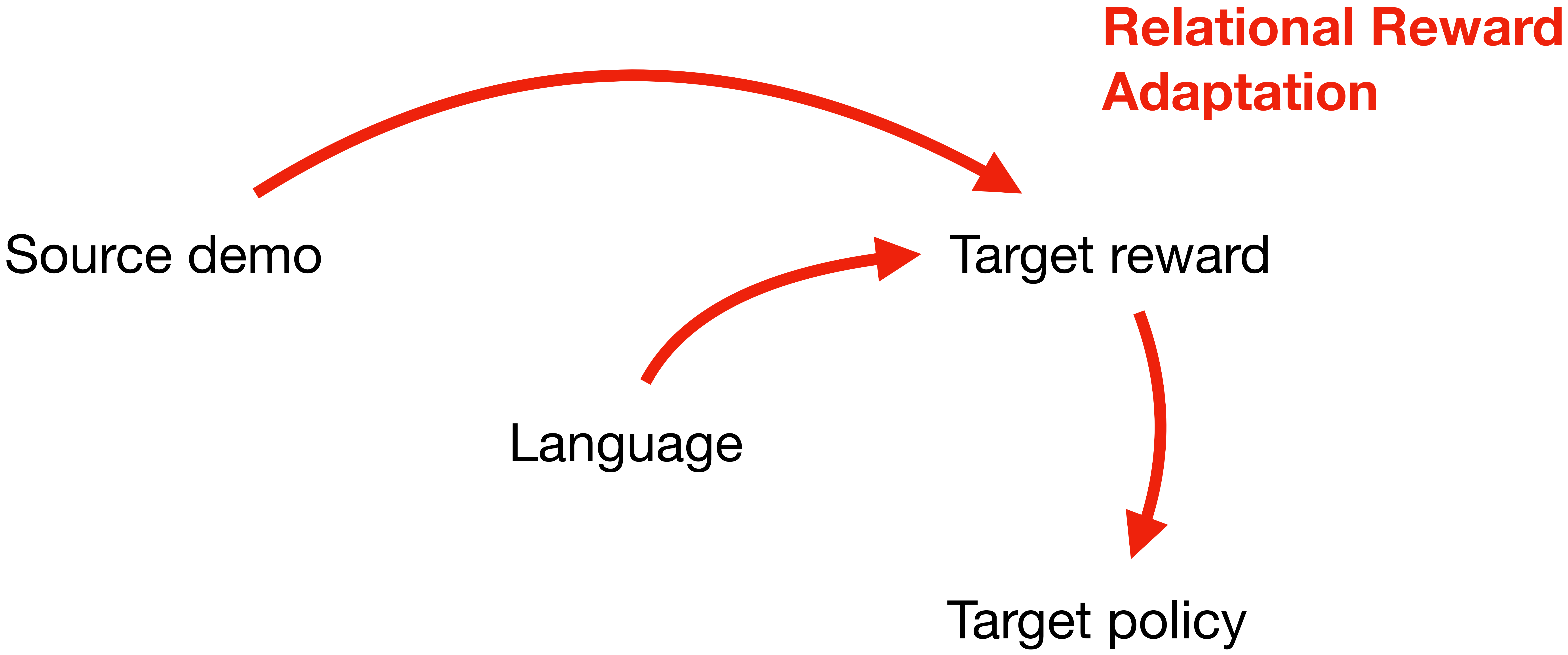
Limitations of LARVA:

- Does not reason about the structure of the tasks or the environment.
- Does not have an explicit policy learning phase for evaluation.
- Assumes data of the form
  (src demo, lang, tgt goal, tgt reward/value function).



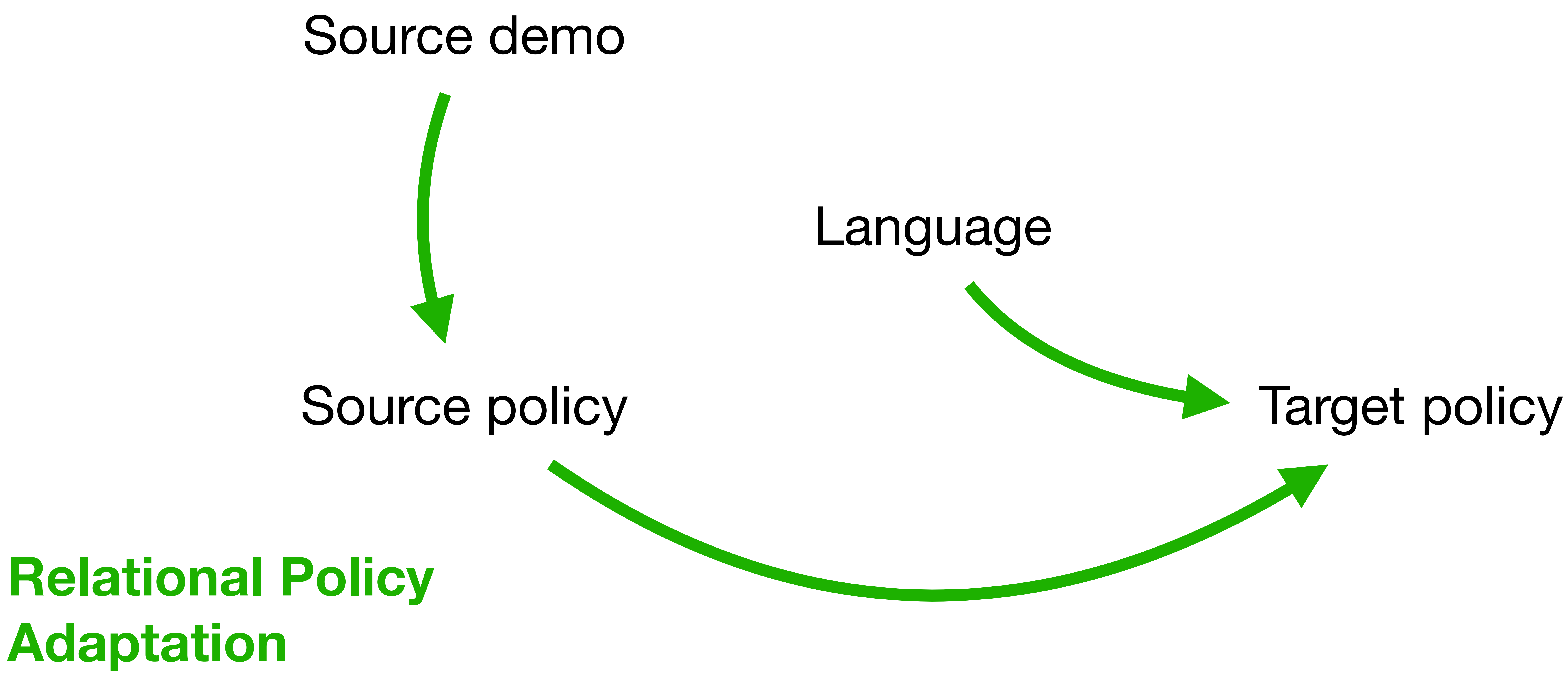
RElational Task Adaptation for Imitation with Language (RETAIL)

Approach Overview



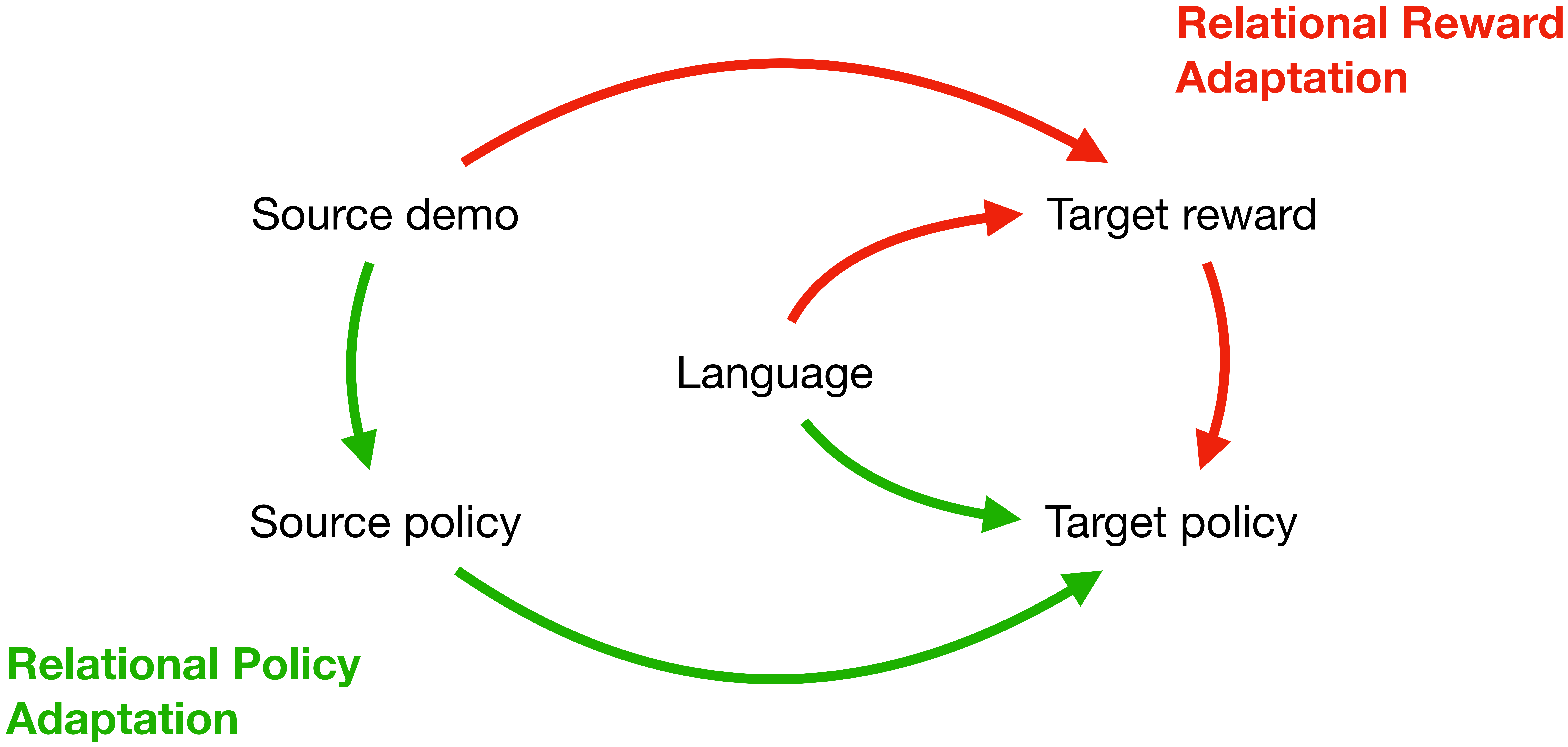
RElational Task Adaptation for Imitation with Language (RETAIL)

Approach Overview



RElational Task Adaptation for Imitation with Language (RETAIL)

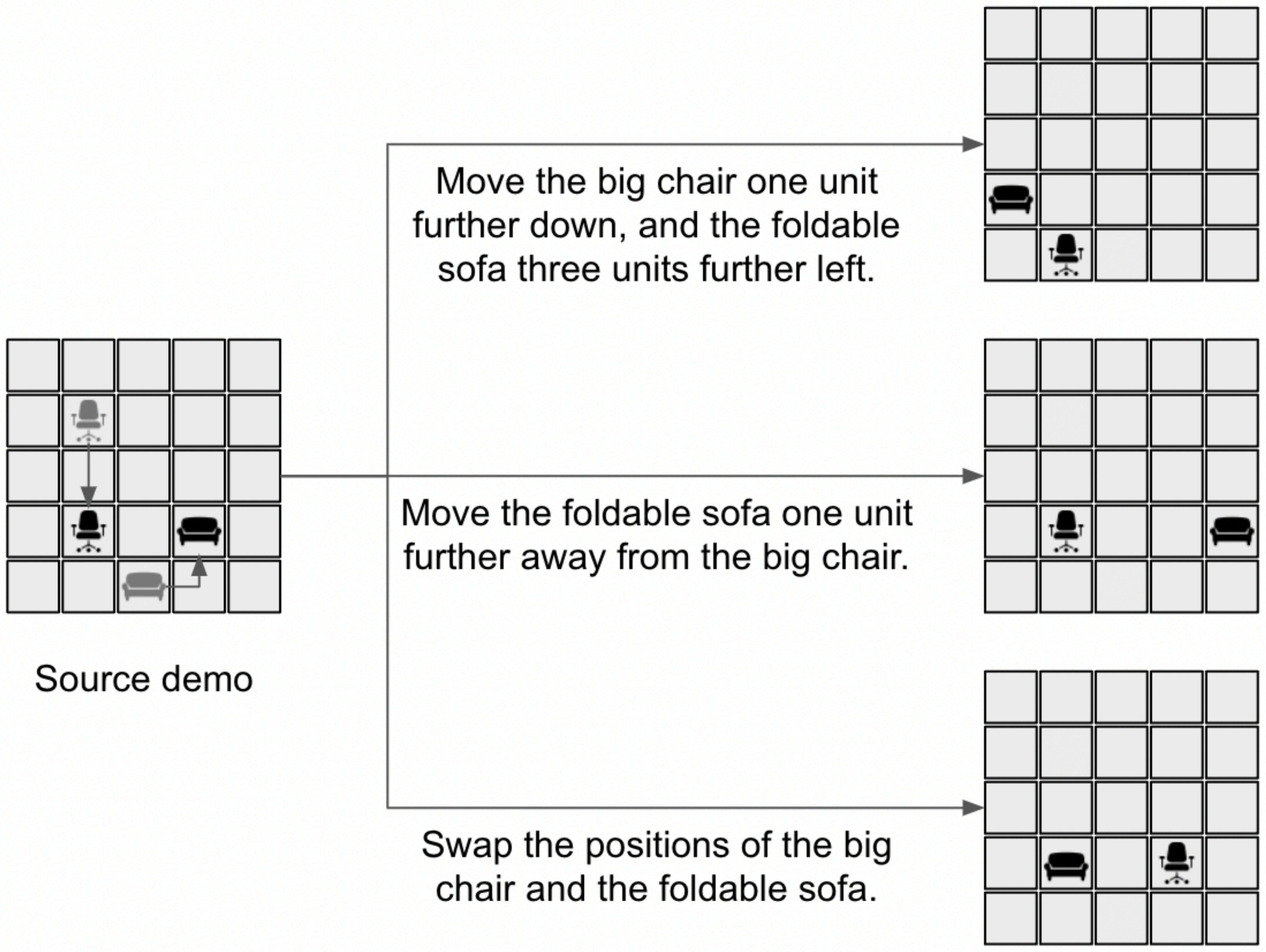
Approach Overview



REIational Task Adaptation for Imitation with Language (RETAIL)

Domains: Room Rearrangement

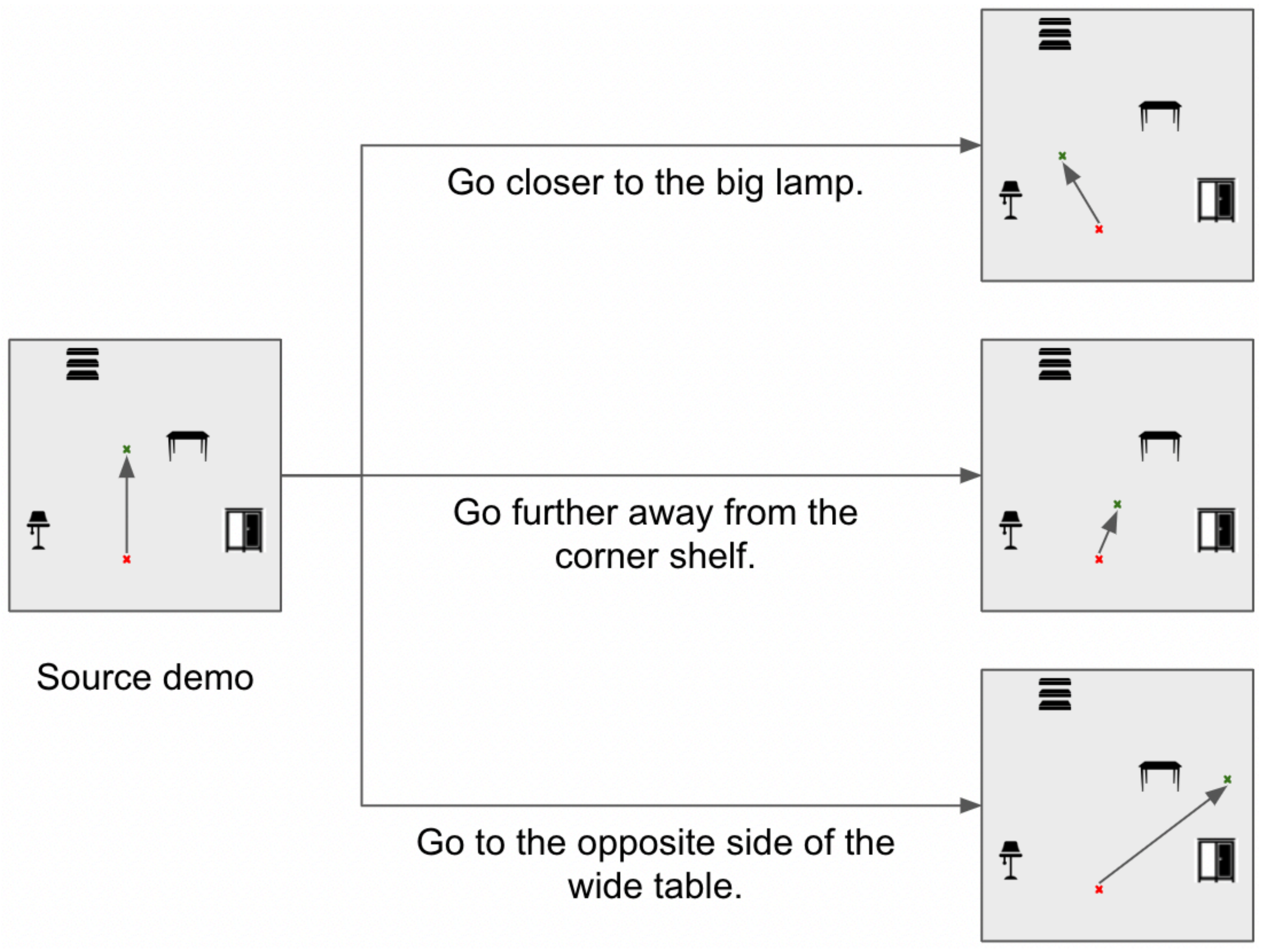
- Agent + 2 objects
- Actions:
  - Up, Down, Left, Right
  - Grasp, Release
  - Stop
- Tasks: Moving each object to desired goal locations.



REIational Task Adaptation for Imitation with Language (RETAIL)

Domains: Room Navigation

- Agent + 4 objects
- Actions:
  - (Δx, Δy)
- Tasks: Navigating to a desired goal location.



