CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation

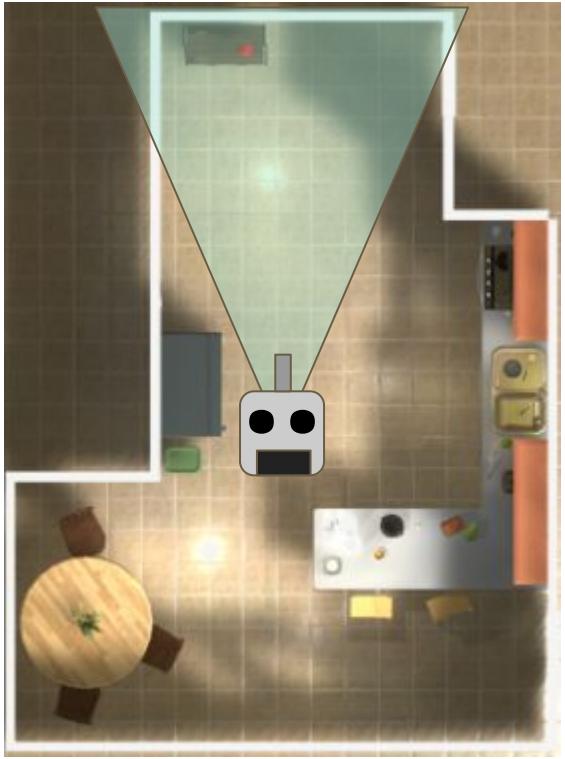
Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, Shuran Song

Presented by: Chase Vickery

Overview

- Problem/Background/Motivation
  - Embodied AI Intro
  - Embodied CLIP
- Methods
- Experimental Setup
- Results
- Strengths / Weaknesses
- Discussion

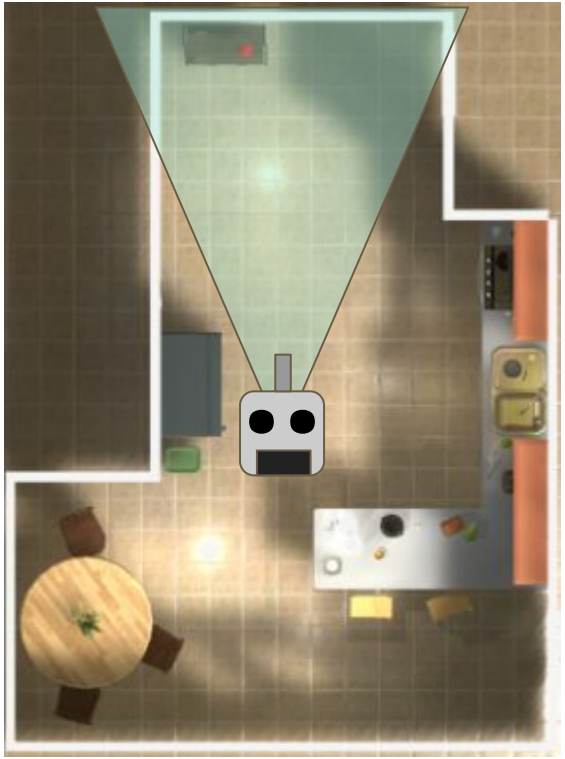
Background (Embodied AI)



Explore?
Go Somewhere?
Find Something?
Move Something?
…

https://ai2thor.allenai.org/demo

Background (Embodied AI)



