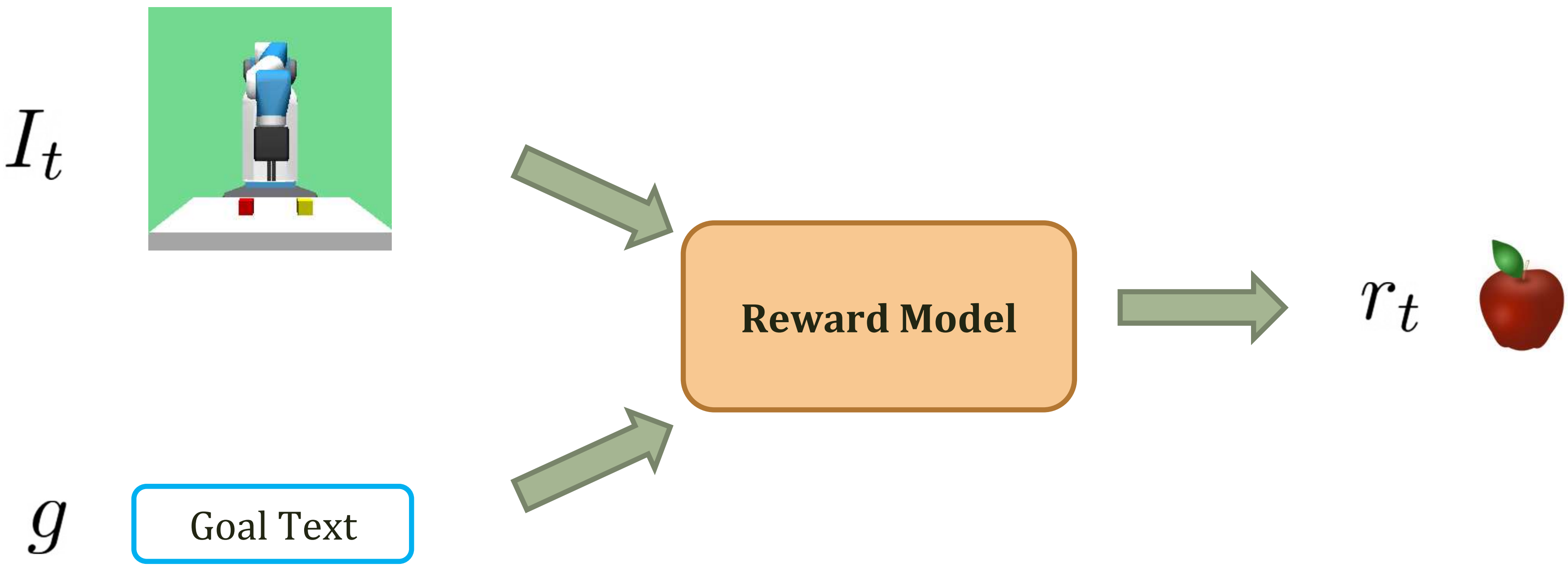
Zero-shot Reward Specification via Grounded Natural Language

Parsa Mahmoudieh, Deepak Pathak, Trevor Darrell

Language Conditioned Reward

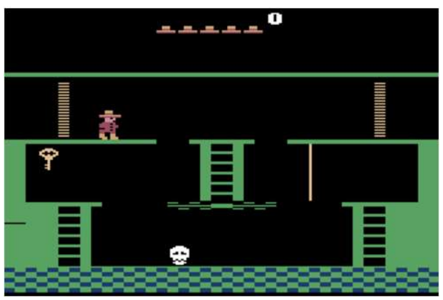


Reward Signal Resources

Reward Signals: typically need access to state or a human

Reward Signal Resources

Reward Signals: typically need access to state or a human

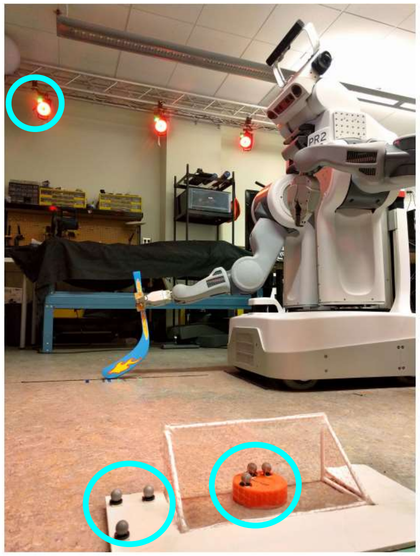


Mnih et al. (2013)