REIational Task Adaptation for Imitation with Language (RETAIL)

Domains

Room Rearrangement

- Discrete states and actions
- Short horizon tasks (~30 steps)
- Multiple optimal trajectories

Room Navigation

- Continuous states and actions
- Long horizon tasks (~150 steps)
- Unique optimal trajectory

REIational Task Adaptation for Imitation with Language (RETAIL)

Domains: Objects

	Table	Chair	Sofa	Light	Shelf	Wardrobe
Large	36 total (attribute, object) combinations 24 for training 6 for validation 6 for testing					
Wide						
Wooden						
Metallic						
Corner						
Foldable						

REIational Task Adaptation for Imitation with Language (RETAIL)

Domains: Data Collection

- Planner to generate source and target demos

- For each type of adaptation:
  - Training set: 5000
  - Validation set for supervised learning: 100
  - Validation set for RL: 5
  - Test set: 10

- Language:
  - Synthetic: Using templates
  - Natural: Paraphrases using Amazon Mechanical Turk

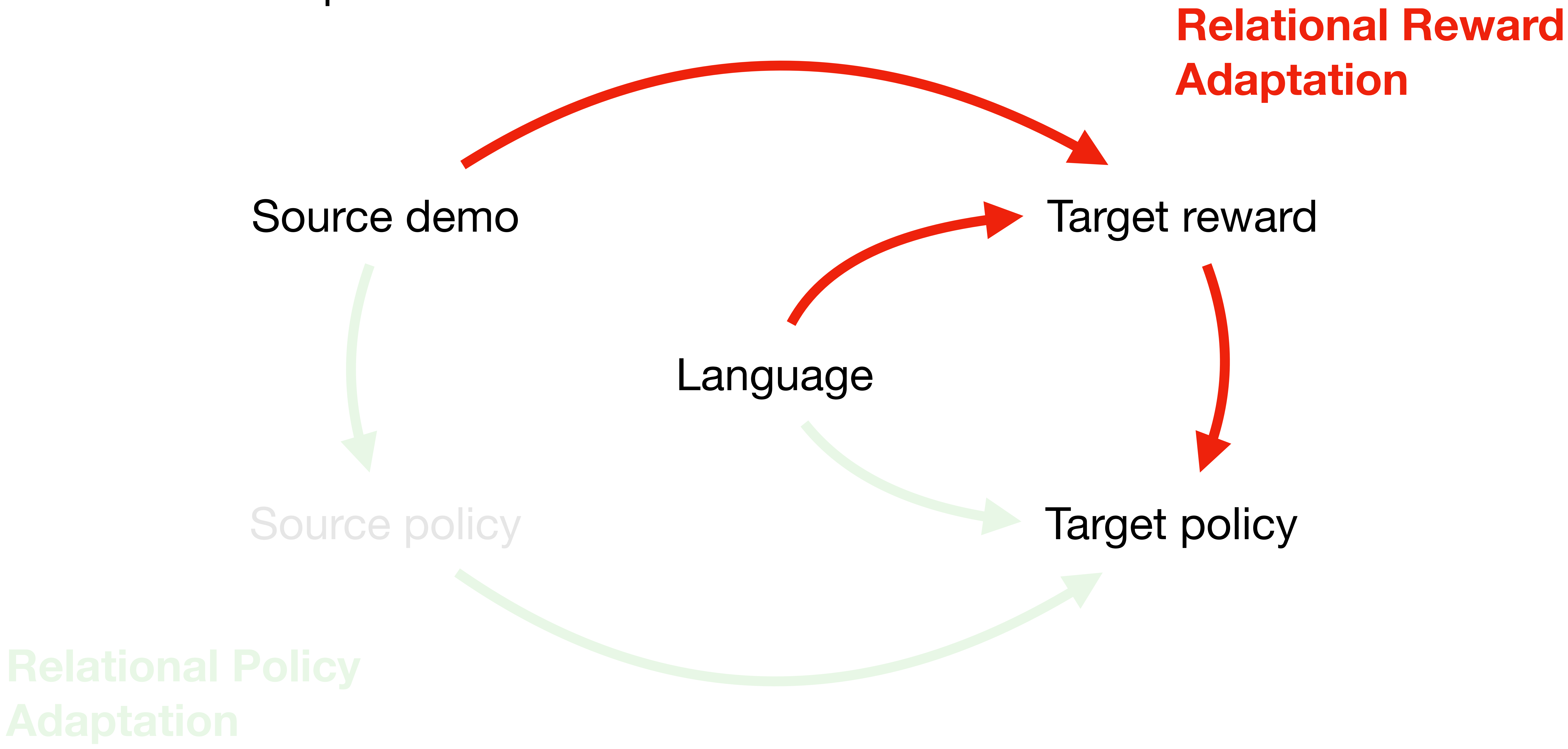
REIational Task Adaptation for Imitation with Language (RETAIL)

Domains: Data Collection

	Template	Natural language paraphrase
1.	go further away from the metallic table	Increase your distance from the metallic table.
2.	go closer to the foldable light	Move in the direction of the light that is foldable
3.	go to the opposite side of the corner light	Move across from the corner light.
4.	move the large chair one unit farther from the wide couch	Increment the distance of the big chair from the wide couch by one.
5.	move corner table two units further left and metallic shelf one unit further backward	slide the corner table two units left and move the metal shelf a single unit back
6.	move the large table to where the large sofa was moved, and vice versa	swap the place of the table with the sofa

RElational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation



REIational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation: Approach

Reward:

$$R(s, s') = \phi(s') - \phi(s)$$

g_{src} : Goal state for the source task

g_{tgt} : Goal state for the target task

l : Language describing the difference

REIational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation: Approach

Reward:

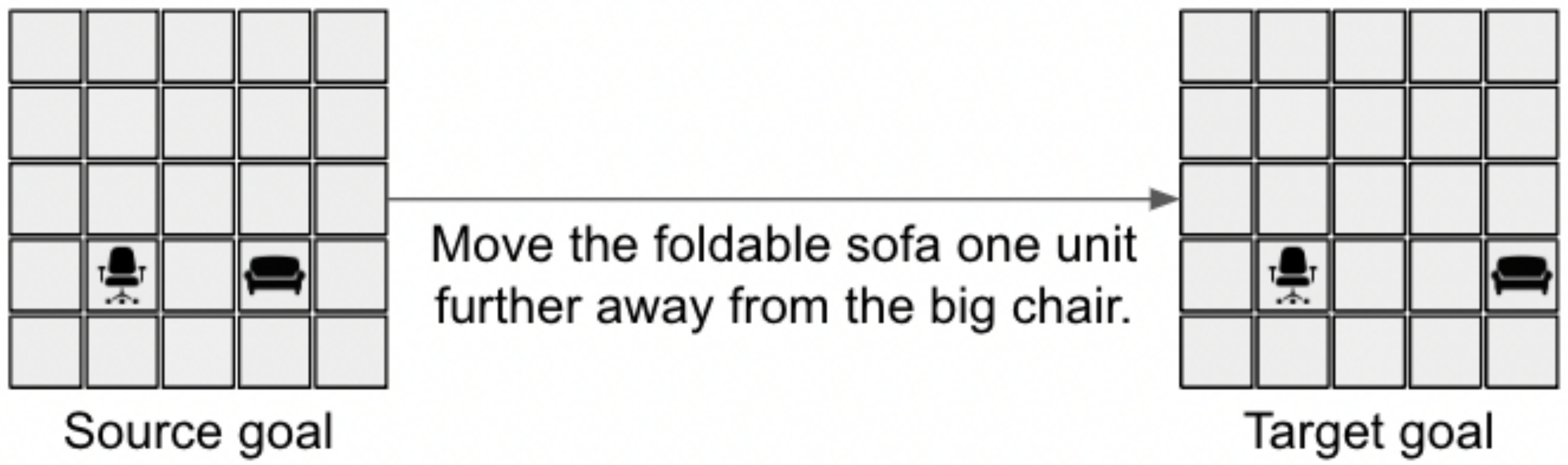
$$R(s, s') = \phi(s') - \phi(s)$$

g_{src} : Goal state for the source task

g_{tgt} : Goal state for the target task

l : Language describing the difference

Goal prediction: $g_{tgt} = \text{Adapt}(g_{src}, l)$



Distance function: $d(s, s')$

REIational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation: Approach

Reward:

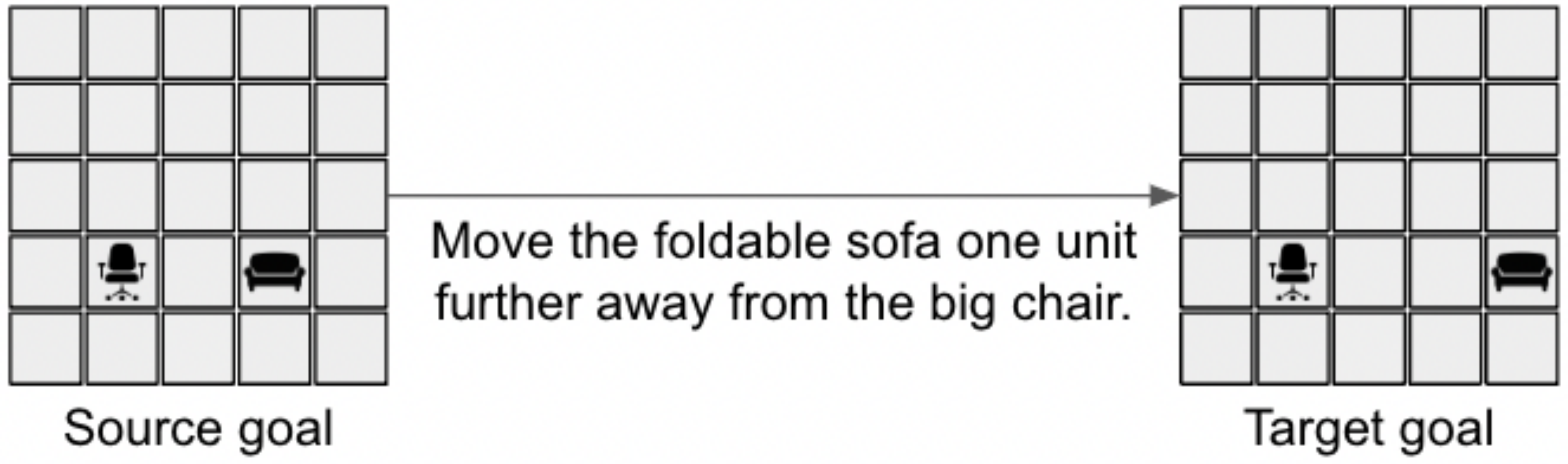
$$R(s, s') = \phi(s') - \phi(s)$$

g_{src} : Goal state for the source task

g_{tgt} : Goal state for the target task

l : Language describing the difference

Goal prediction: $g_{tgt} = \text{Adapt}(g_{src}, l)$

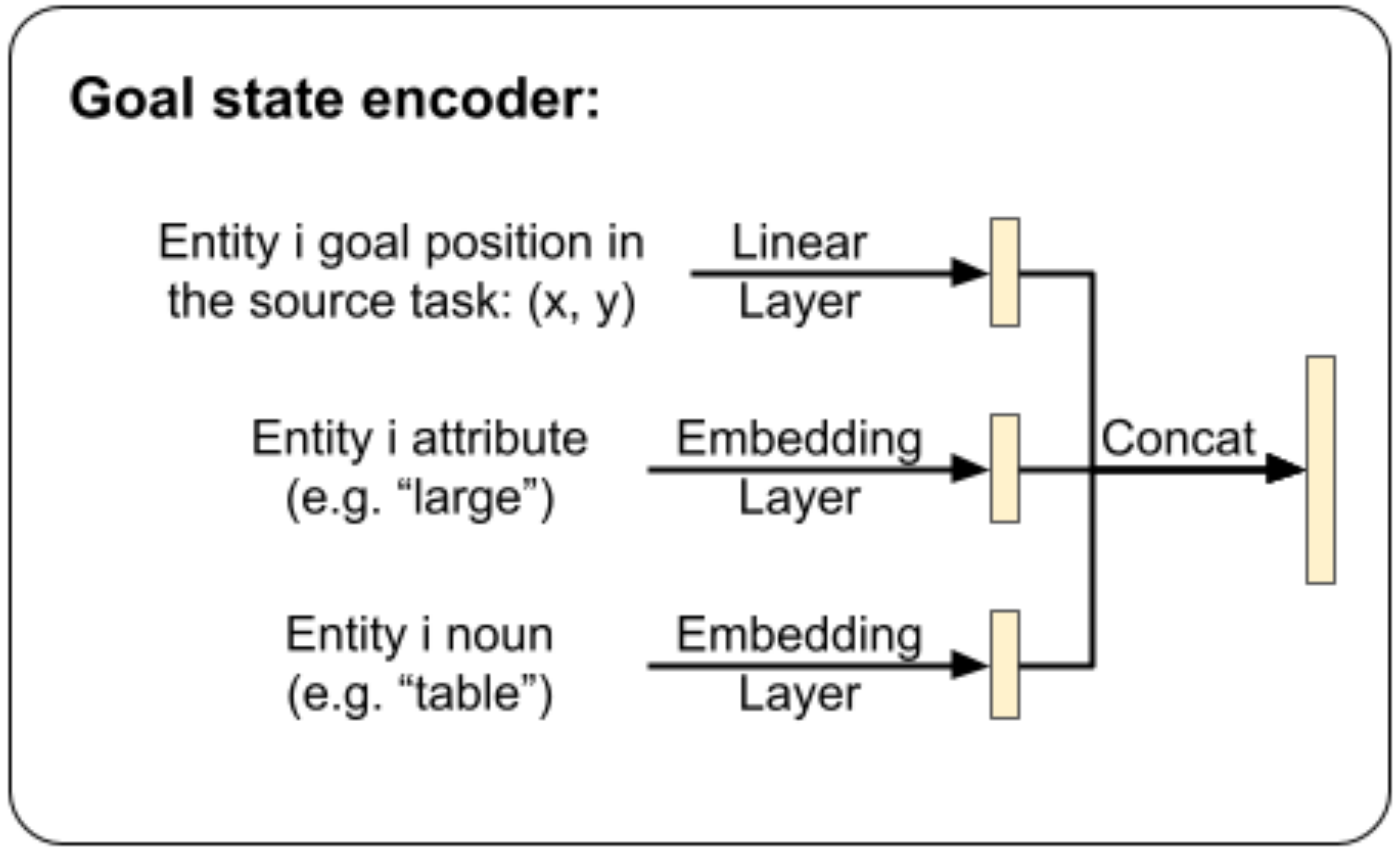


Distance function: $d(s, s')$

$$\implies \phi_{tgt}(s \mid g_{src}, l) = -d(s, \text{Adapt}(g_{src}, l))$$

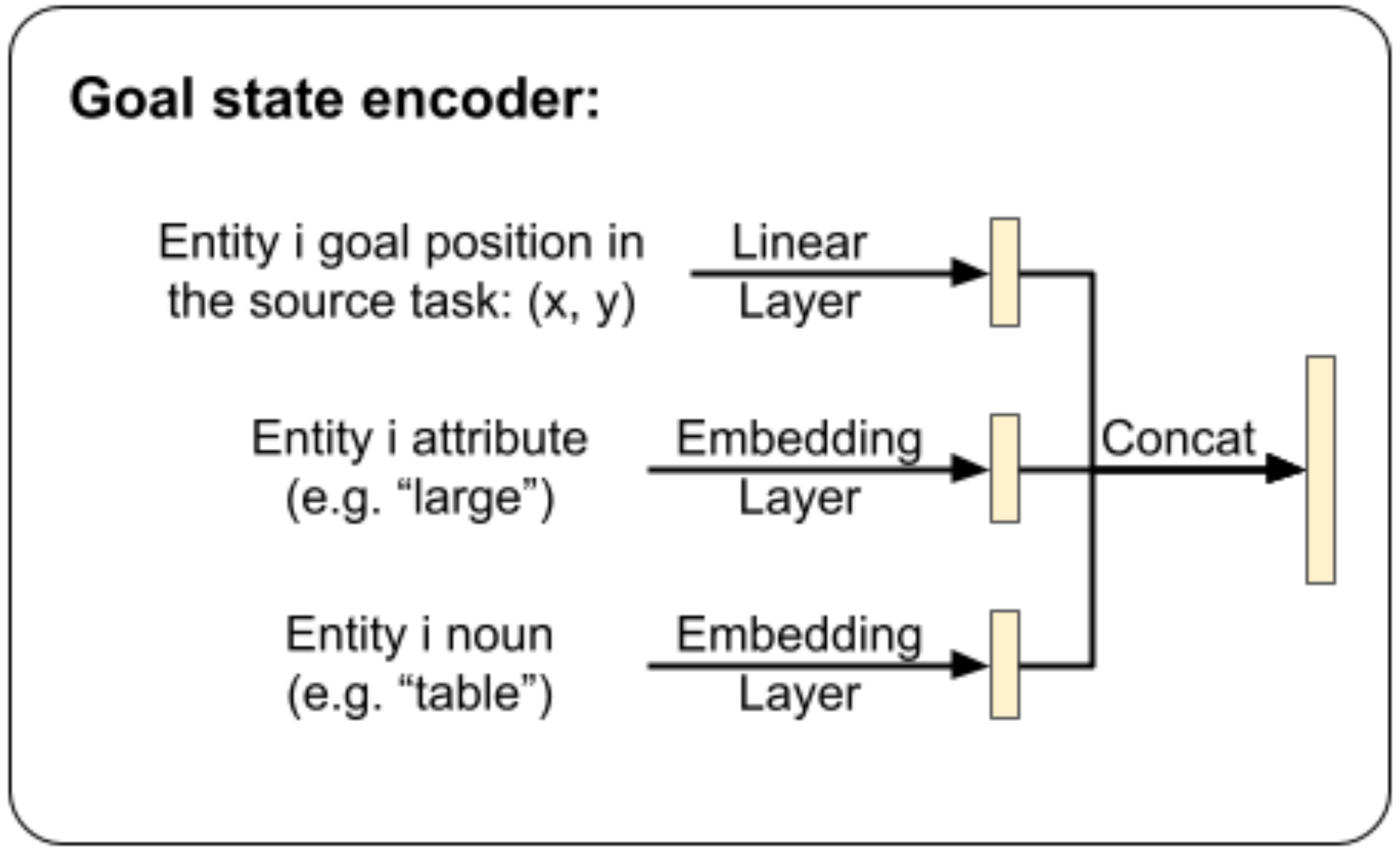
REIational Task Adaptation for Imitation with Language (RETAIL)