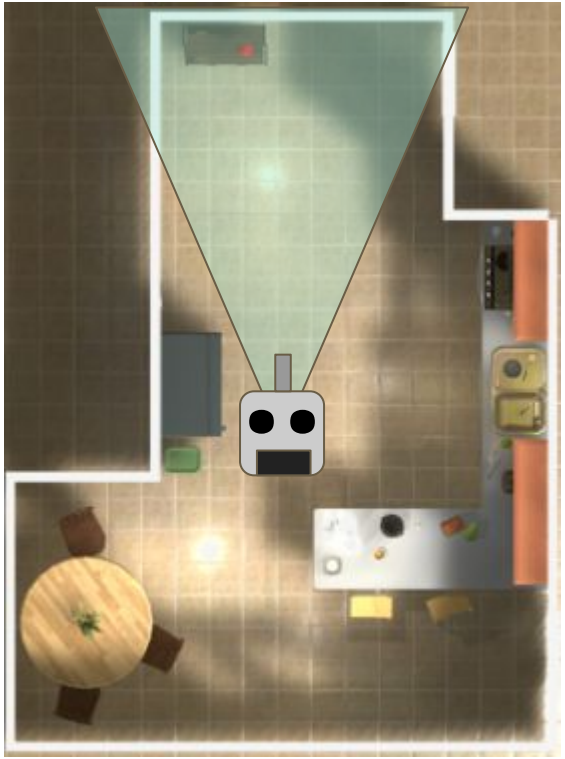
Helpful (Potentially)

- Navigate Unseen/Dangerous Areas
- Find Something/Someone
- Minimize Risk to Humans

For now we have simulations.

https://ai2thor.allenai.org/demo

Background (Embodied AI)



Needs Training!
(Millions of Steps)

https://ai2thor.allenai.org/demo

Simple but Effective: CLIP Embeddings for Embodied AI

Environments

- RoboTHOR
  - Modular, Asset-based
- Habitat
  - Matterport 3D Dataset

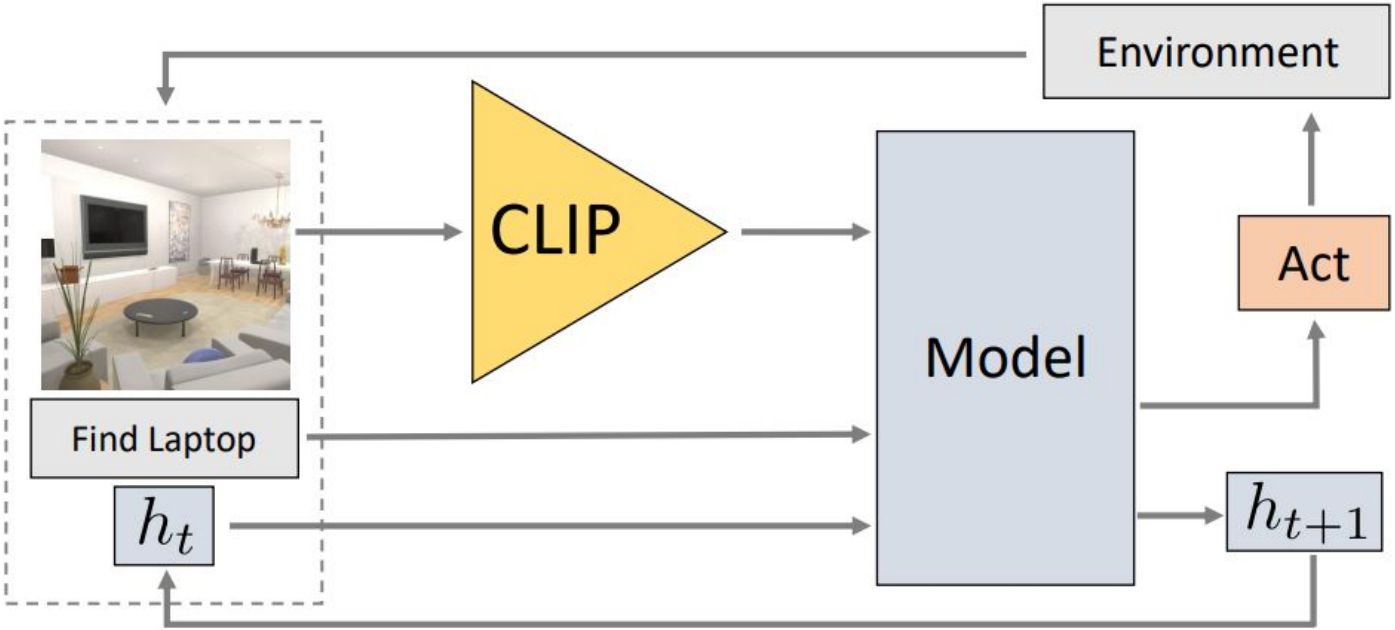
https://ai2thor.allenai.org/robothor/ https://aihabitat.org/