Reward Signal Resources

Reward Signals: typically need access to state or a human

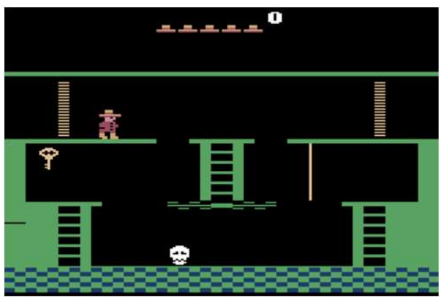
Mnih et al. (2013)



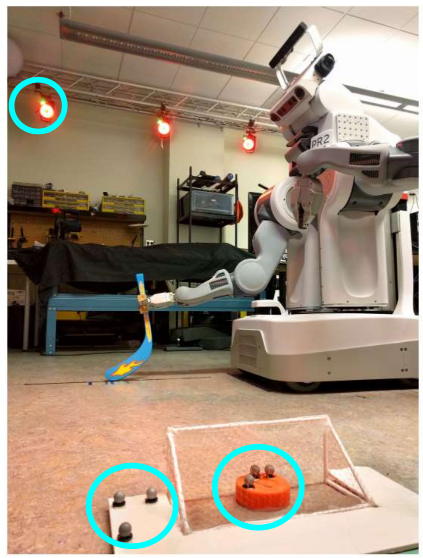
Chebotar et al (2017)

Reward Signal Resources

Reward Signals: typically need access to state or a human



Mnih et al. (2013)



Chebotar et al (2017)

Reward Signal Resources

How can we avoid reward functions that need access to

state, human evaluator, demonstrations, or goal images

Reward Signal Resources

How can we avoid reward functions that need access to

state, human evaluator, demonstrations, or goal images

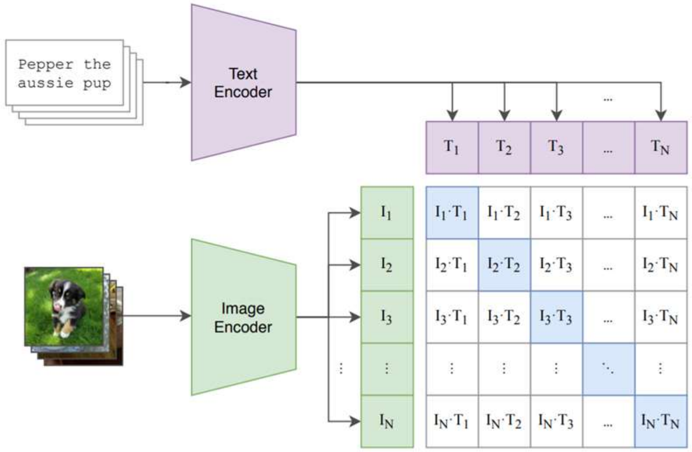
Can we leverage large language vision models to avoid this?

Reward Signal Resources

How can we avoid reward functions that need access to

state, human evaluator, demonstrations, or goal images

Can we leverage large language vision models to avoid this?



CLIP: Radford et al. Learning Transferable Visual Models From Natural Language Supervision