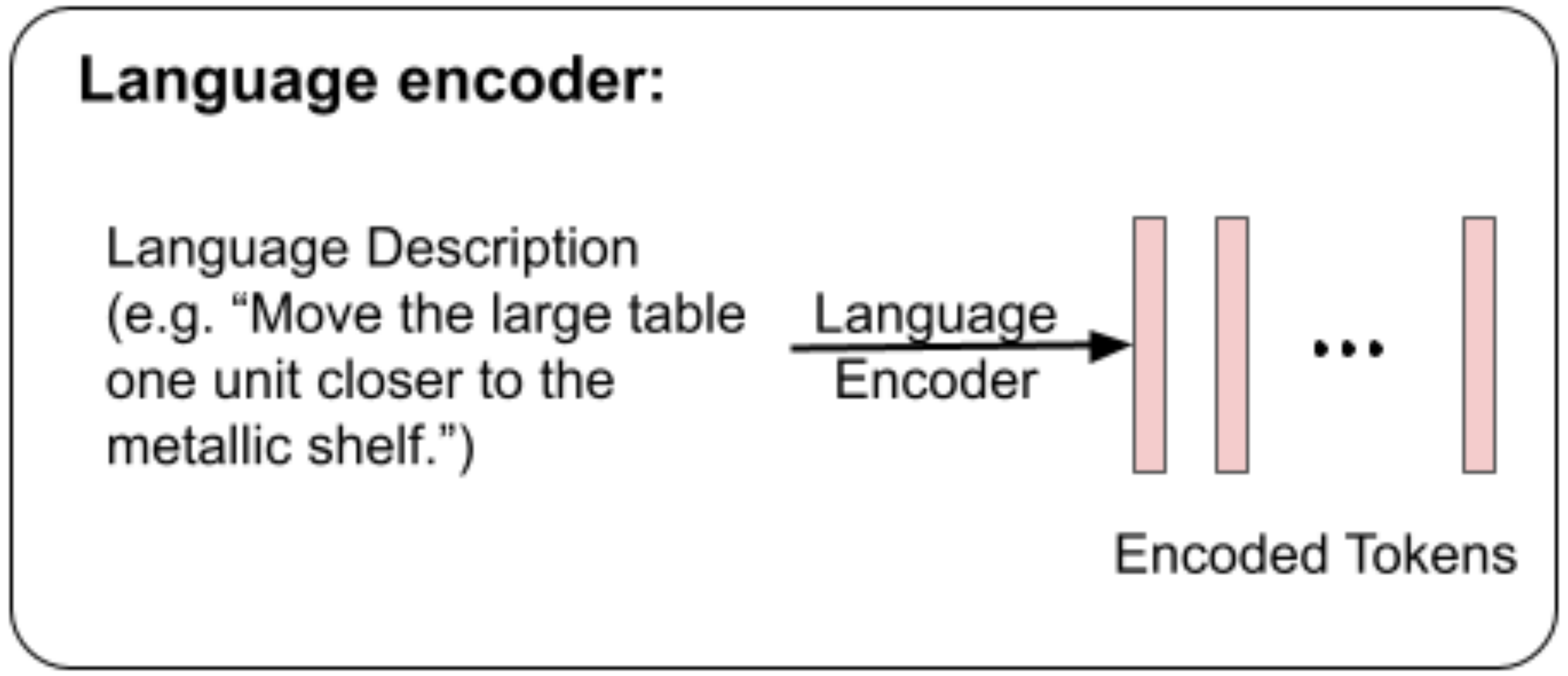
Relational Reward Adaptation: Goal Prediction



RElational Task Adaptation for Imitation with Language (RETAIL)

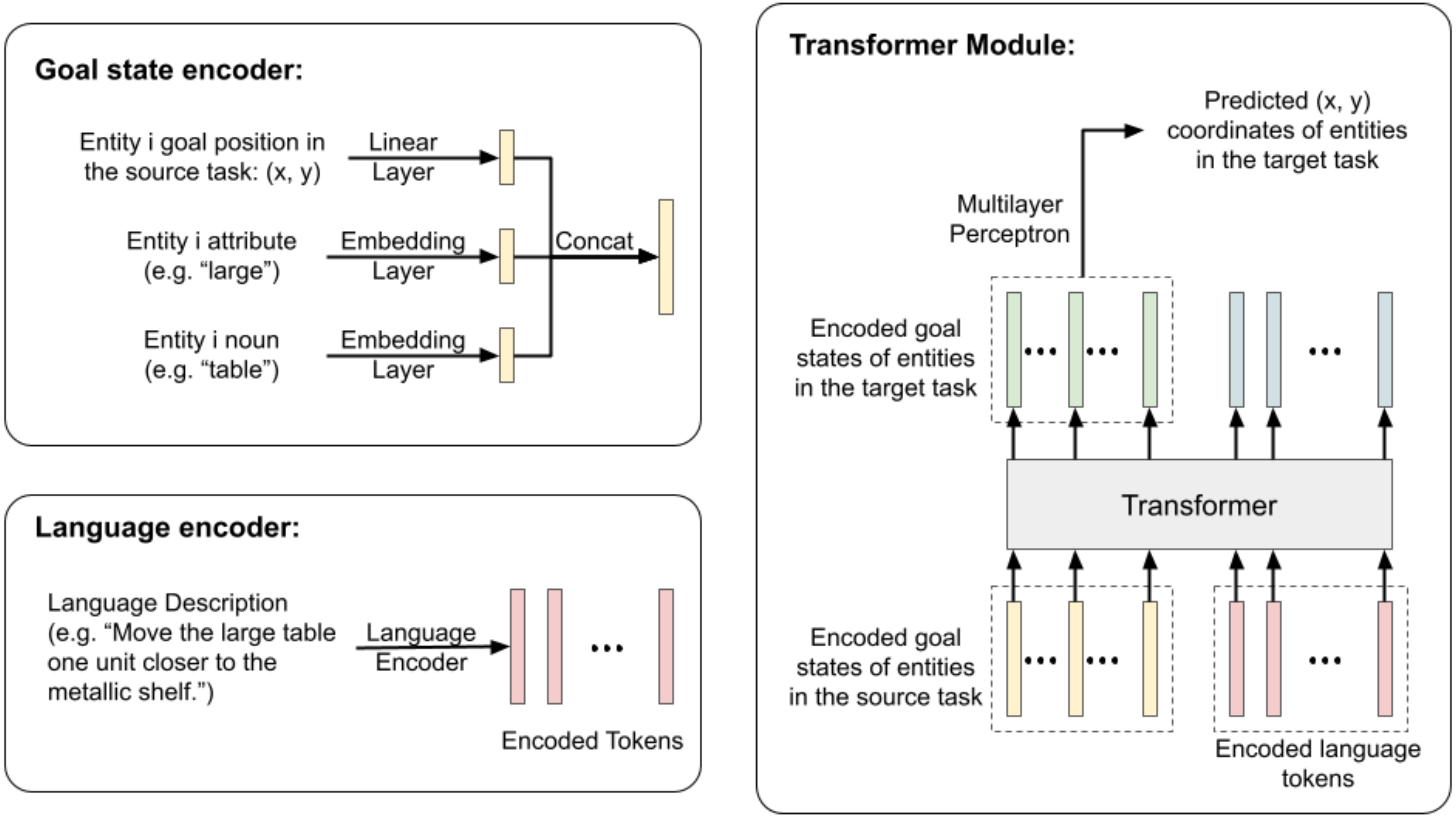
Relational Reward Adaptation: Goal Prediction





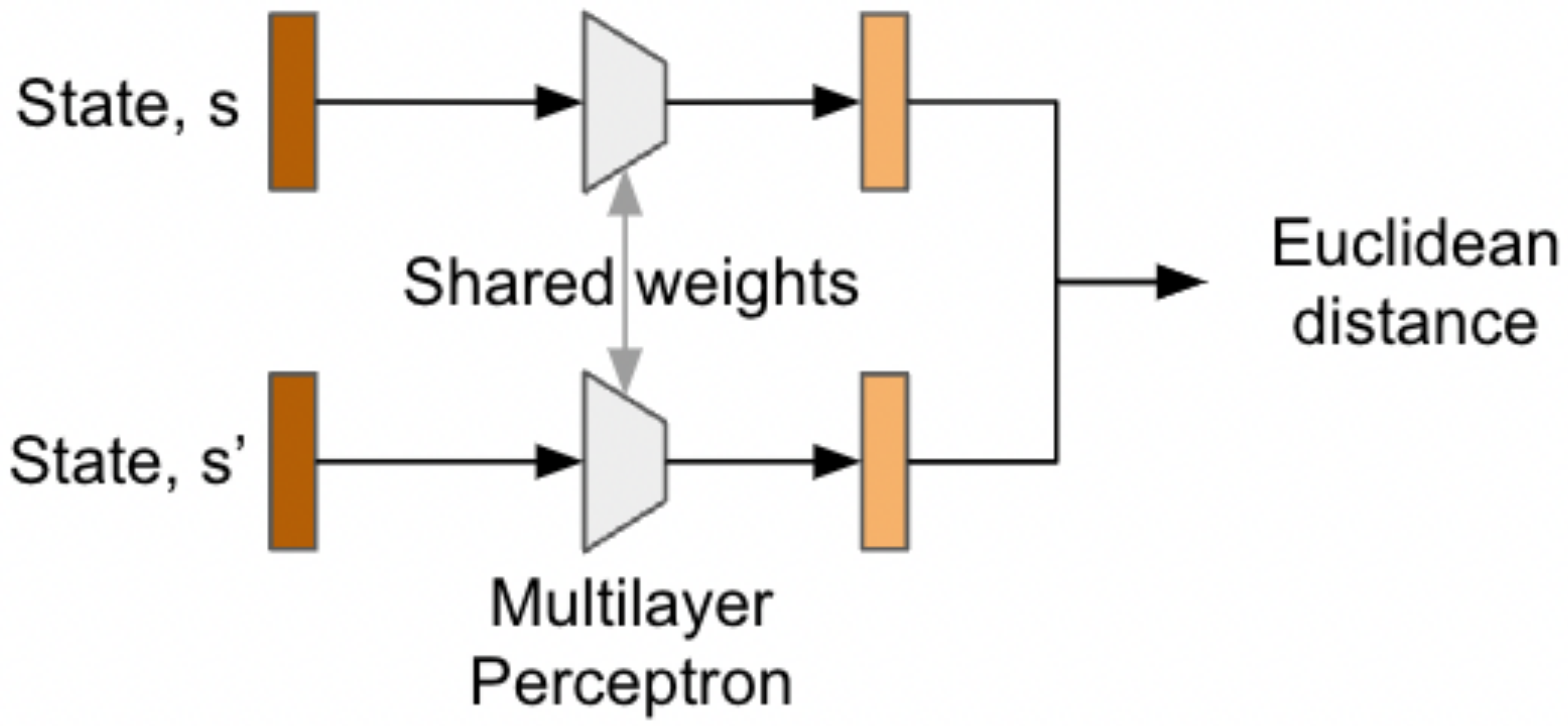
RElational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation: Goal Prediction



REIational Task Adaptation for Imitation with Language (RETAIL)

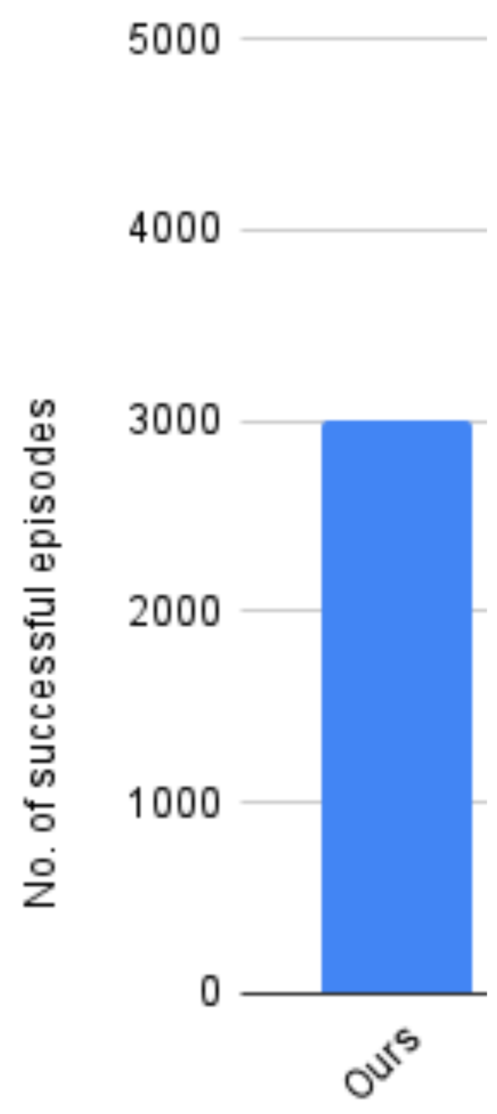
Relational Reward Adaptation: Distance Function



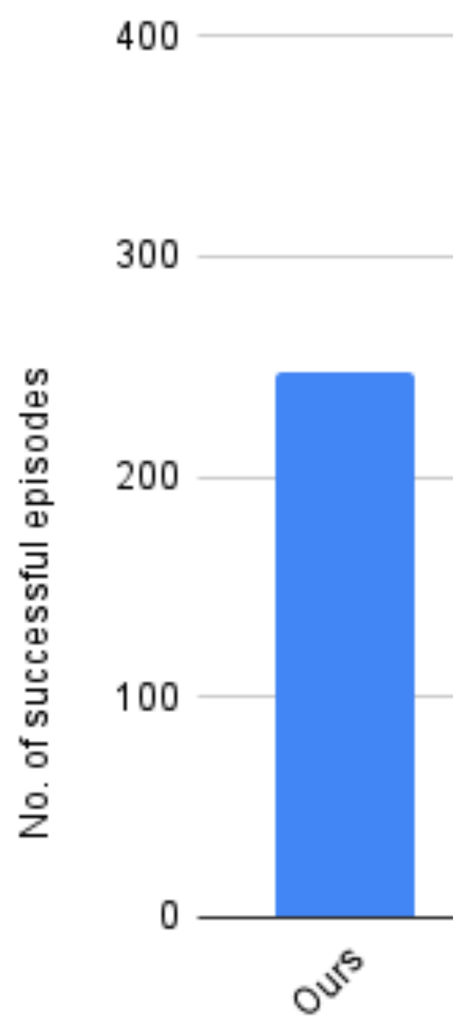
REIational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation: Experiments

Rearrangement



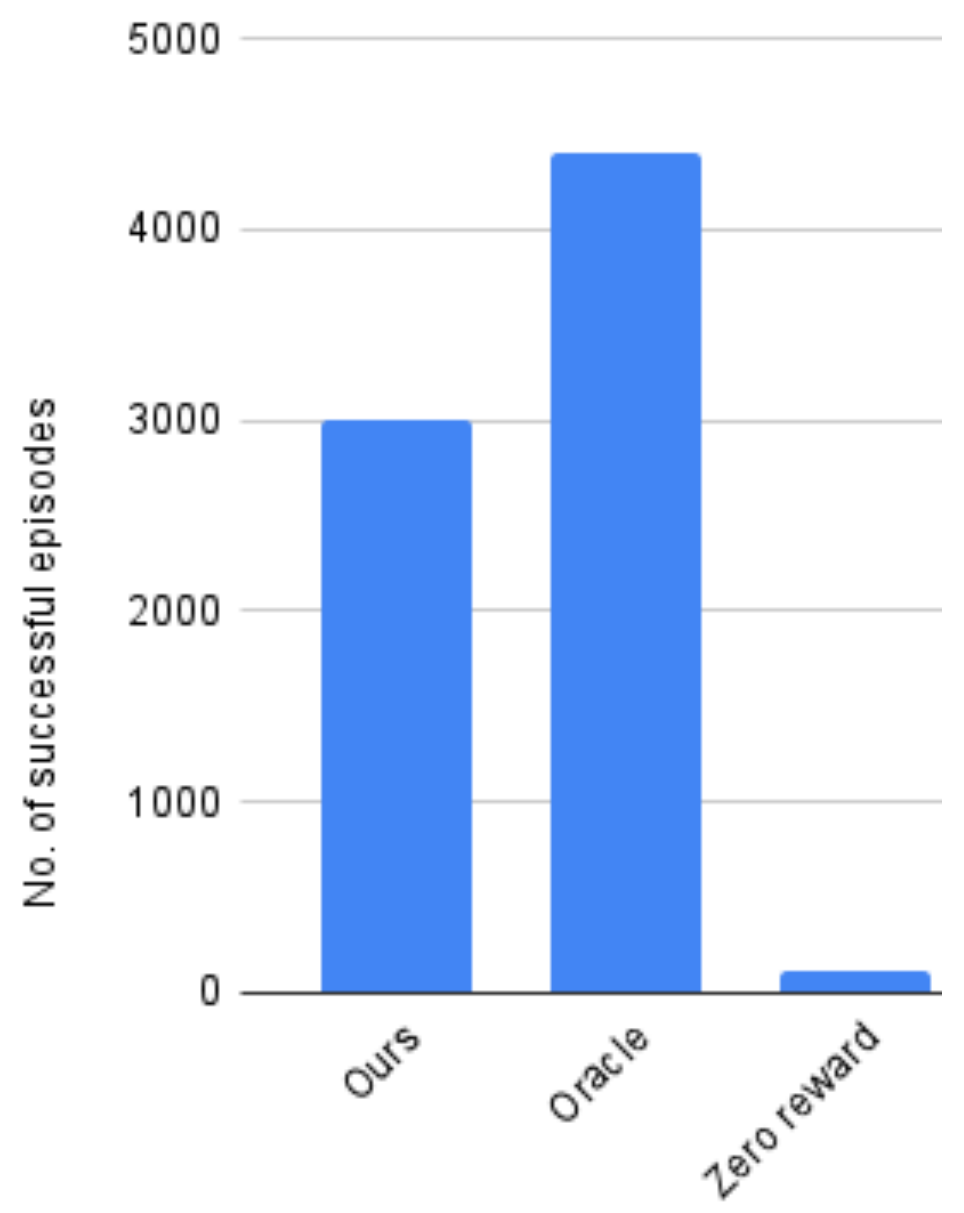
Navigation



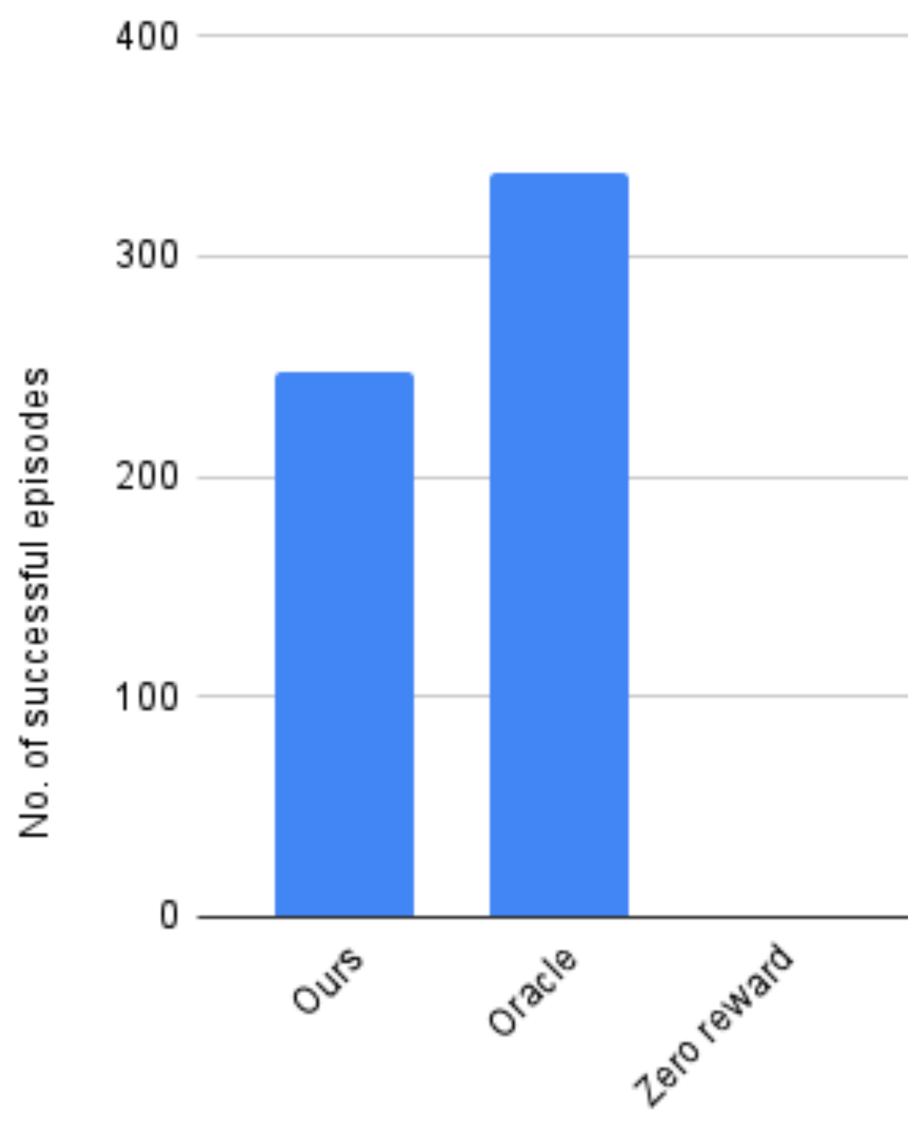
REIational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation: Experiments