Tasks

- RoboTHOR
  - ObjectNav
  - Room Rearrangement
- Habitat
  - ObjectNav
  - PointNav

https://ai2thor.allenai.org/robothor/ https://aihabitat.org/

Methods



Methods

ResNets Pretrained on ImageNet

ResNets Pretrained w/CLIP

4 Tasks

Evaluations

Results

Models	SPL	SR	SPL Prox	SR Prox
ResNet-50 (CLIP)	**0.20**	**0.47**	**0.20**	**0.48**
ResNet-50 (ImageNet)	0.15	0.34	0.15	0.35
(1) EmbCLIP (Ours)	**0.20**	**0.47**	**0.20**	**0.48**
(2) Action Boost	0.12	0.28	0.12	0.30
(3) RGB+D ResNet18	0.11	0.26	0.12	0.28
		...		

Results

Model	FS	SR	E	M
ResNet-50 (CLIP)	**0.17**	**0.08**	**0.89**	0.88
ResNet-50 (ImageNet)	0.07	0.03	1.06	1.05
(1) EmbCLIP (Ours)	**0.17**	**0.08**	**0.89**	0.88
(2) RN18 + ANM IL [31]	0.09	0.03	1.04	1.05
(3) RN18 + IL [31]	0.06	0.03	1.09	1.11
...				