Vanilla Dot Product

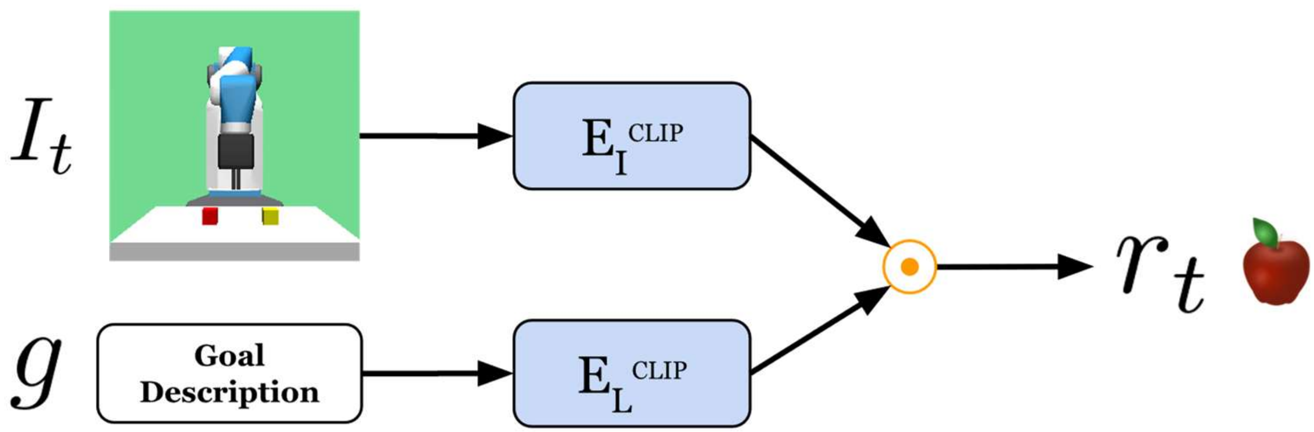
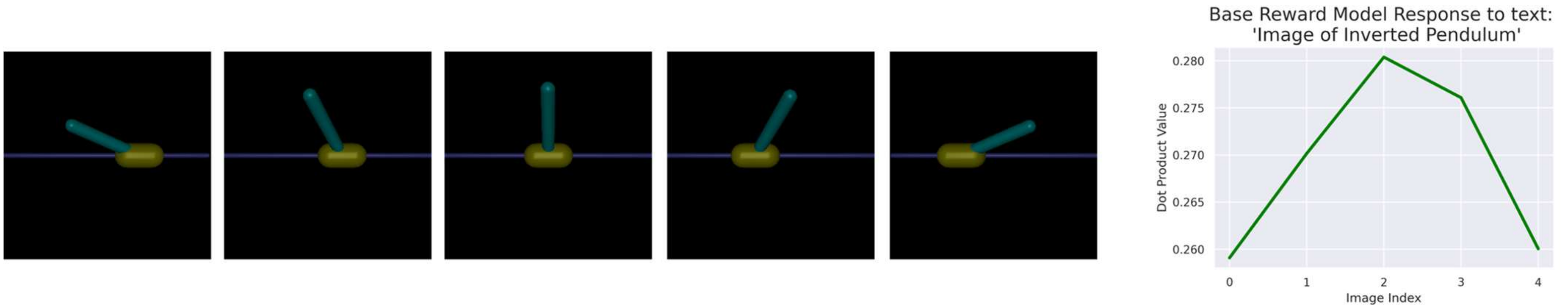
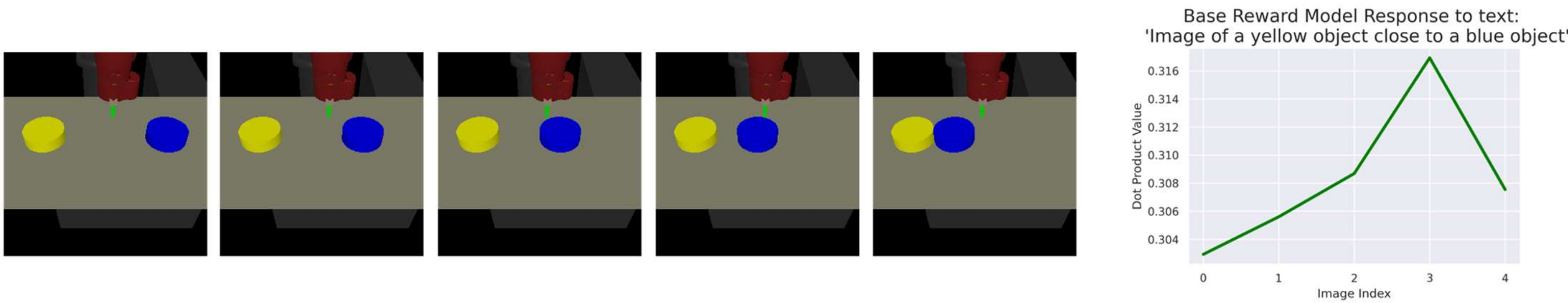