Rearrangement



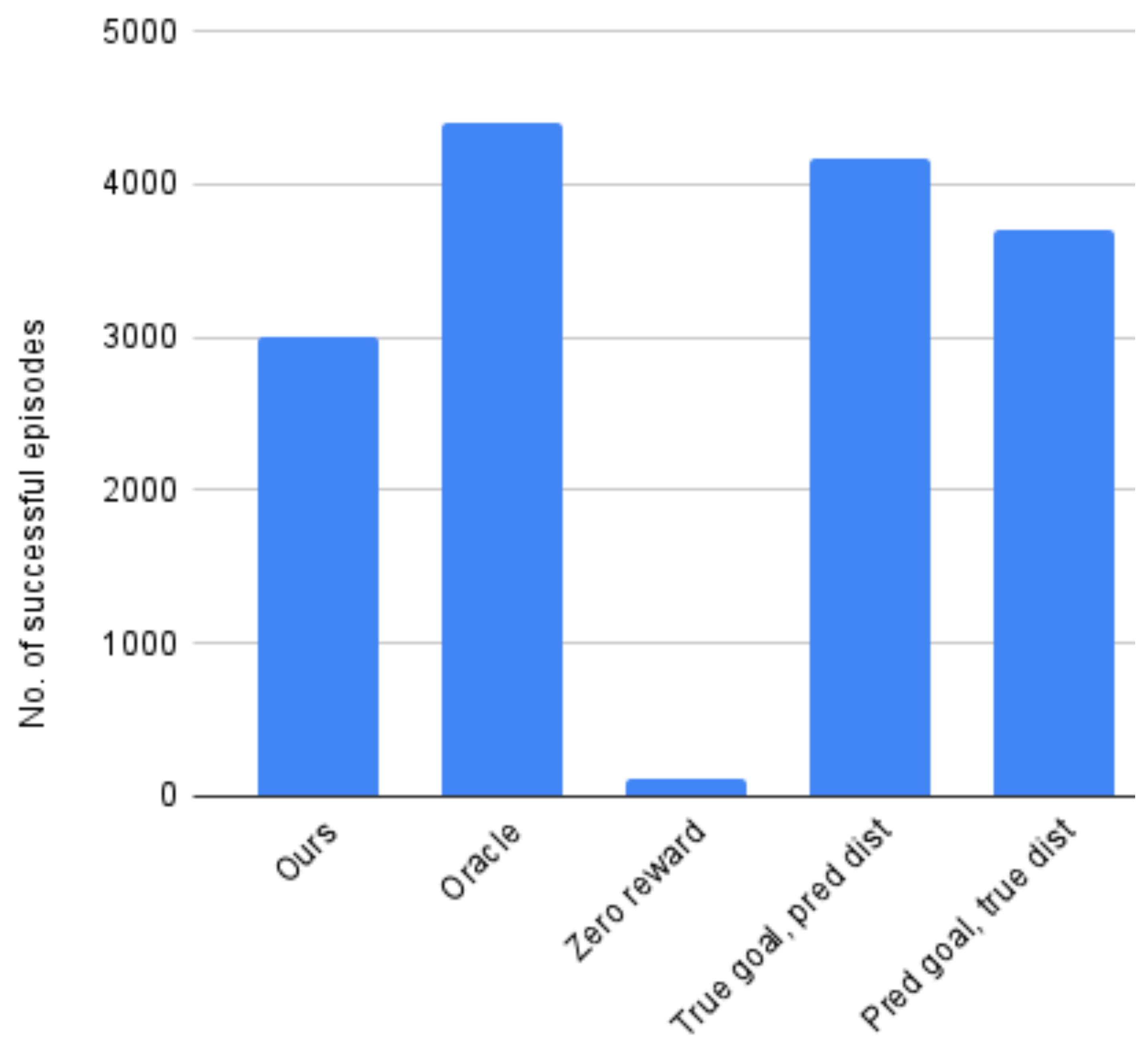
Navigation



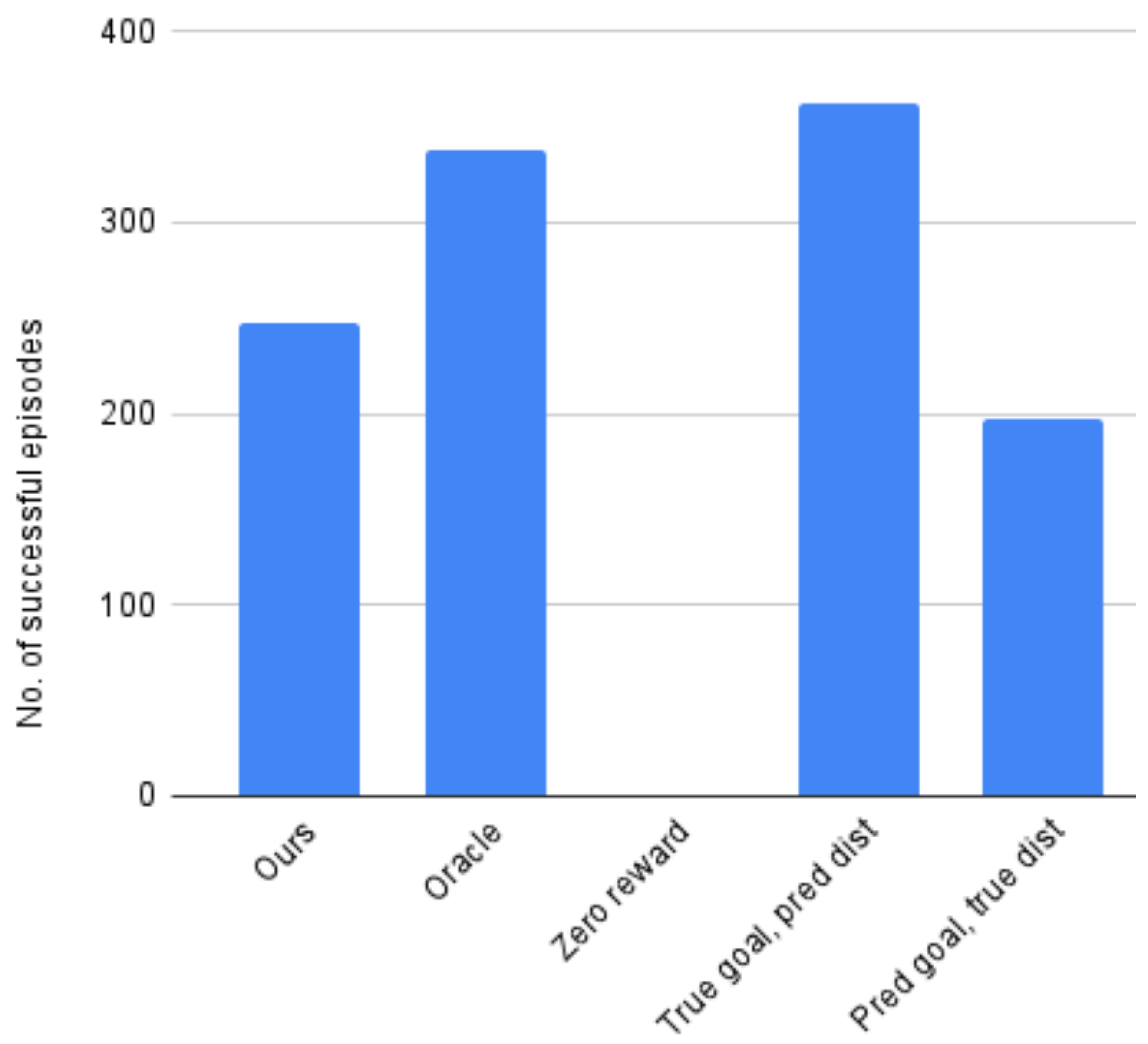
REIational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation: Experiments

Rearrangement



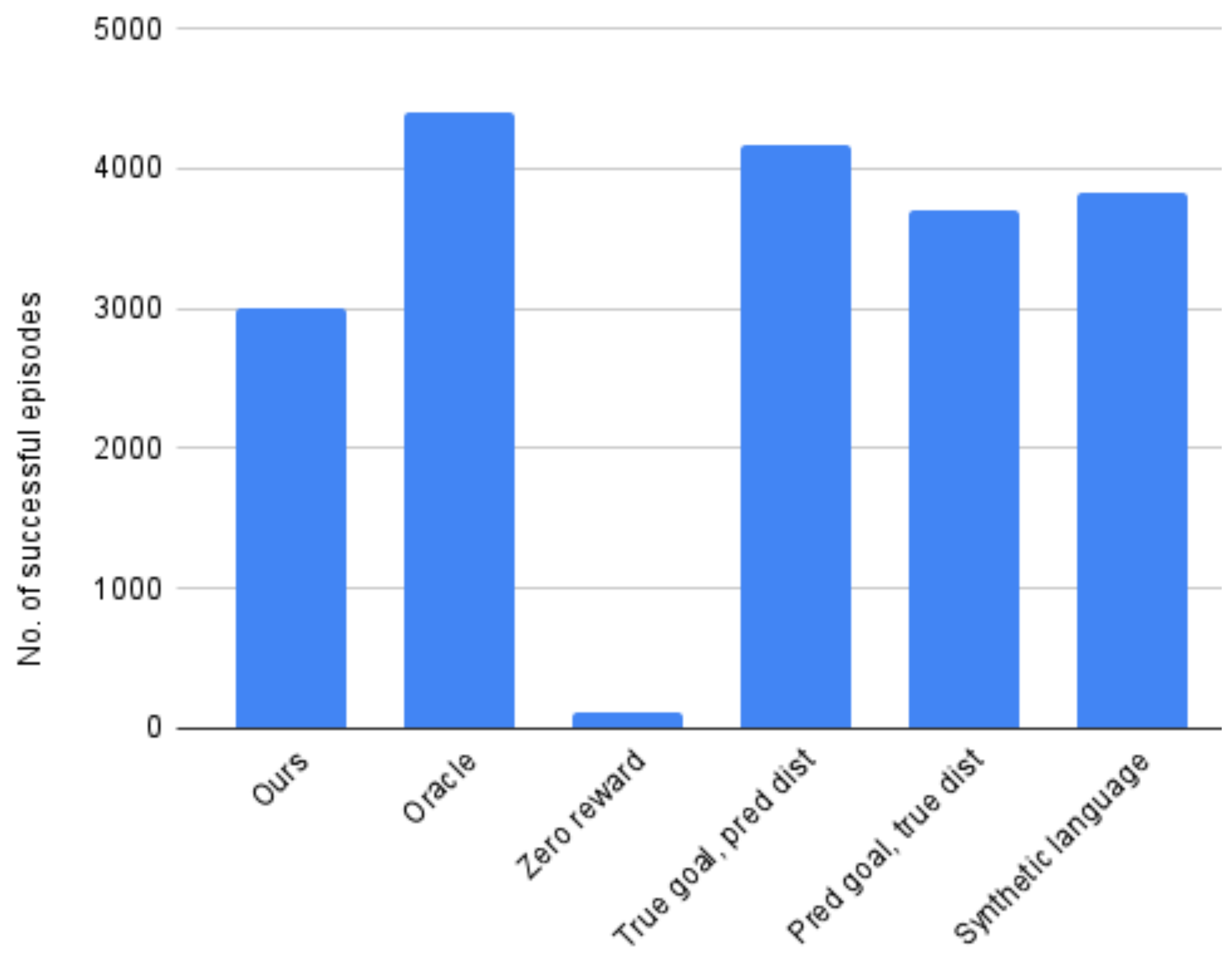
Navigation



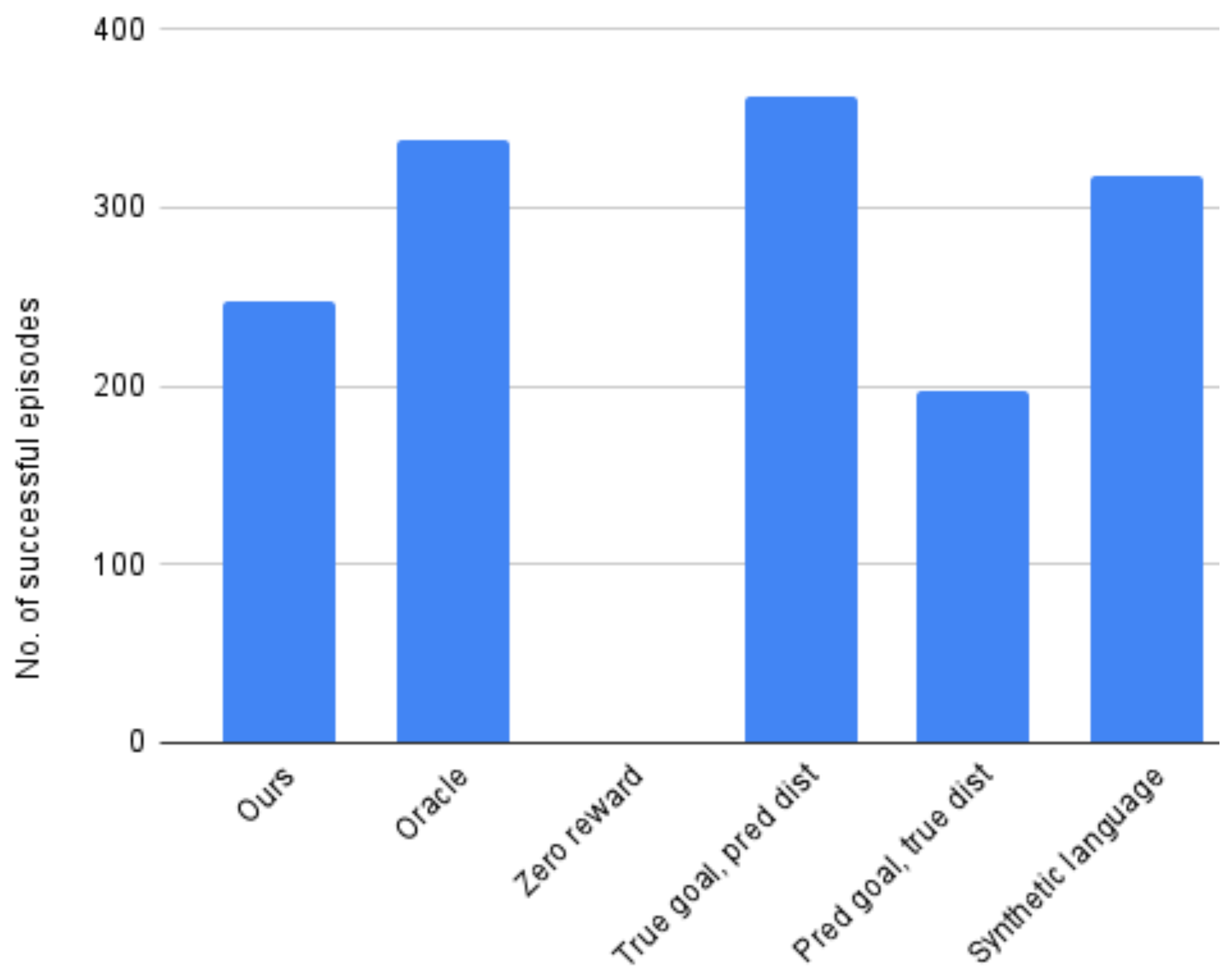
RElational Task Adaptation for Imitation with Language (RETAIL)