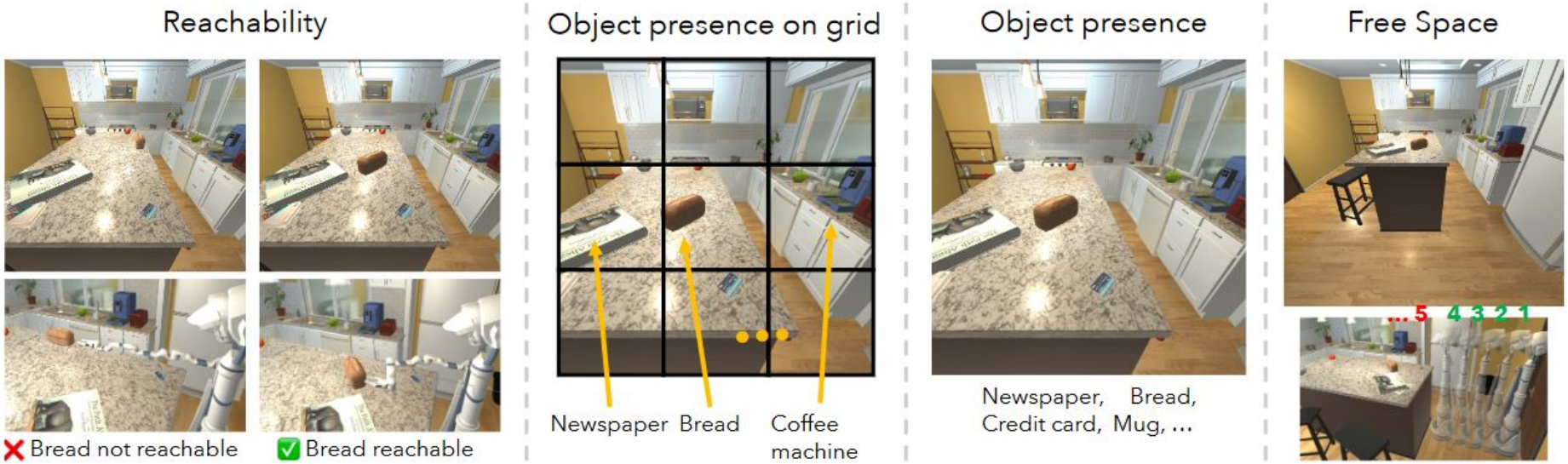
Results

Models	SPL	SR	SoftSPL	Goal Dist
ResNet-50 (CLIP)	**0.08**	**0.18**	**0.20**	**7.92**
ResNet-50 (ImageNet)	0.05	0.13	0.17	8.69
(1) yuumi_the_magic_cat [18]	**0.10**	0.22	0.18	9.17
(2) TreasureHunt [19]	0.09	0.21	0.17	9.20
(3) Habitat on Web (IL-HD) [24]	0.08	**0.24**	0.16	**7.88**
(4) EmbCLIP (Ours)	0.08	0.18	**0.20**	7.92
(-) Habitat on Web^{2021} [24]	0.07	0.21	0.15	8.26
(5) Red Rabbit^{2021} [37]	0.06	0.24	0.12	9.15
...				
(9) DD-PPO	0.00	0.00	0.01	10.326

Results

Models	SPL	SR	Goal Dist
ResNet-50 (CLIP)	**0.87**	**0.97**	**0.40**
ResNet-50 (ImageNet)	0.82	0.94	0.73