Image + Goal Description => Task completion score

Vanilla Dot Product

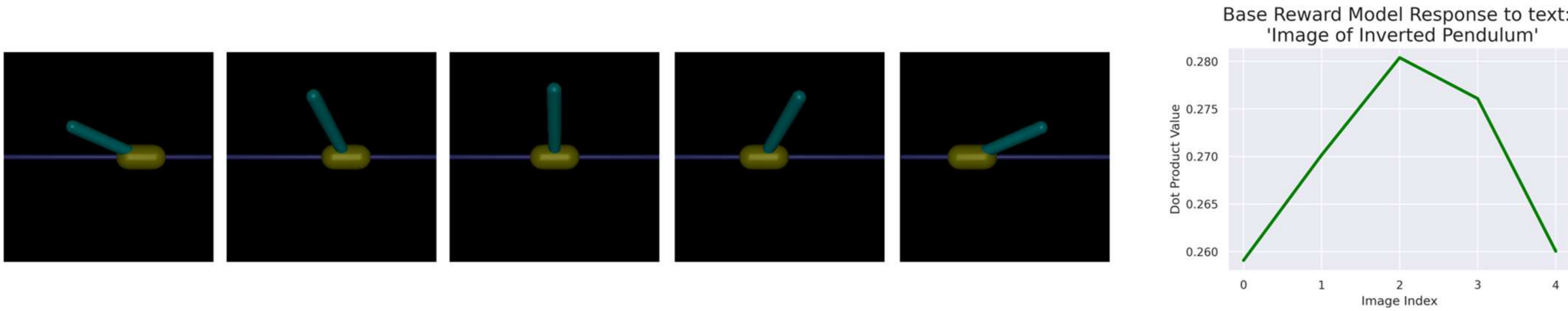


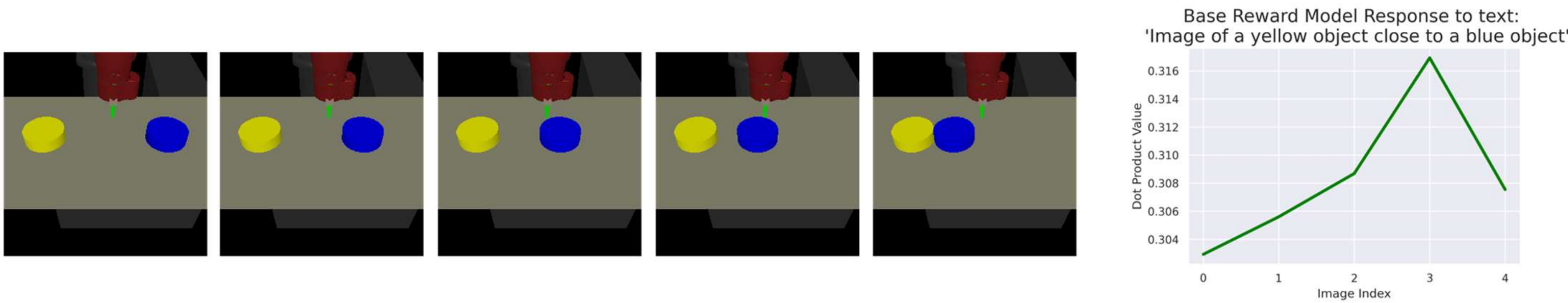


Bad at spatial relationships

Good at discriminating Nouns

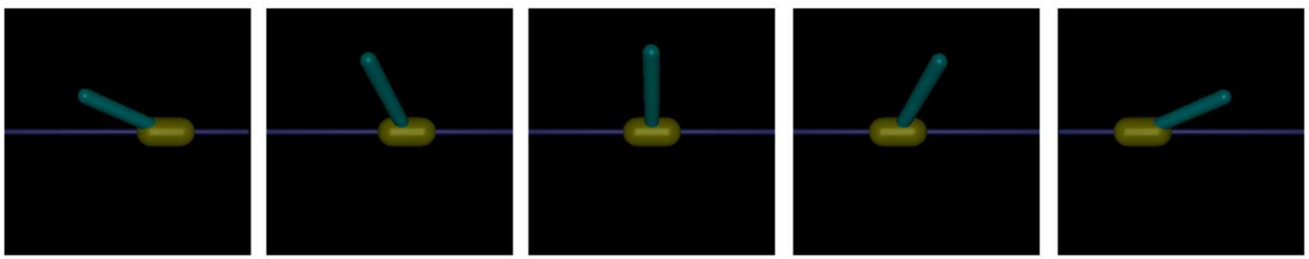
Vanilla Dot Product

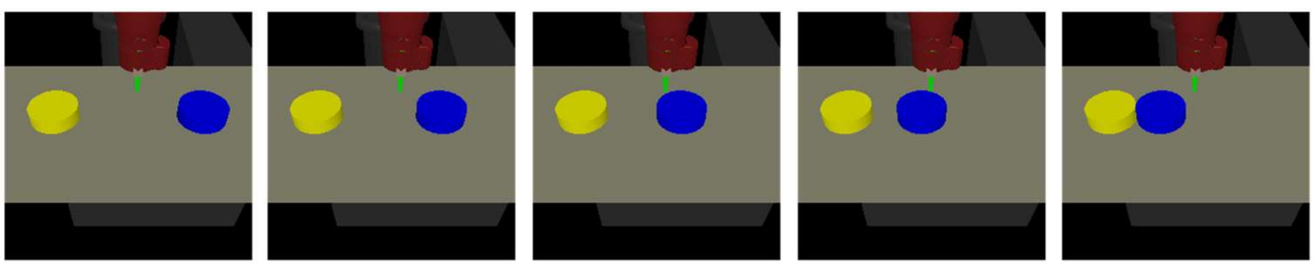




How can we leverage this?