Relational Reward Adaptation: Experiments

Rearrangement

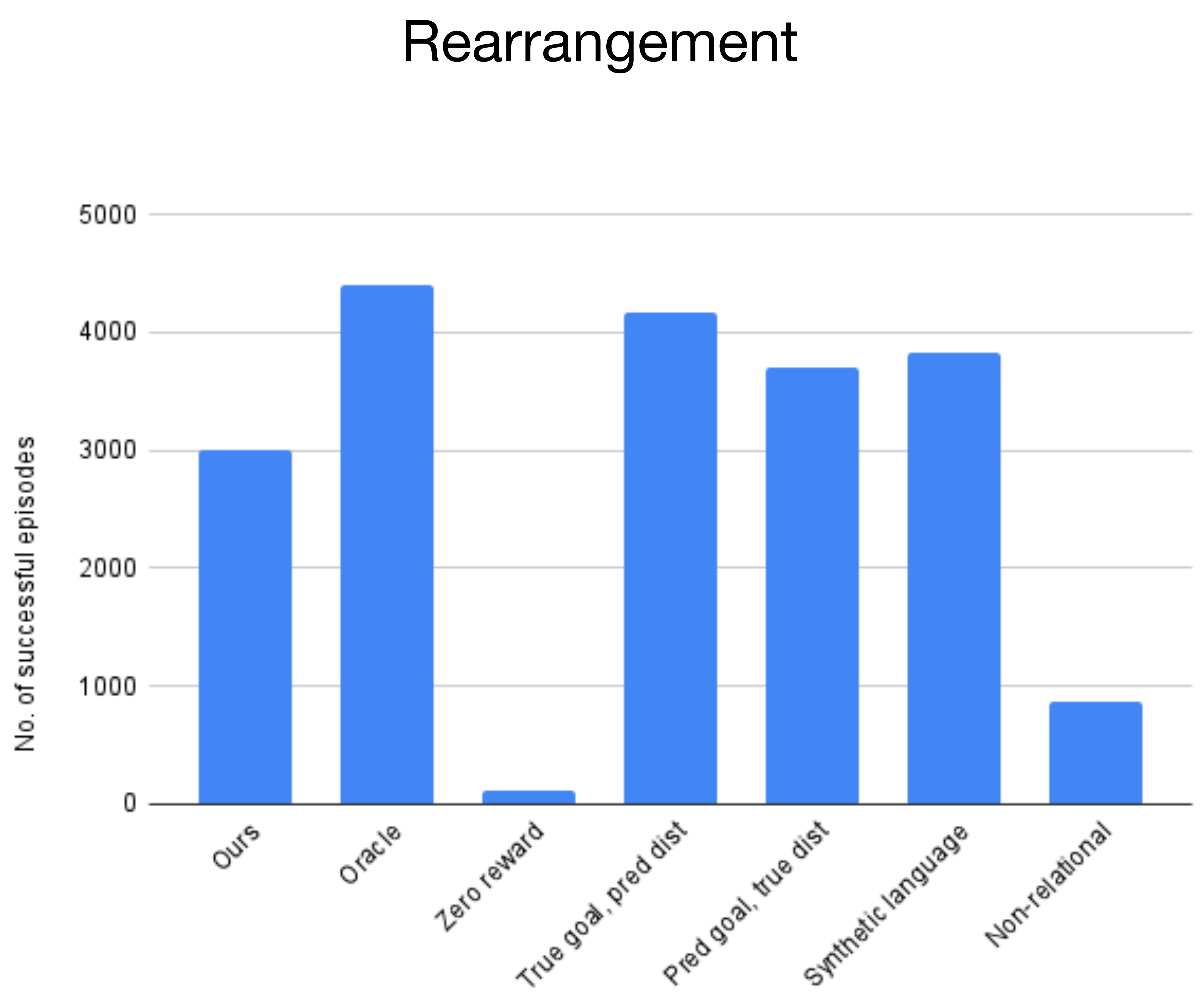


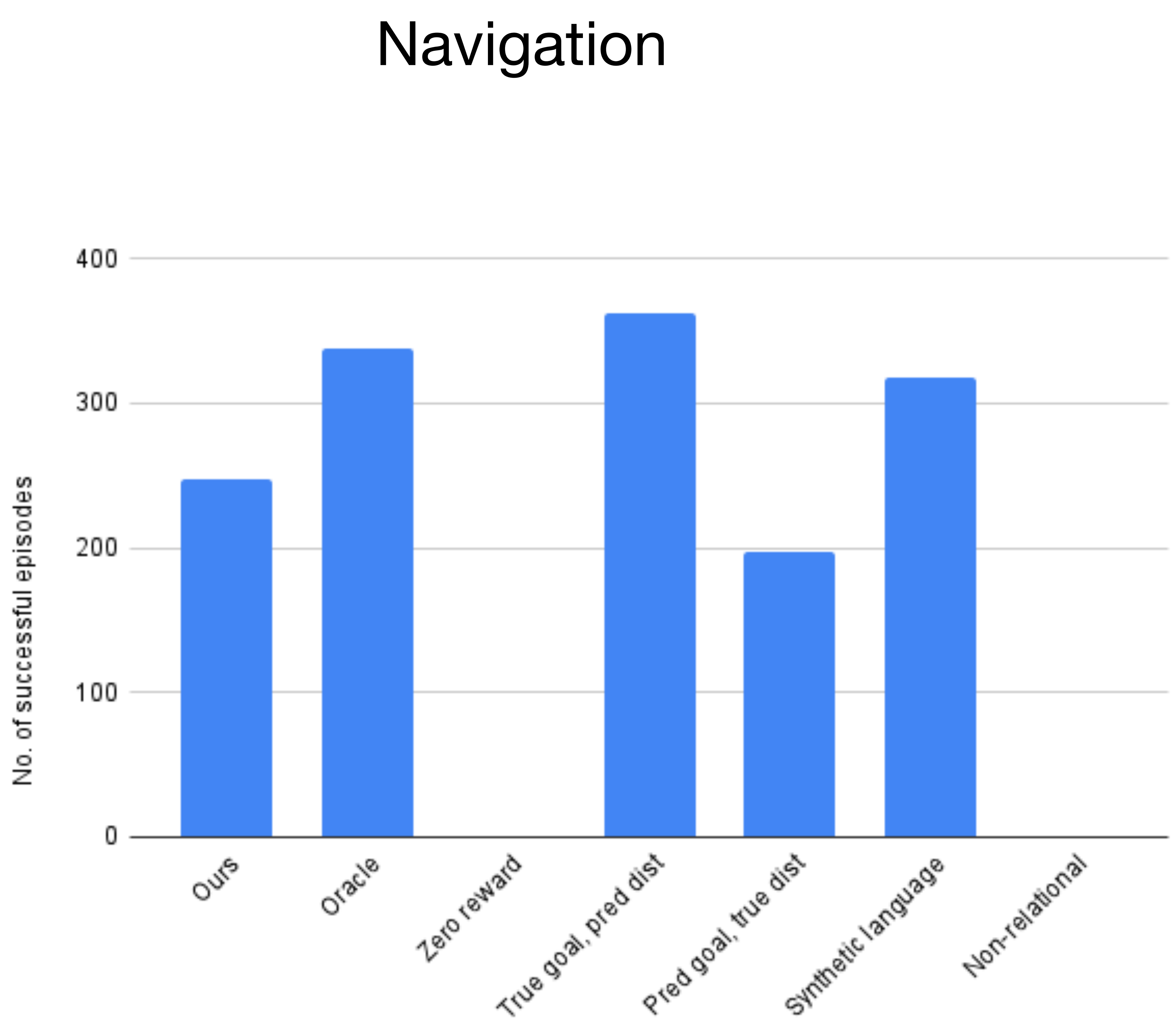
Navigation



REIational Task Adaptation for Imitation with Language (RETAIL)

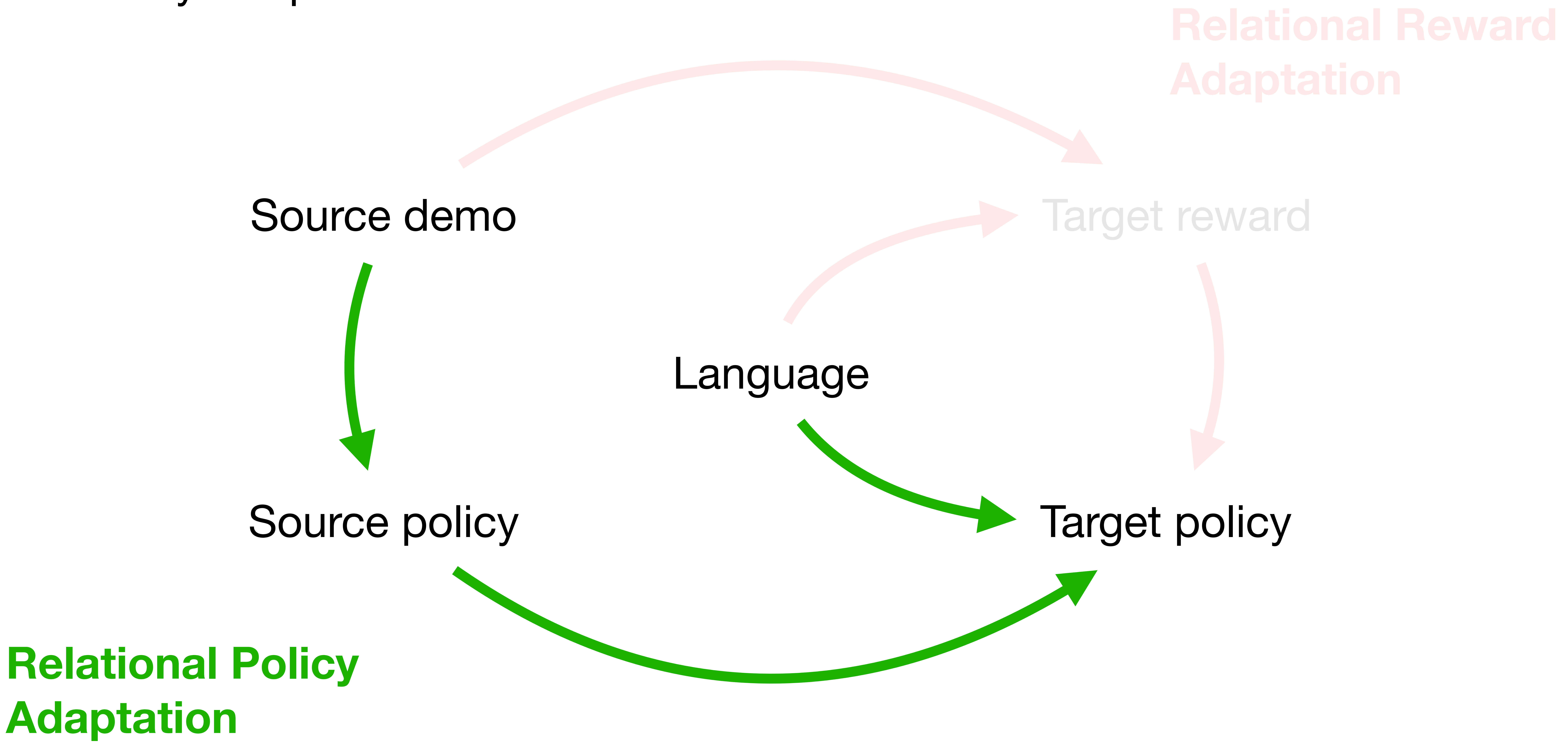
Relational Reward Adaptation: Experiments





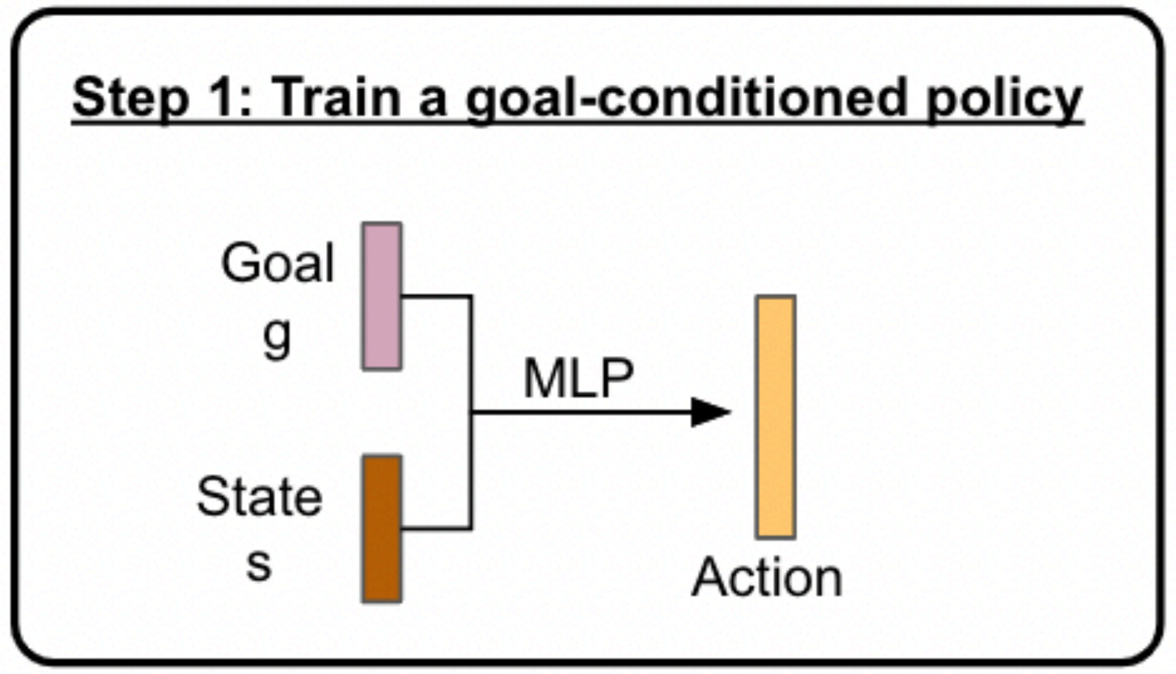
RElational Task Adaptation for Imitation with Language (RETAIL)

Relational Policy Adaptation



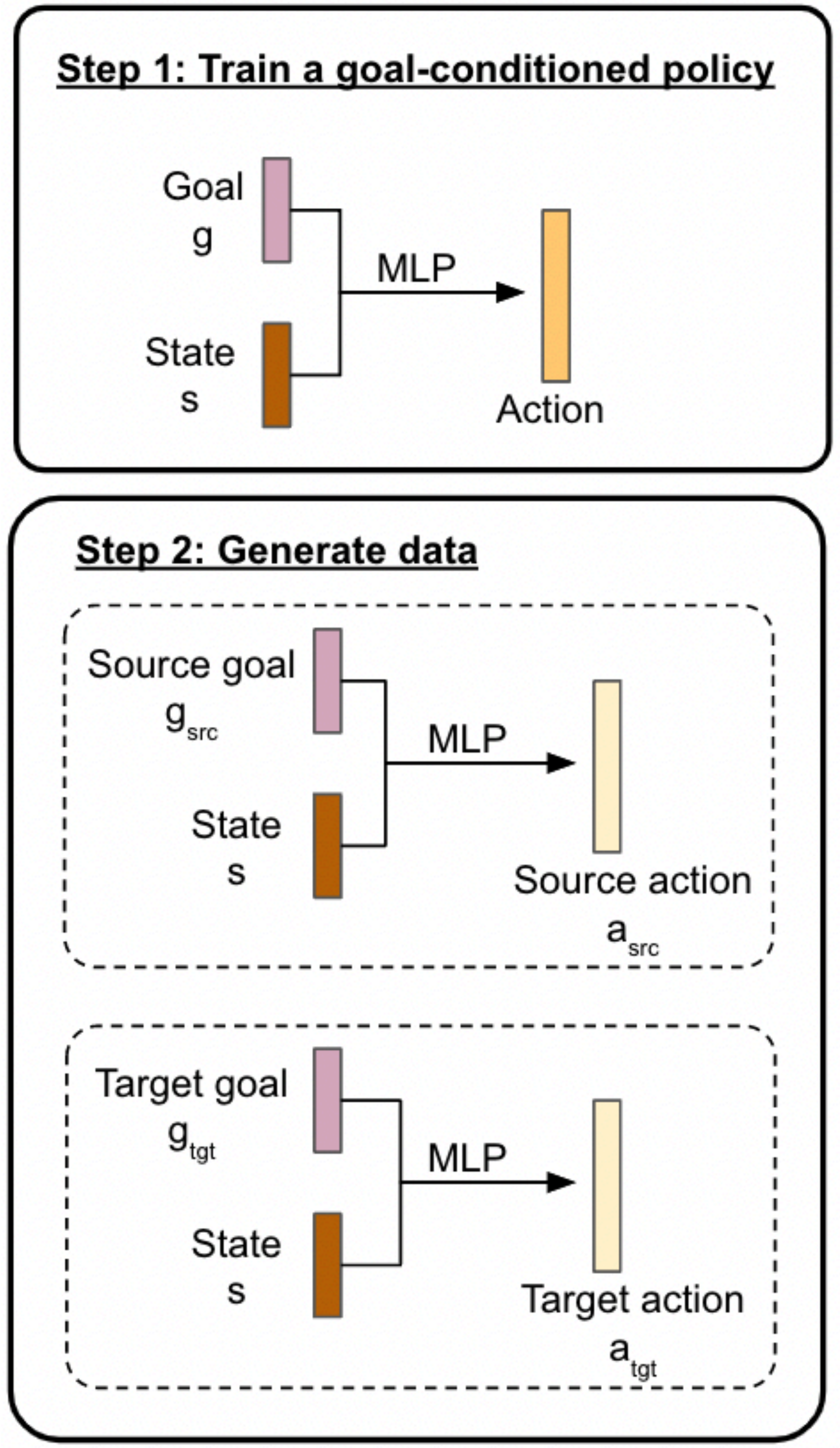
REllational Task Adaptation for Imitation with Language (RETAIL)

Relational Policy Adaptation: Approach



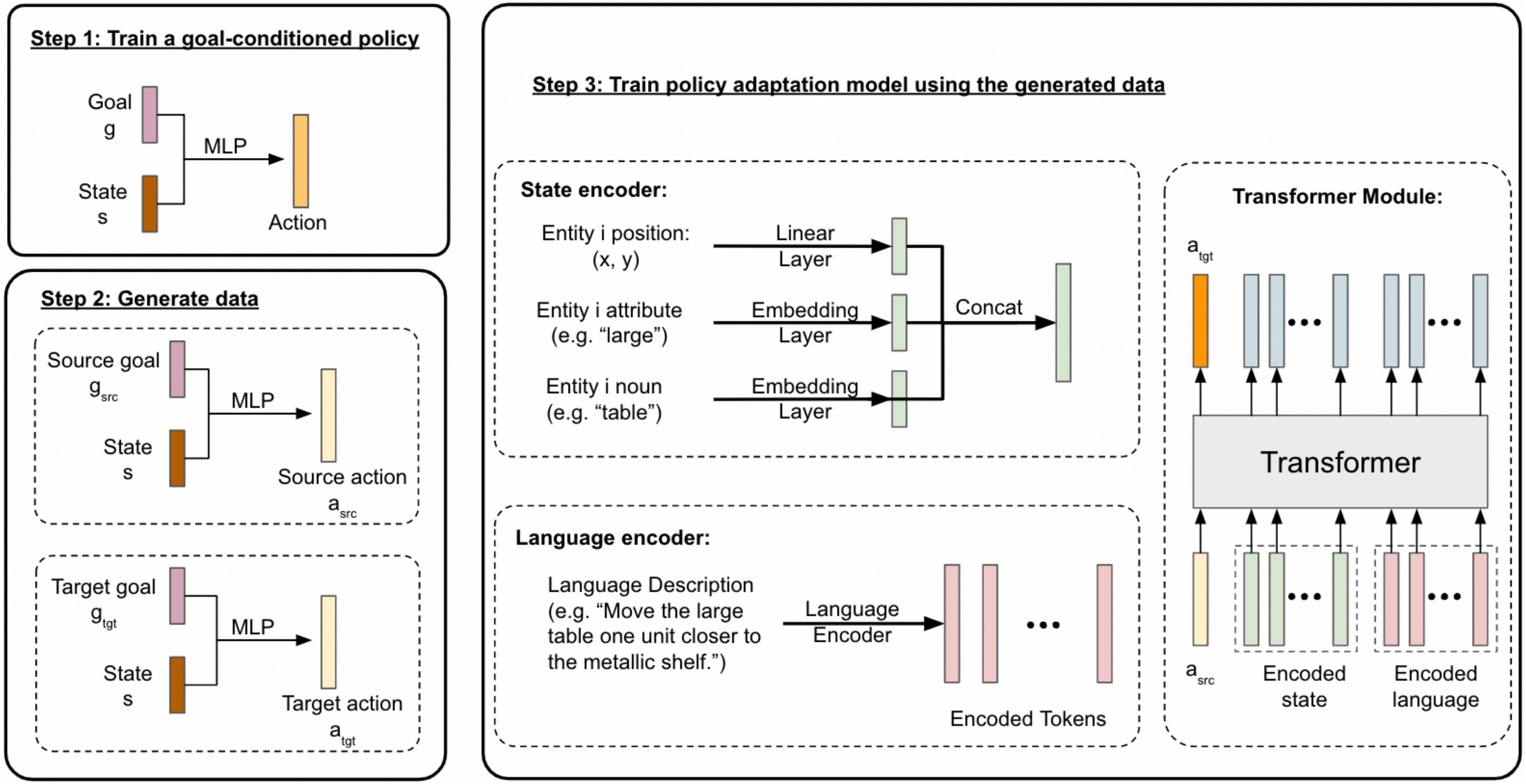
RElational Task Adaptation for Imitation with Language (RETAIL)