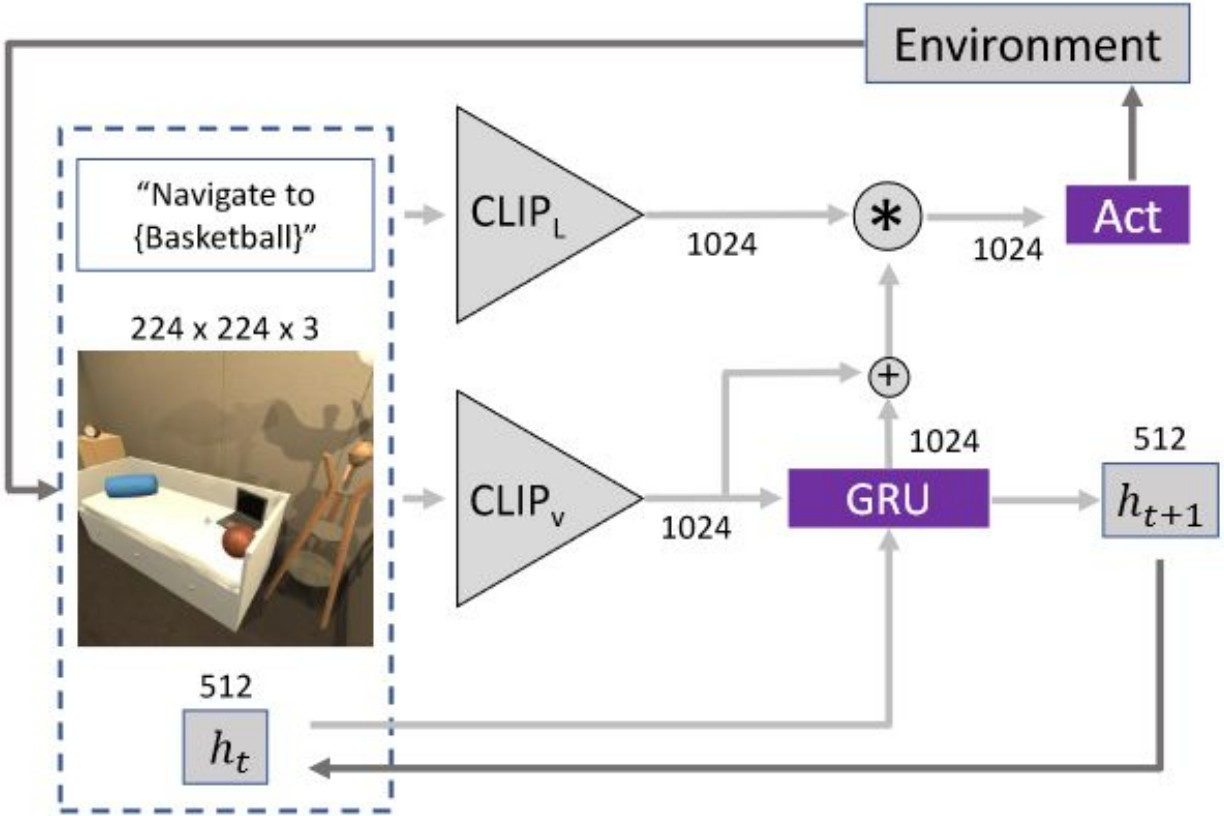
Move Evaluations



Evaluation Results

Task	Pretraining	Pooling	Score
Object Presence	ImageNet	Average	0.502
	CLIP	**Average**	**0.530**
	CLIP	Attention	0.529
Object Localization	ImageNet	Average	0.387
	CLIP	**Average**	**0.452**
Reachability	ImageNet	Average	0.638
	CLIP	**Average**	**0.677**
	CLIP	Attention	0.668
Free Space	ImageNet	Average	0.287
	CLIP	**Average**	**0.315**
	CLIP	Attention	0.257

ZSON

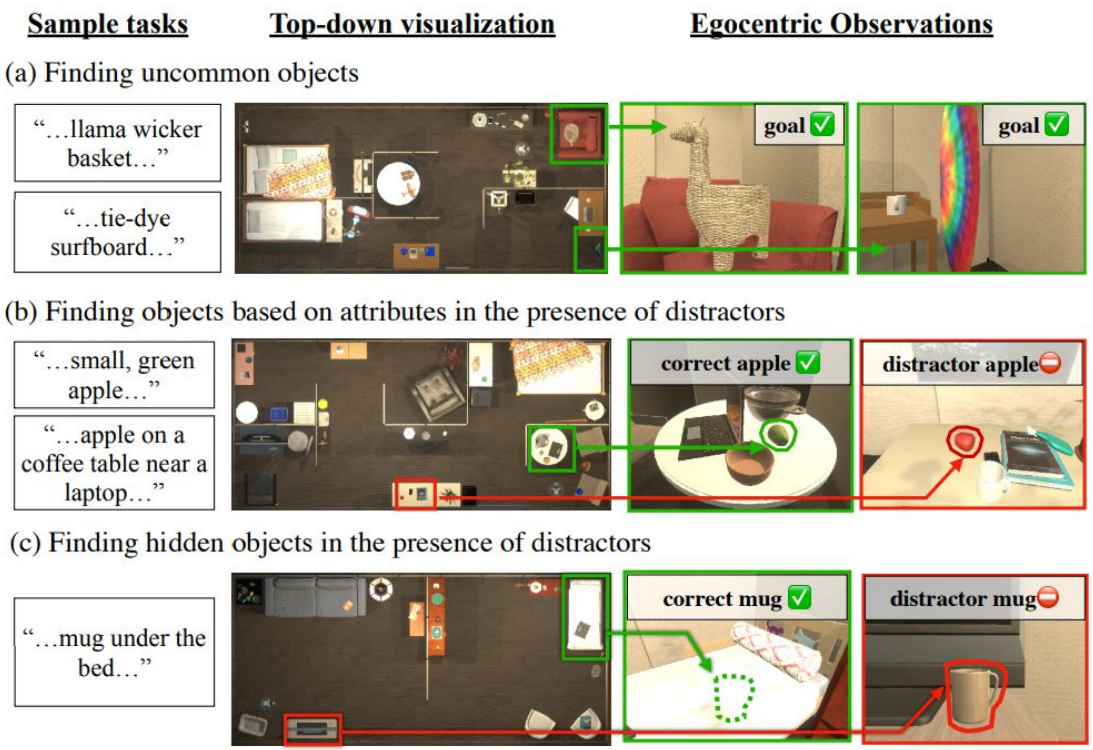

Environment
"Navigate to {Basketball}"
224 x 224 x 3
CLIP$_L$
CLIP$_v$
1024
1024
1024
1024
512
512
GRU
Act
h_t
h_{t+1}

ZSON (Results)

	Seen Objects	Unseen Objects				
Method	All	All	Apple	Basketball	House Plant	Television
Random	0.016	0.02	0.013	0.007	0.047	0.013
Ours	**0.170**	**0.081**	**0.147**	**0.067**	**0.053**	**0.060**

CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation

Overview (L-ZSON)



Methods

Baselines (CoWs)