Vanilla Dot Product





GradCAM can extract spatial information of semantics in Conv layers

Grad-CAM Background

Grad-CAM provides a way to see what part of a spatial feature map contributes the most to predicting a certain class

1. Selvaraju et al. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization

Grad-CAM Background

Grad-CAM provides a way to see what part of a spatial feature map contributes the most to predicting a certain class

$$\frac{\partial y^c}{\partial A_{ij}^k}$$

Delta in Probability output for class C

Delta in K^{th} feature map in activation layer A

1. Selvaraju et al. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization

Grad-CAM Background

Avg Class Score Response for Feature map k

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

1. Selvaraju et al. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization

Grad-CAM Background

Avg Class Score Response for Feature map k

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

Weighted fprop HeatMap

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k \right)}_{\text{linear combination}}$$

1. Selvaraju et al. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization

Grad-CAM Background

Avg Class Score Response for Feature map k

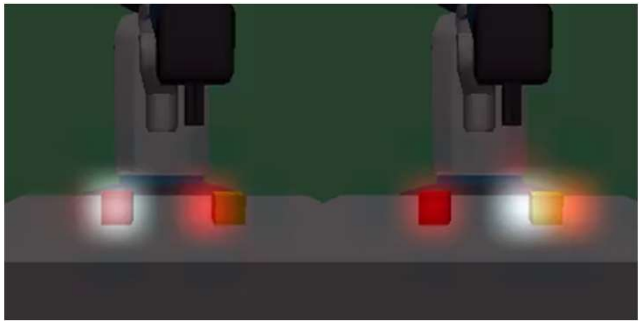
$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

Weighted fprop HeatMap

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

Grad-CAM on CLIP

Text emb: a red block

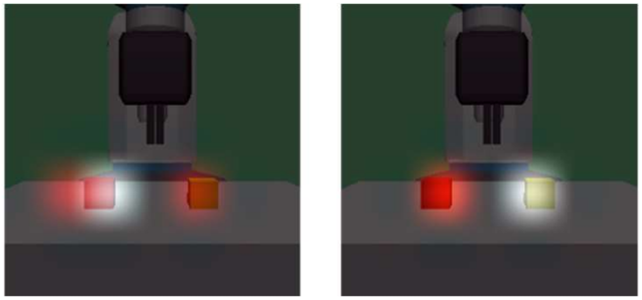


Text emb: a yellow block

1. Selvaraju et al. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization

Spatial language data generation

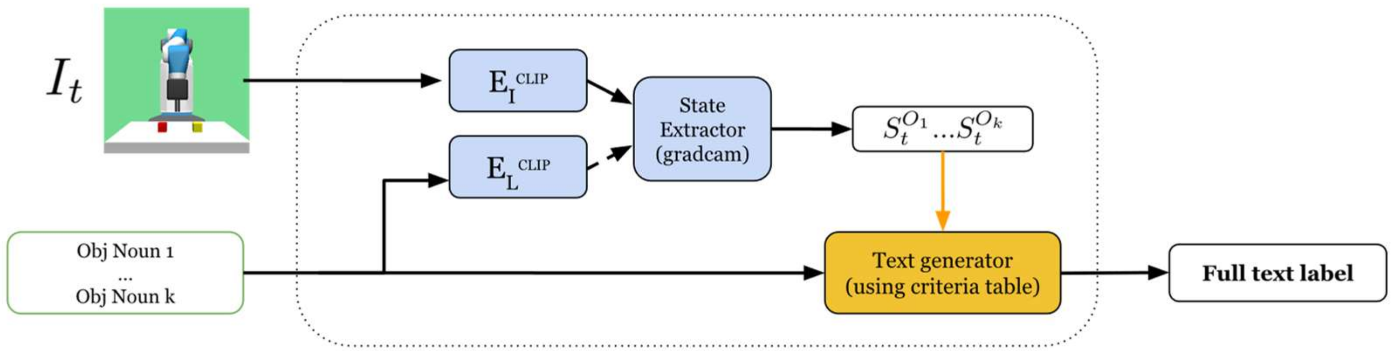
Spatial Language Label	Label Grounding Criteria
Obj1 on the left of Obj2	$O_x^2 > O_x^1$
Obj1 on the right of Obj2	$O_x^1 > O_x^2$
Obj1 on top of Obj2	$\|O_x^1 - O_x^2\| < \epsilon_1$ & $O_y^2 < O_y^1 < O_y^2 + \epsilon_2$
Obj1 below Obj2	$\|O_x^1 - O_x^2\| < \epsilon_1$ & $O_y^1 < O_y^2 < O_y^1 + \epsilon_2$
Obj1 in between Obj2, Obj3	$min(O_x^2, O_x^3) < O_x^1 < max(O_x^2, O_x^3)$
Obj1 in front of Obj2	$O_{x2}^1 > O_{x2}^2$
Obj1 behind Obj2	$O_{x2}^2 > O_{x2}^1$
Obj1 close to Obj2	$\|\|O_{xy}^1 - O_{xy}^2\|\|_2 < \epsilon$
Obj1 inside of Obj2	$\|\|O_{xy}^1 - O_{xy}^2\|\|_2 < \epsilon$