Relational Policy Adaptation: Approach



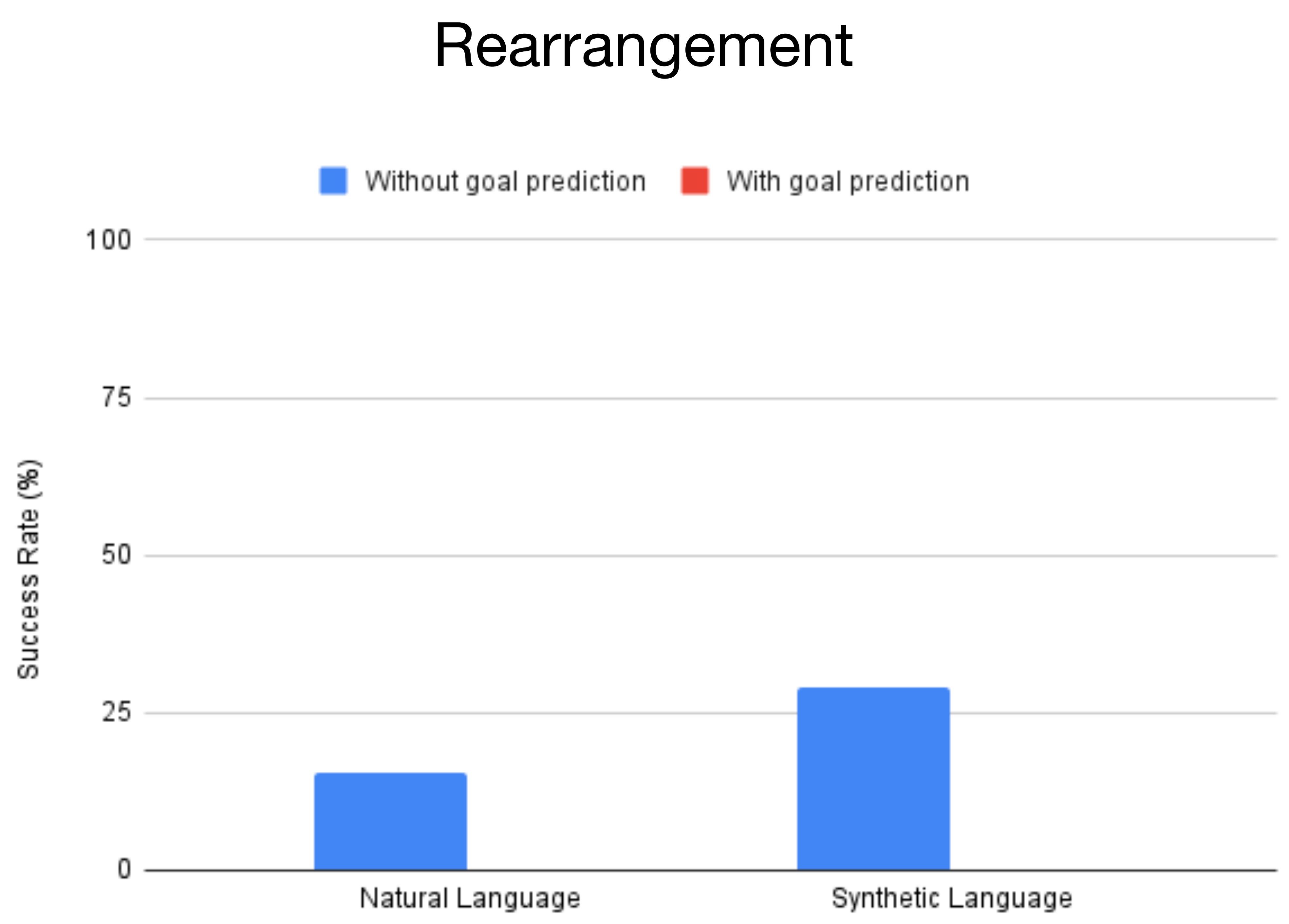
RElational Task Adaptation for Imitation with Language (RETAIL)

Relational Policy Adaptation: Approach



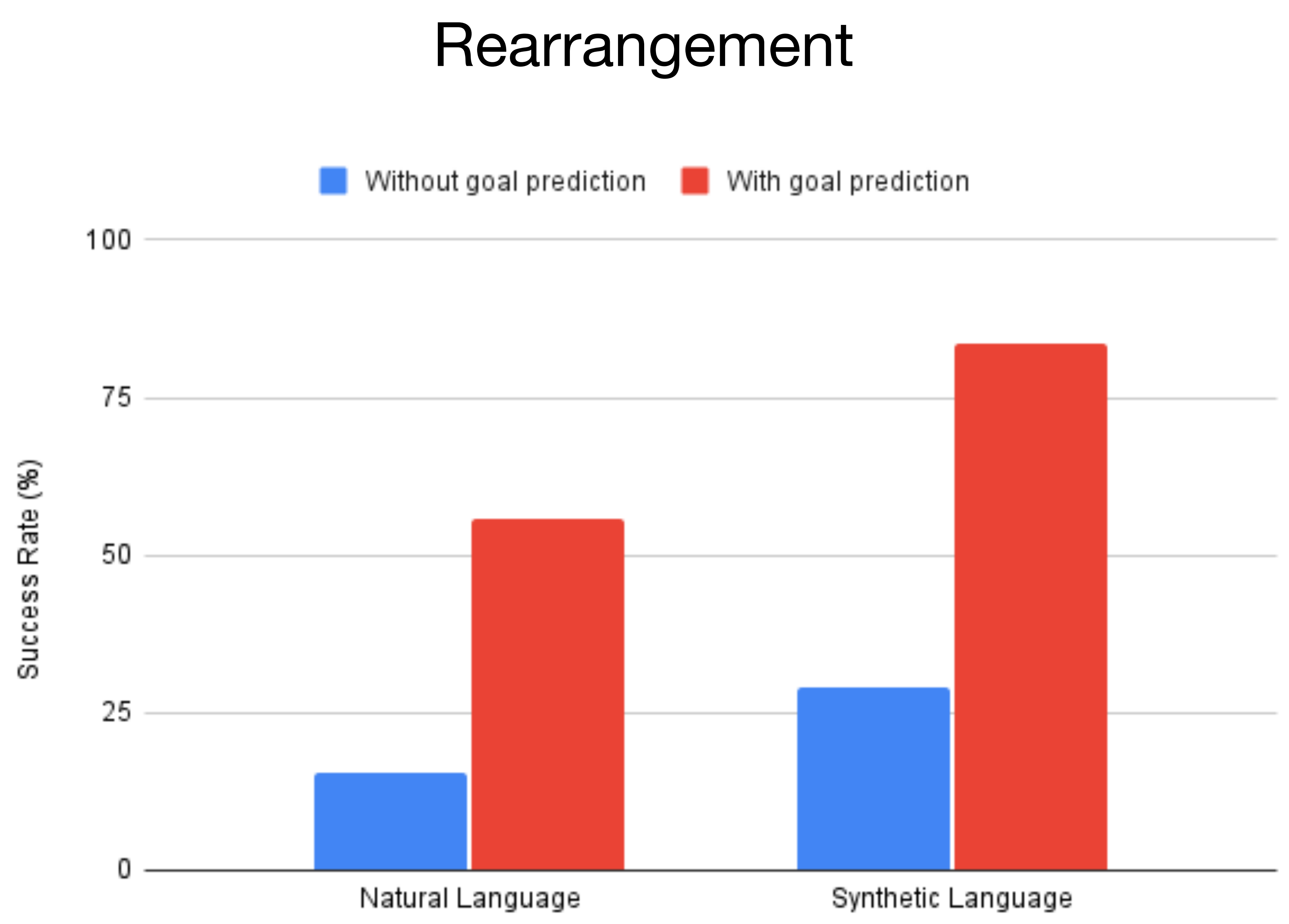
RElational Task Adaptation for Imitation with Language (RETAIL)

Relational Policy Adaptation: Experiments



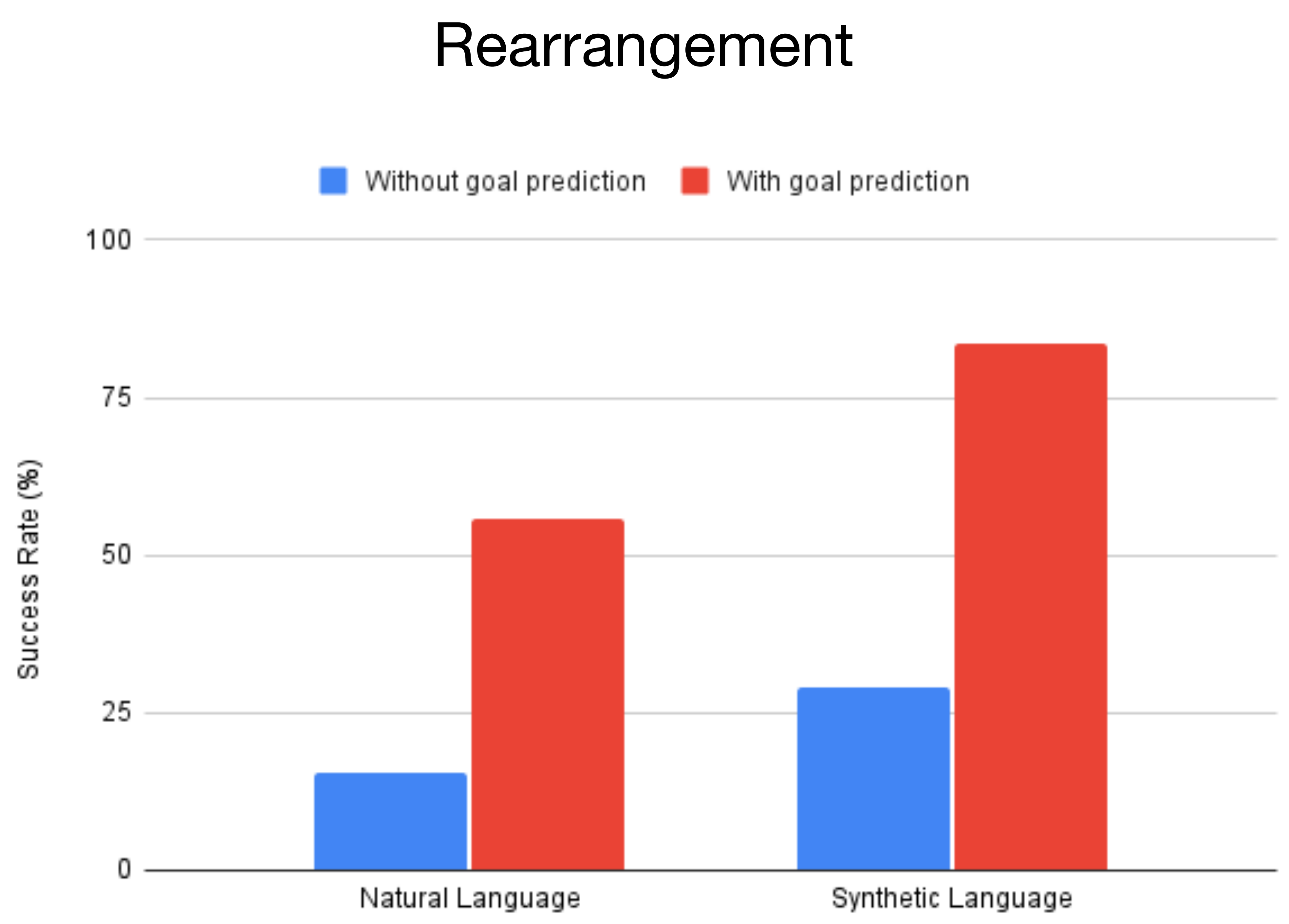
REIational Task Adaptation for Imitation with Language (RETAIL)

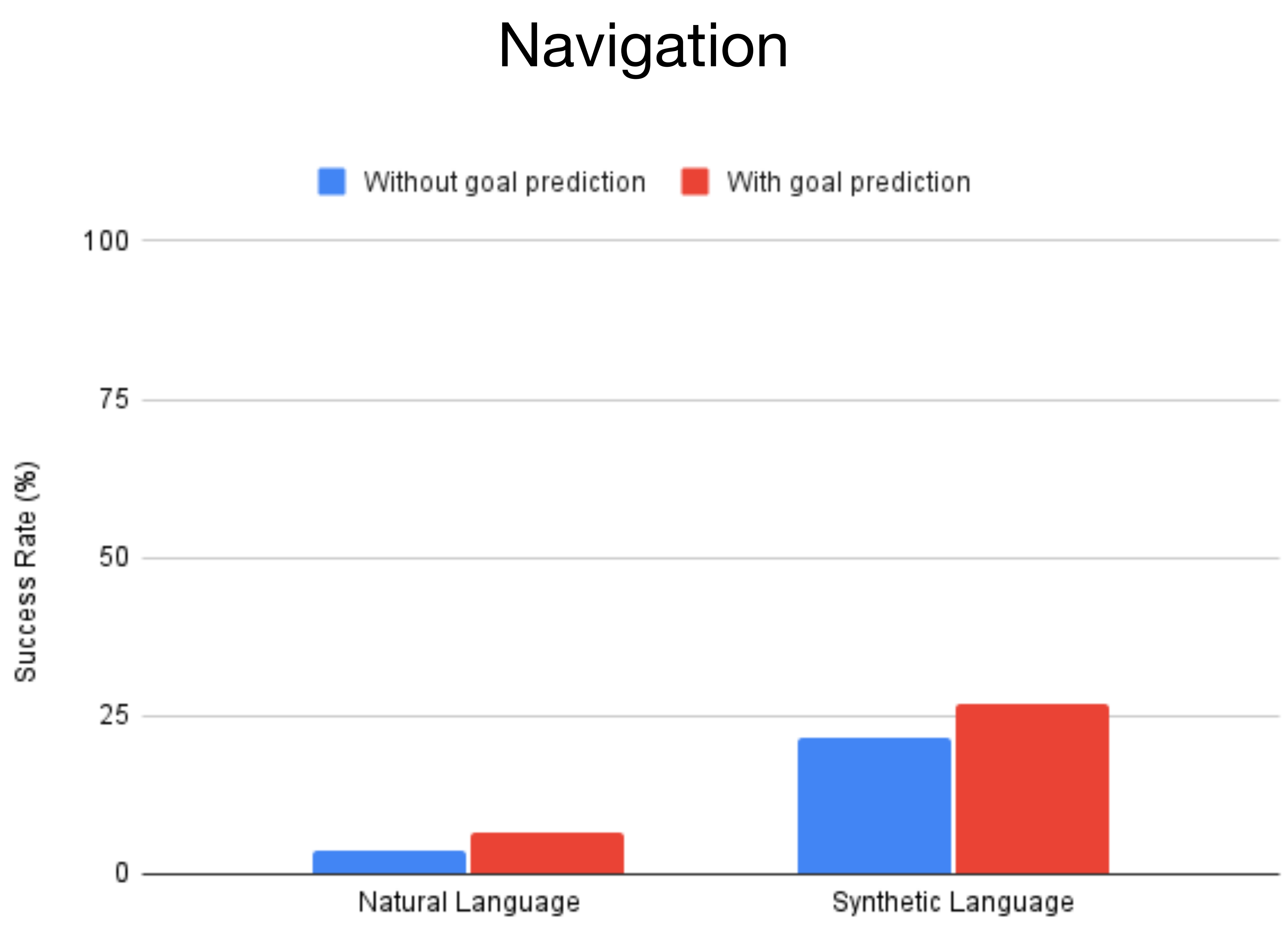
Relational Policy Adaptation: Experiments



RElational Task Adaptation for Imitation with Language (RETAIL)

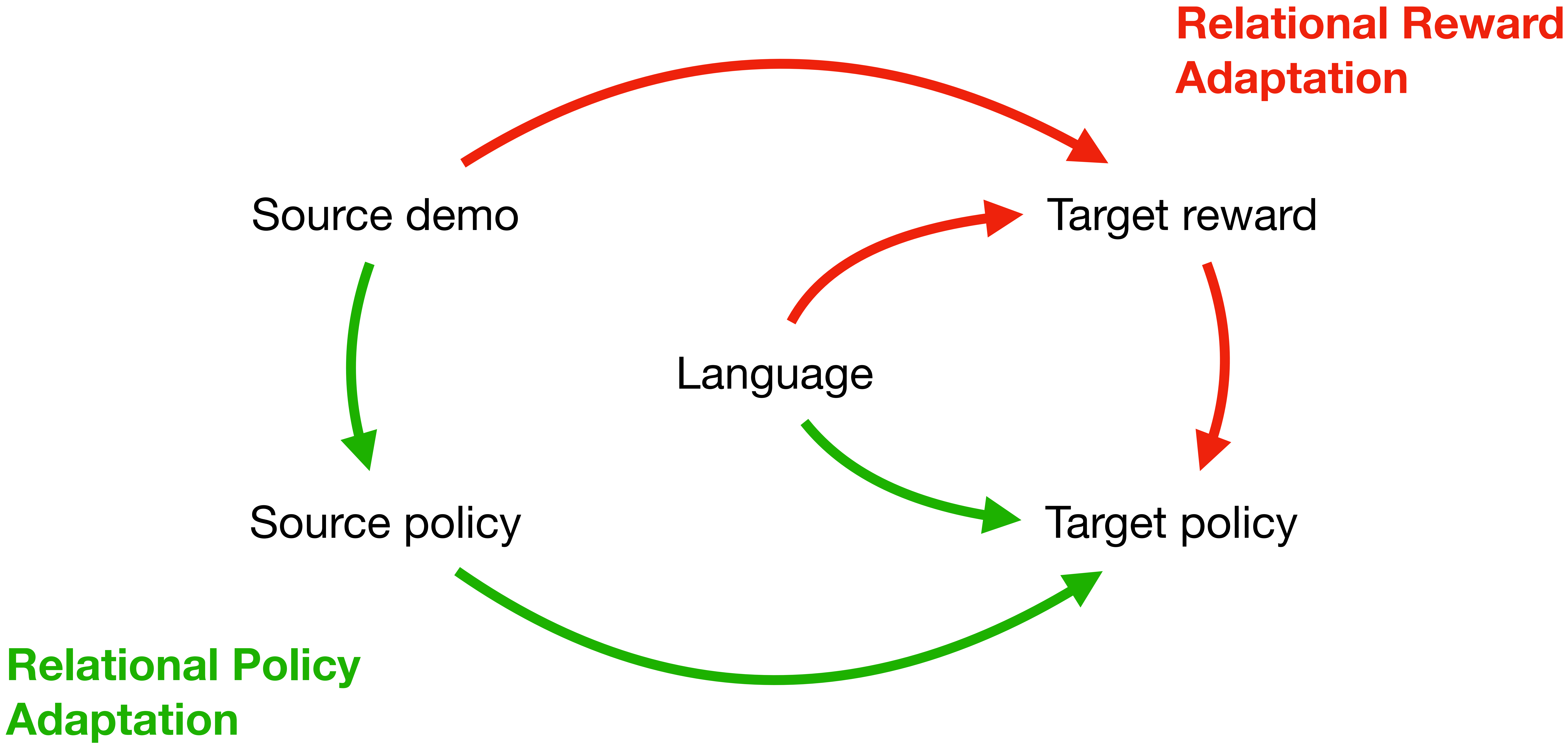
Relational Policy Adaptation: Experiments