Depth-Based Mapping

Open-Vocab Models

RGB-D + Goal Input

Explore vs Exploit

Benchmarks (Pasture)

7 L-ZSON Tasks

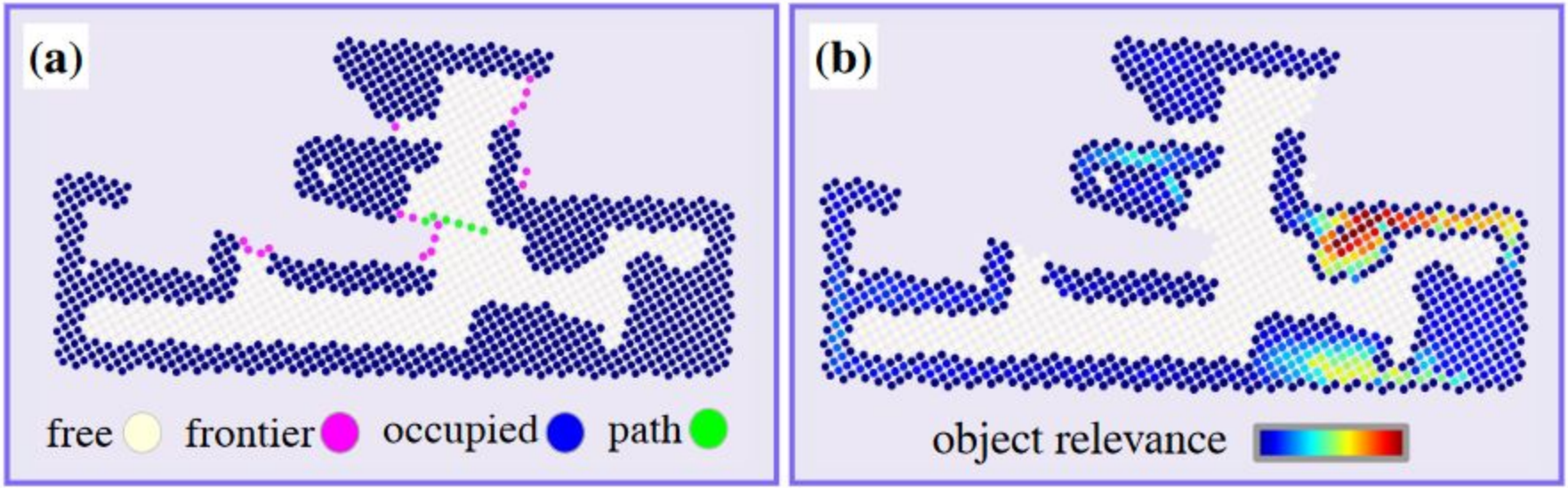
Uncommon Objects

Appearance & Spatial Descriptions

Hidden Objects

Distractors

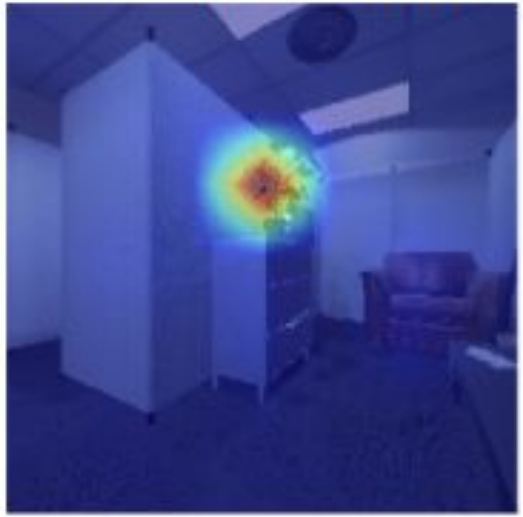
Methods (Mapping)