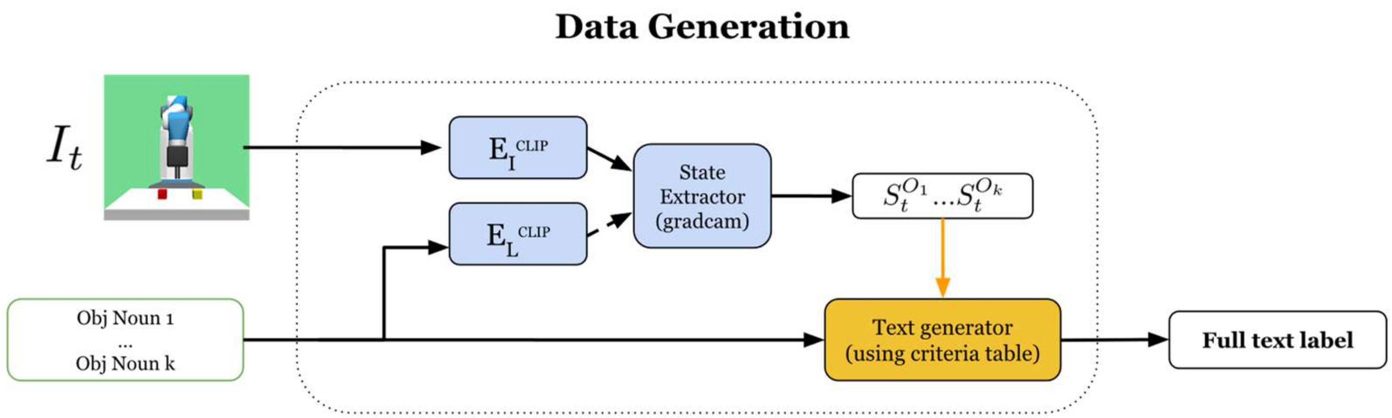
Red block on the left of yellow block

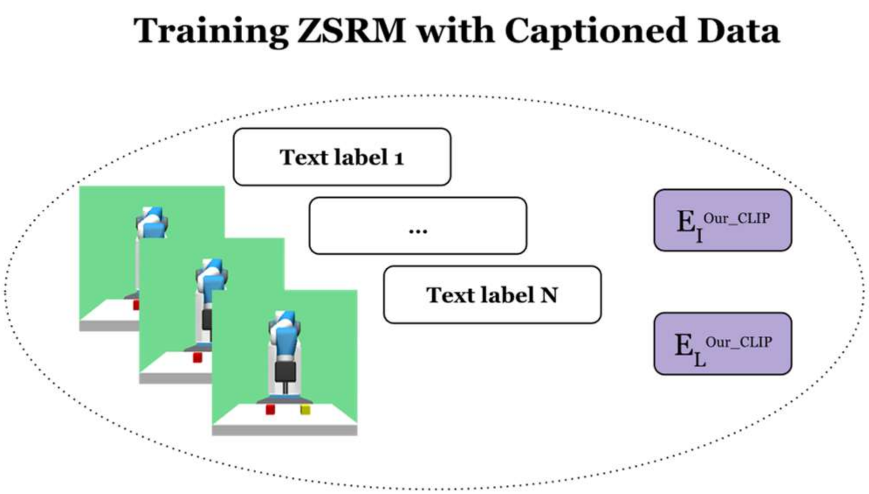
Method Overview

Data Generation

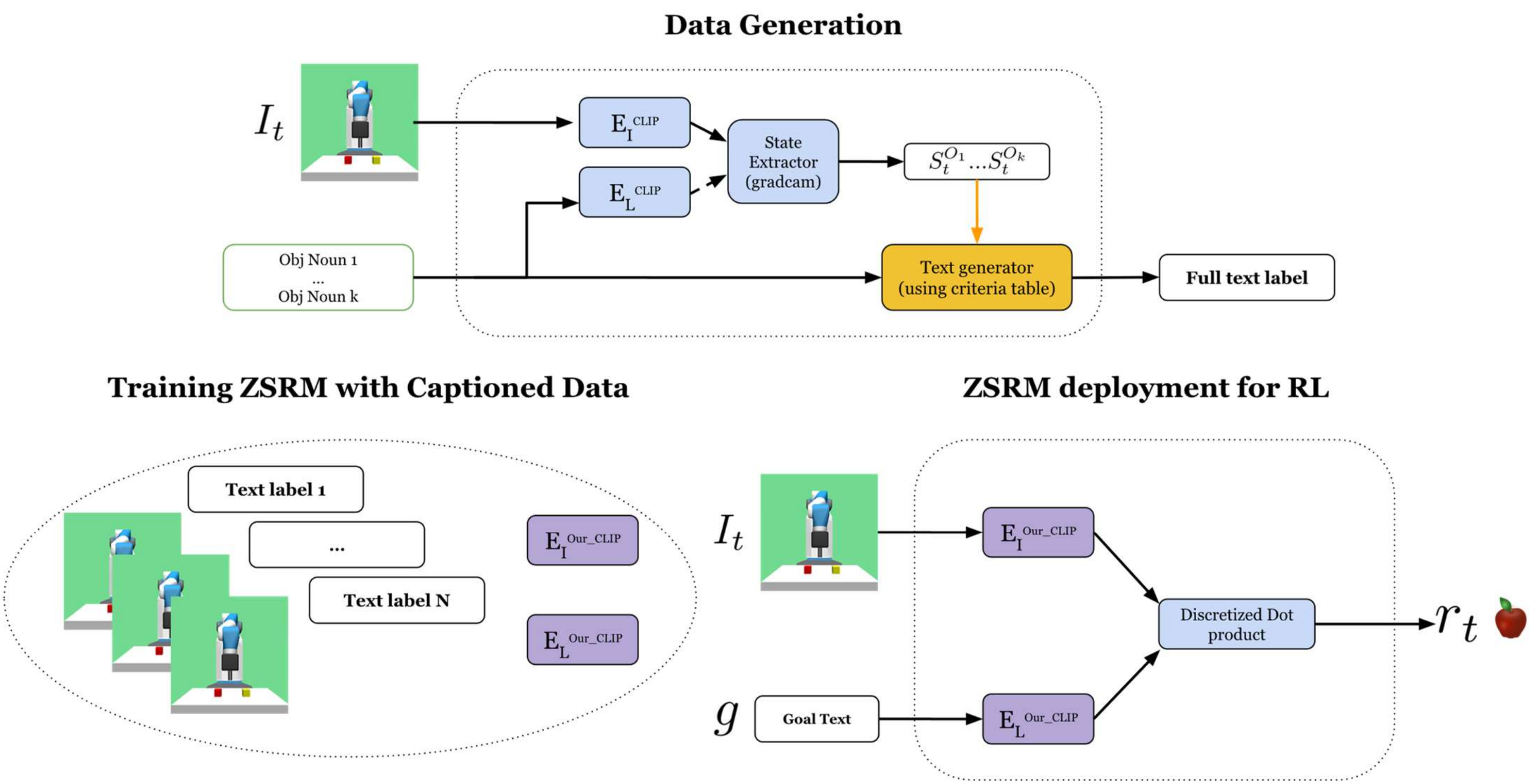


Method Overview

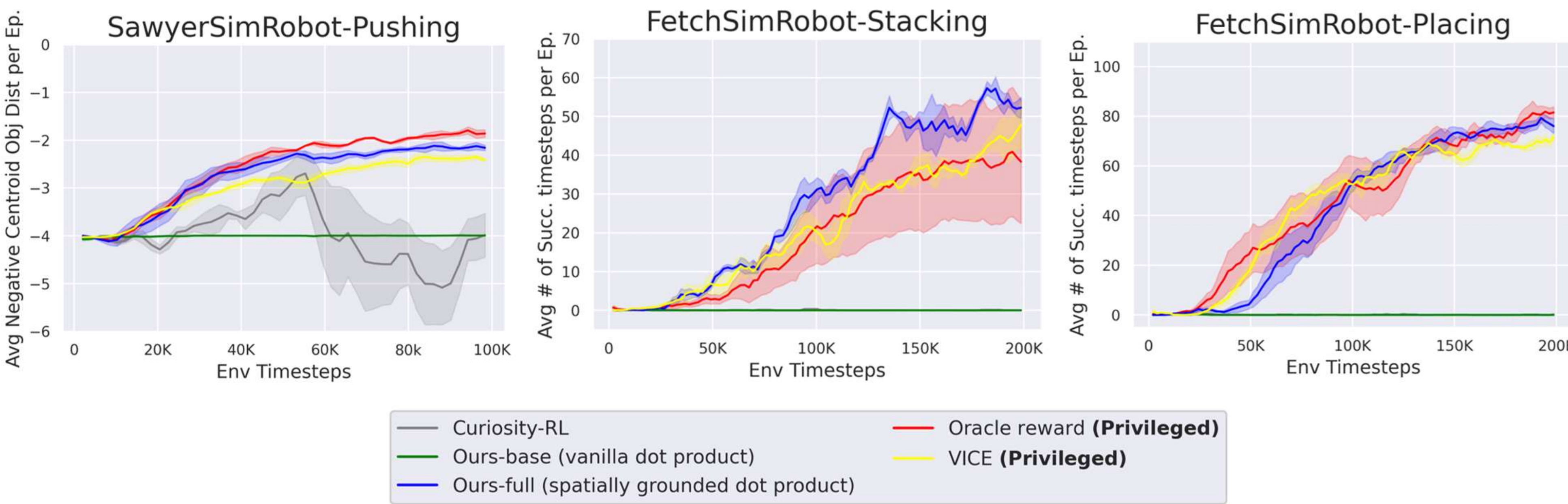




Method Overview

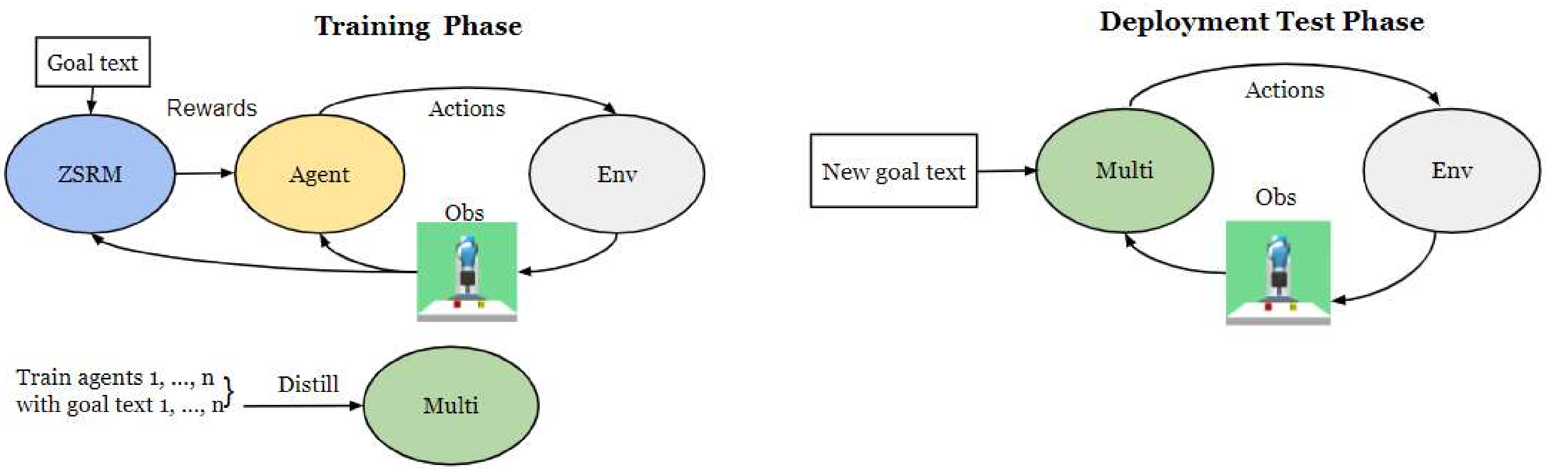