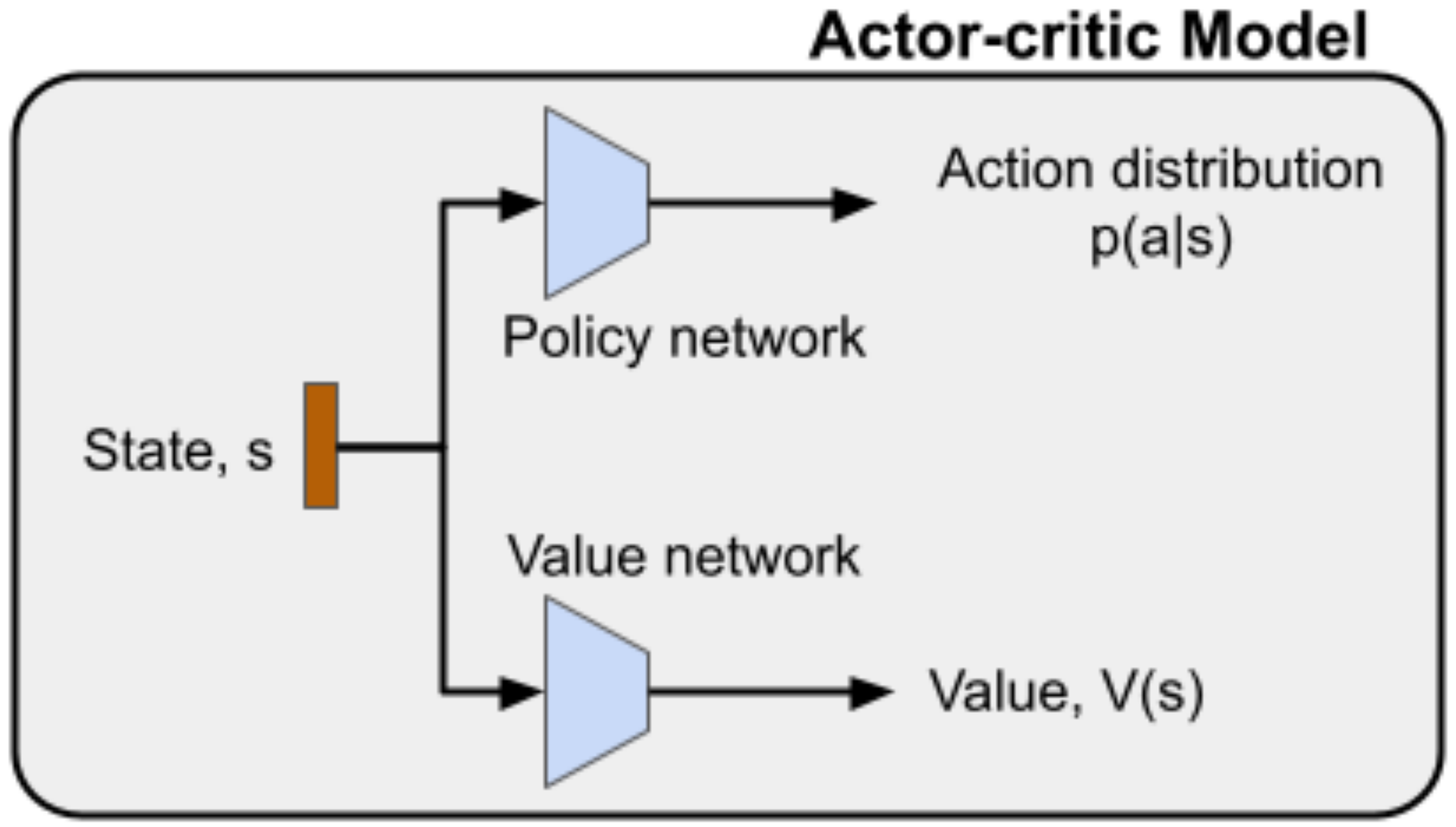
REIational Task Adaptation for Imitation with Language (RETAIL)

Combining Reward and Policy Adaptation



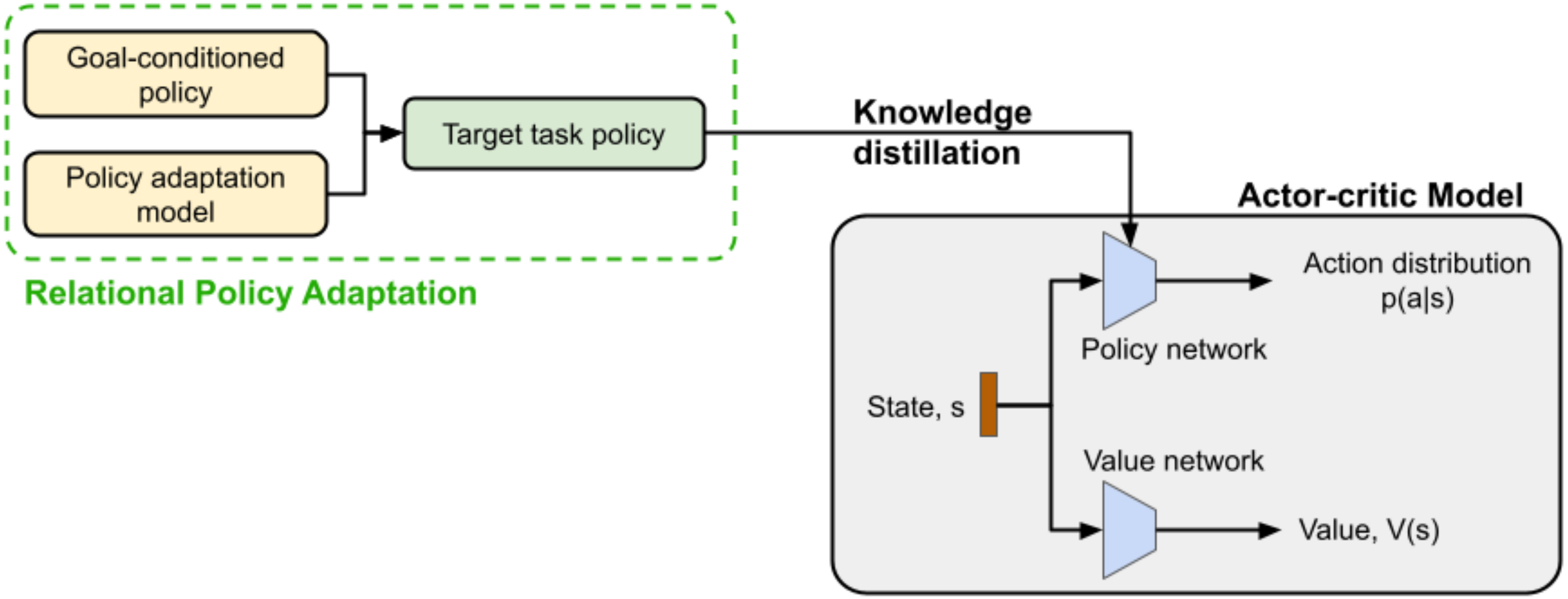
REIational Task Adaptation for Imitation with Language (RETAIL)

Combining Reward and Policy Adaptation



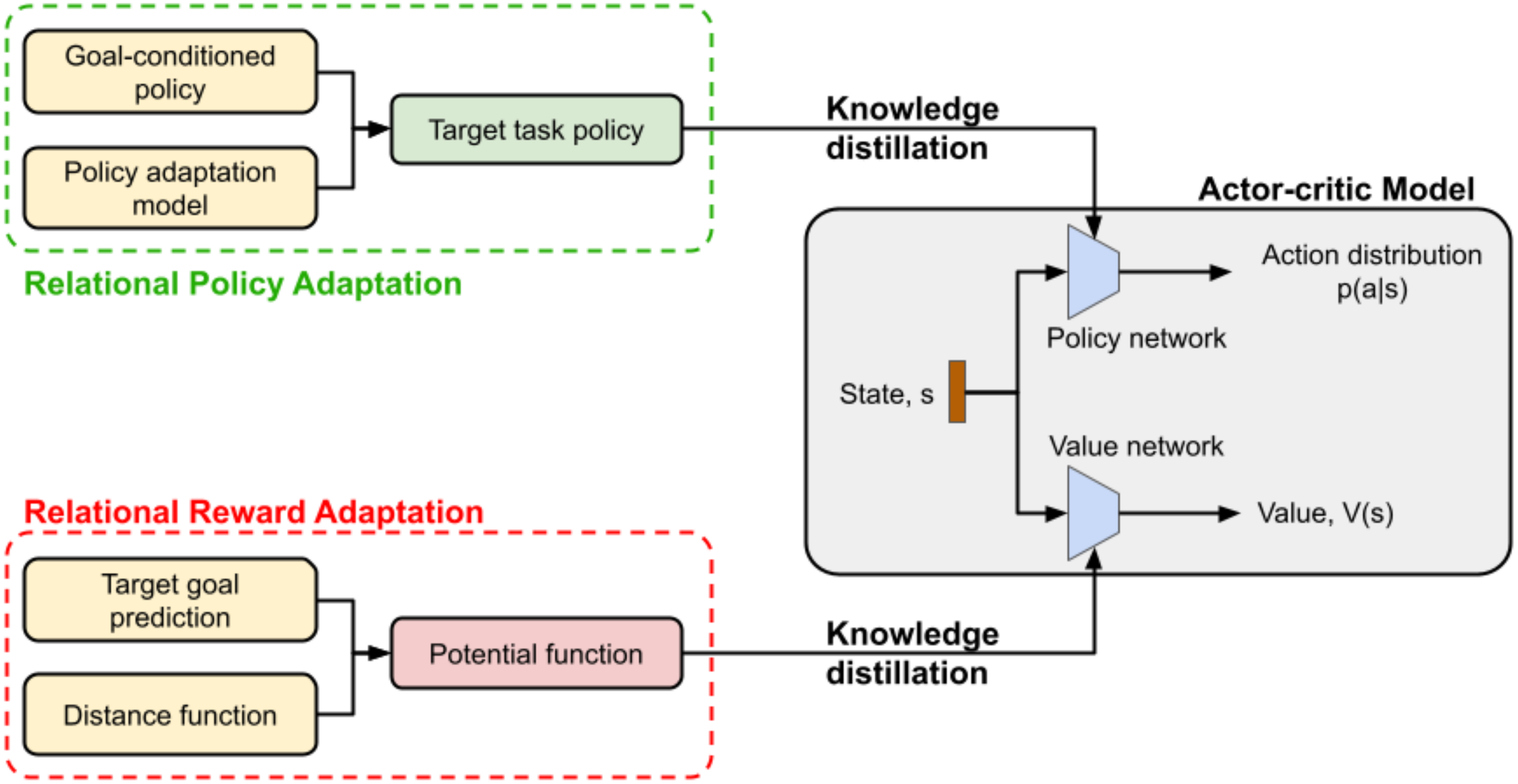
RElational Task Adaptation for Imitation with Language (RETAIL)

Combining Reward and Policy Adaptation



RElational Task Adaptation for Imitation with Language (RETAIL)

Combining Reward and Policy Adaptation



RElational Task Adaptation for Imitation with Language (RETAIL)

Combining Reward and Policy Adaptation

Knowledge Distillation:

- Use states from the demonstration data
- Losses:
  - Value network: Mean squared error
  - Policy network:
    - Mean squared error for continuous actions
    - Cross-entropy for discrete actions

REIational Task Adaptation for Imitation with Language (RETAIL)

Combining Reward and Policy Adaptation

Knowledge Distillation:

- Use states from the demonstration data
- Losses:
  - Value network: Mean squared error
  - Policy network:
    - Mean squared error for continuous actions
    - Cross-entropy for discrete actions

Finetuning using PPO:

- Networks initialized using knowledge distillation have low entropy

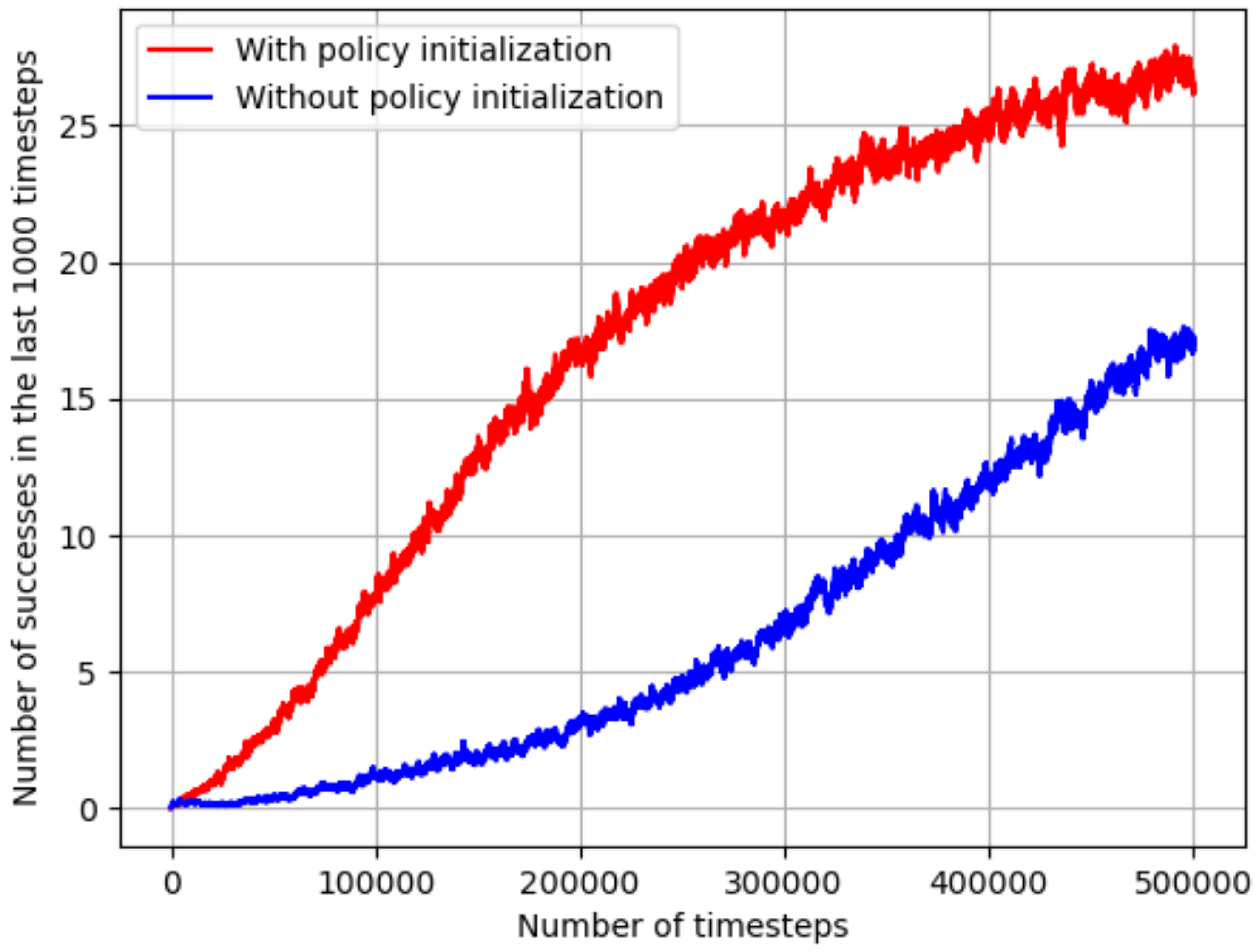
=> Continuous actions: Tune the standard deviation of the policy network

=> Discrete actions: Add an entropy loss during knowledge distillation

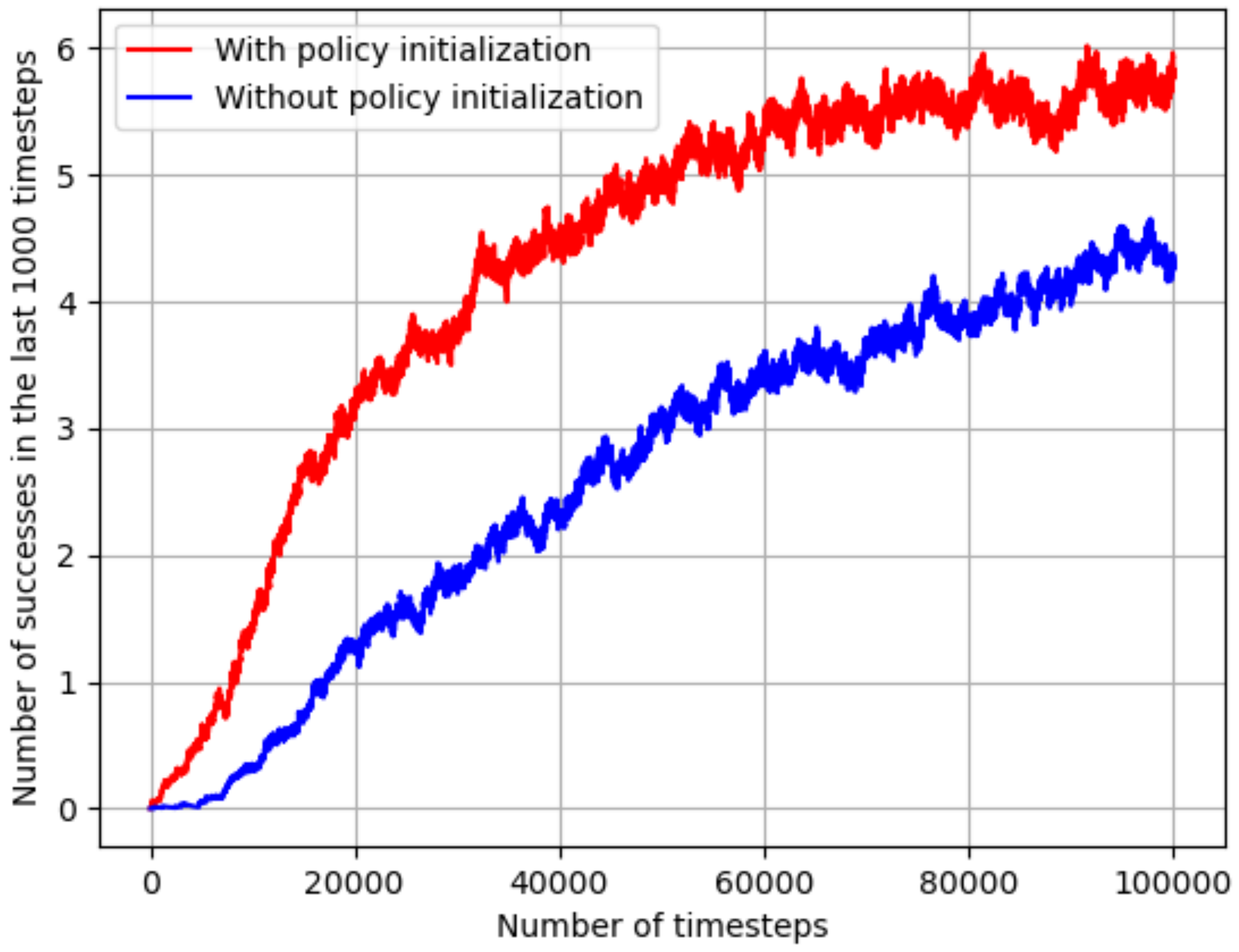
REIational Task Adaptation for Imitation with Language (RETAIL)

Combining Reward and Policy Adaptation

Rearrangement



Navigation



Talk Outline

Background

Core Contributions:

Sequential Decision Making

	Challenge	Solution
RL	Reward design	Language-based Rewards
IL	Many demos needed	Language-guided Task Adaptation

 +

LEARN

PixL2R

 +

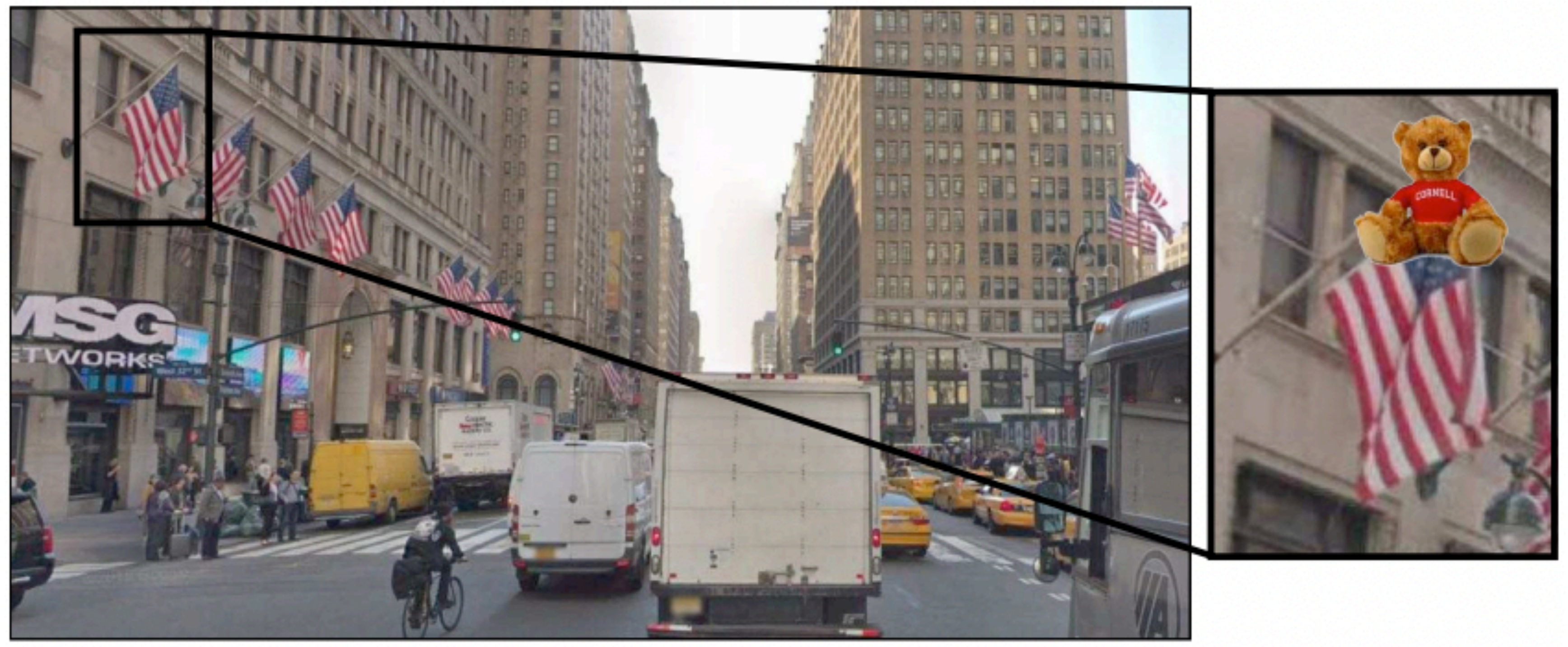
LARVA

RETAIL

Future Directions

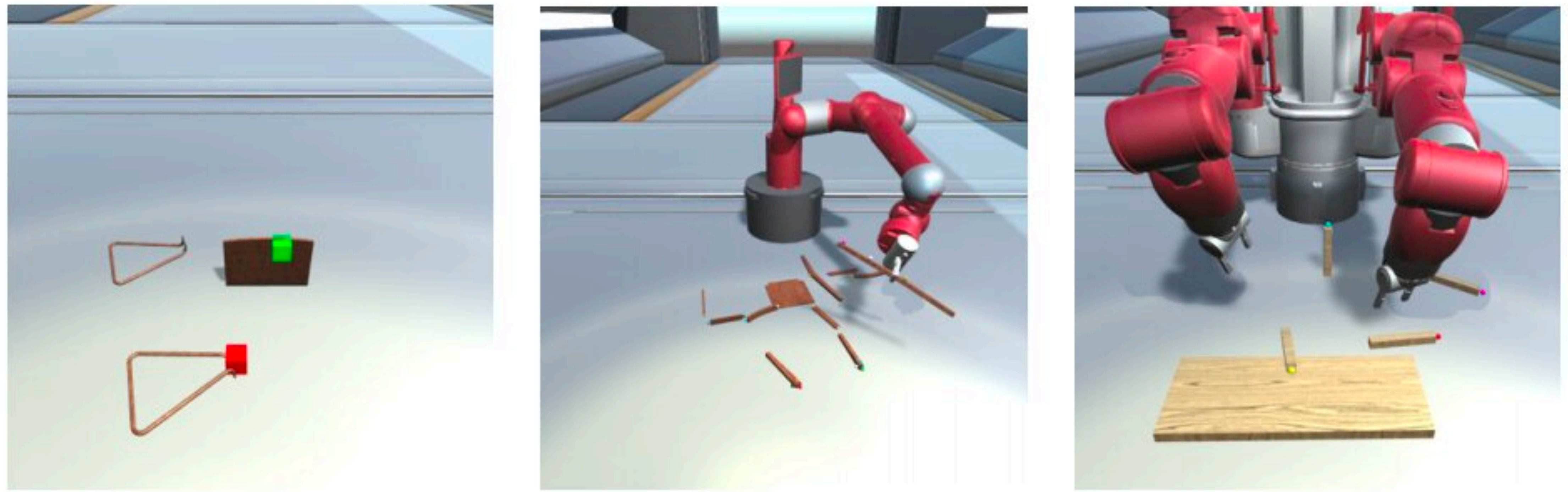
Future Work

Richer Domains



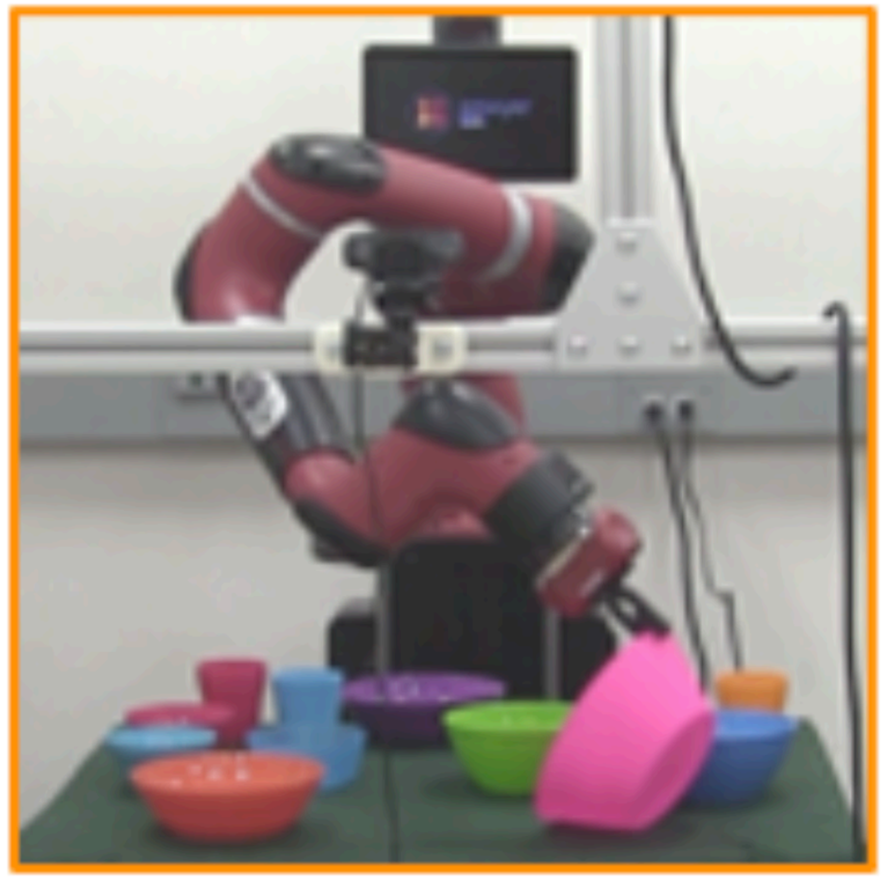
[Chen et al., 2019]

[Kolve et al., 2017]



(a) Cursor (b) Sawyer (c) Baxter

[Lee et al., 2021]



[Mandlekar et al., 2018]

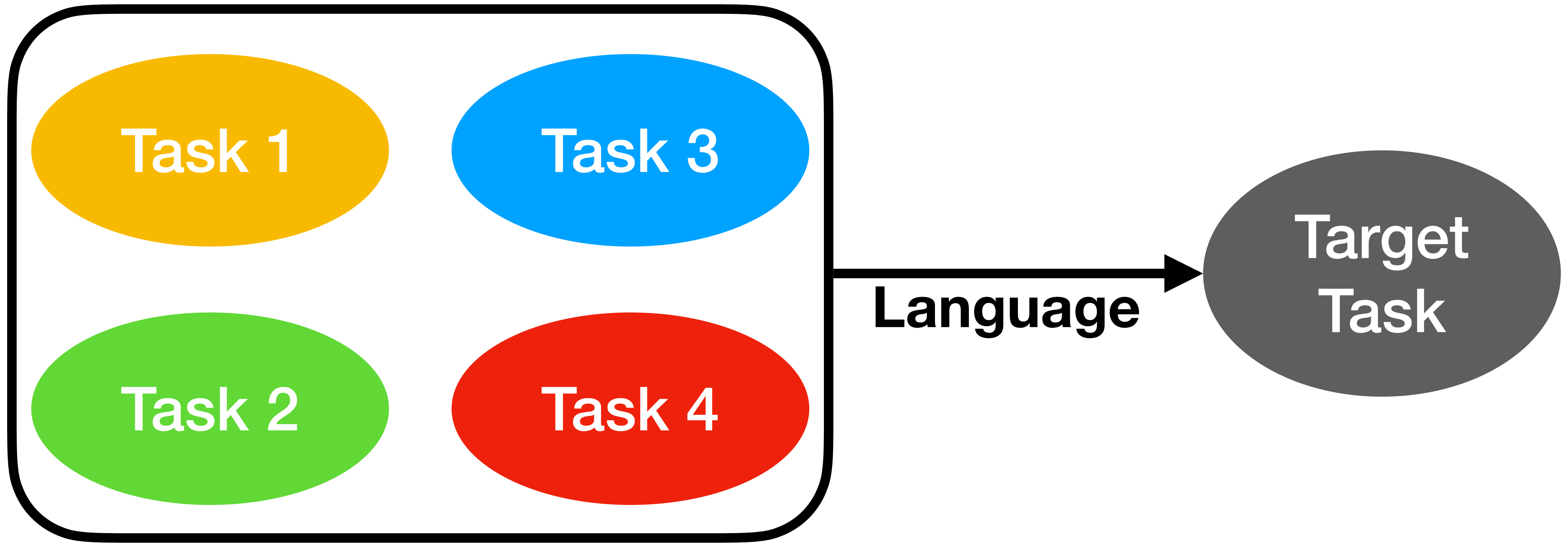
Future Work

Hierarchical Tasks

[Shridhar et al., 2020]

Future Work

Task Adaptation with Multiple Source Tasks

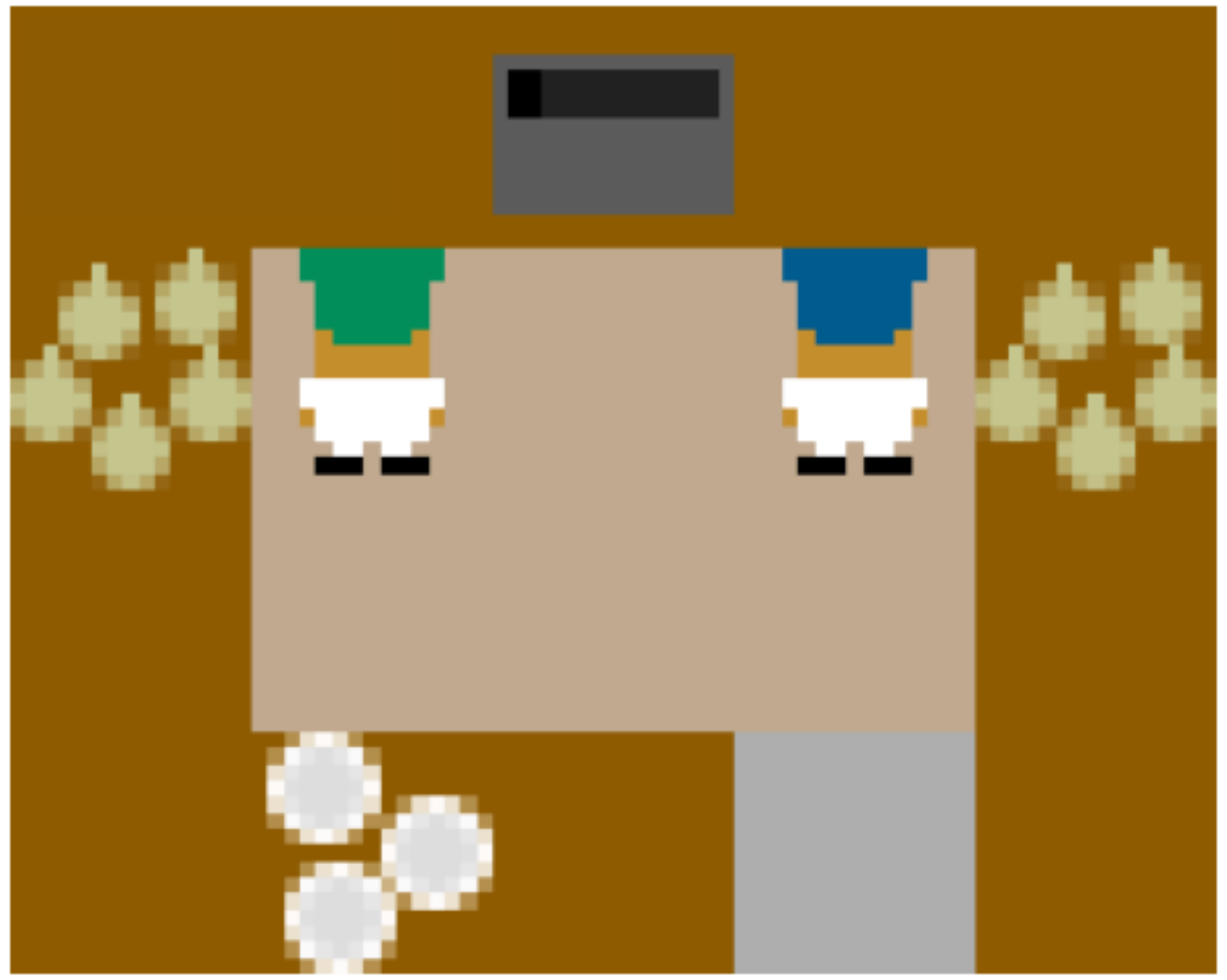


Future Work

Language-aided Imitation Learning

Humans use linguistic cues when giving demonstrations to other humans, e.g., “Turn off the heat when the water starts boiling”

Gridworld Cooking / Repairing:
- Language: What to do, Why
- Demonstrations: How



[Carroll et al., 2019]

The Big Picture…

Language can be used in a lot of ways:

- Communicating the task
- Providing feedback
- Guiding the agent to focus on the important aspects of the task
- Enabling the agent to ask clarification questions

The Big Picture…

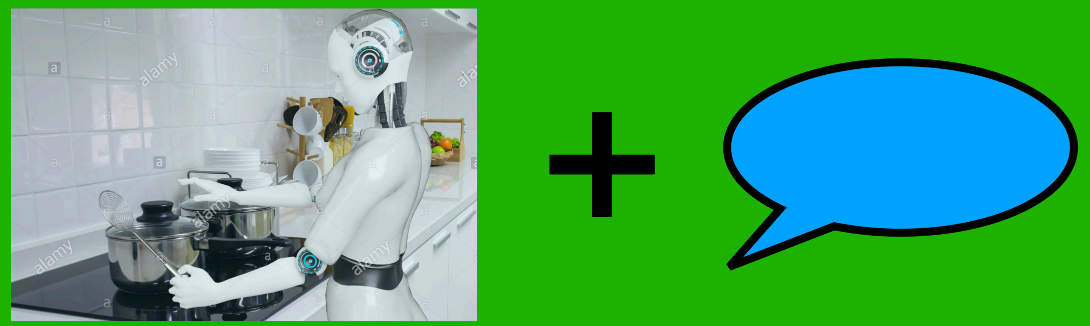
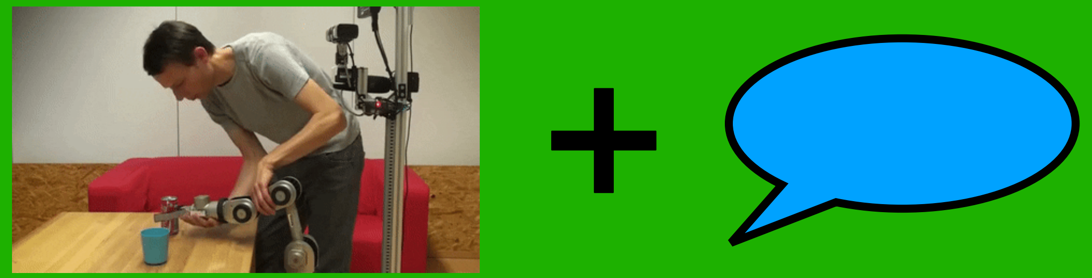
Language can be used in a lot of ways:

- Communicating the task — **Focus of this dissertation**
- Providing feedback
- Guiding the agent to focus on the important aspects of the task

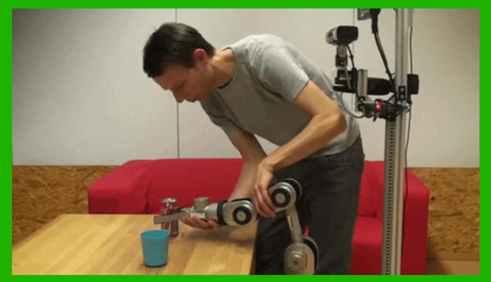
Also related to these (Providing feedback; Guiding the agent to focus on the important aspects of the task)

- Enabling the agent to ask clarification questions

Conclusion

		
Problem Settings	Rewards + Task Description	Source demo + Language describing the difference
Algorithms	LEARN PixL2R	LARVA RETAIL
Benchmarks	Montezuma's Revenge Robot Manipulation	Organizer Relational Rearrangement Relational Navigation

Conclusion

	 +	 +
Problem Settings	Rewards + Task Description	Source demo + Language describing the difference
Algorithms	LEARN PixL2R	LARVA RETAIL
Benchmarks	Montezuma's Revenge Robot Manipulation	Organizer Relational Rearrangement Relational Navigation

Acknowledgements

Raymond Mooney

Scott Niekum