Main Results



Same performance as Oracle Reward

Multi-task Policy



	Seen distribution				Unseen distribution			
	train tasks		test tasks		train tasks		test tasks	
(episode reward stats)	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
No Conditioning	17.91	1.11	14.82	0.97	14.81	1.02	10.79	0.85
Primitive Code Cond.	26.71	1.23	17.20	1.03	17.03	1.07	11.74	0.87
Language Cond.	29.89	1.28	22.41	1.09	21.14	1.16	15.69	0.98

Future Work

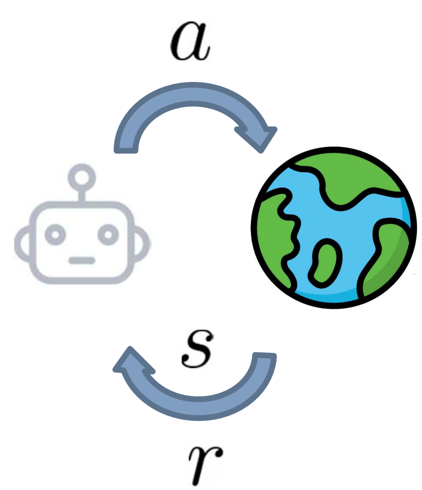
What's missing?

- pose tasks, semantic tasks like closed door, …

Future directions:

1. Leverage Simulators

2. Improve image & text alignment of LLVM



Thanks!