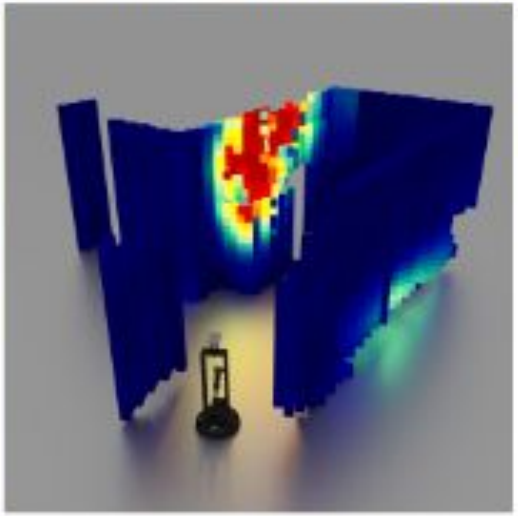
Methods (Mapping)

(a) Egocentric RGB

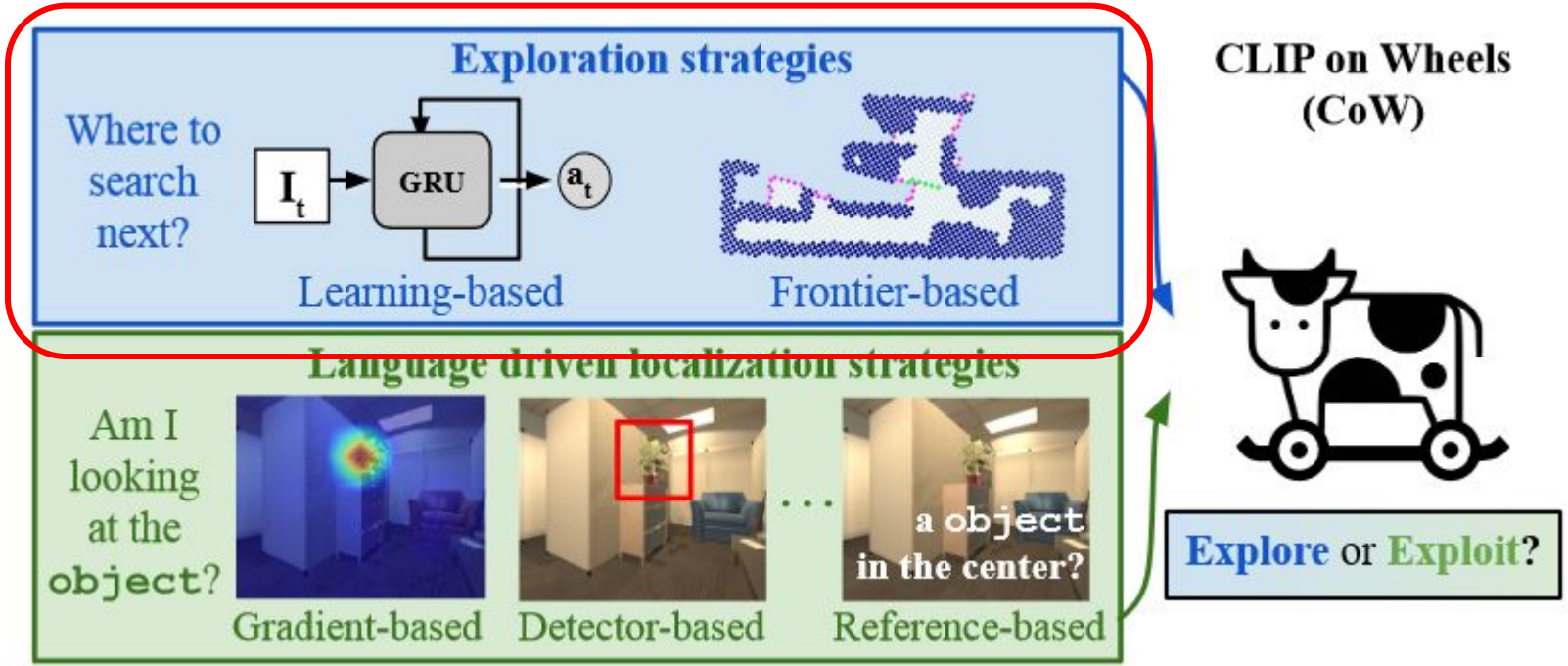
(b) CLIP-Grad. B/32 relevance overlay



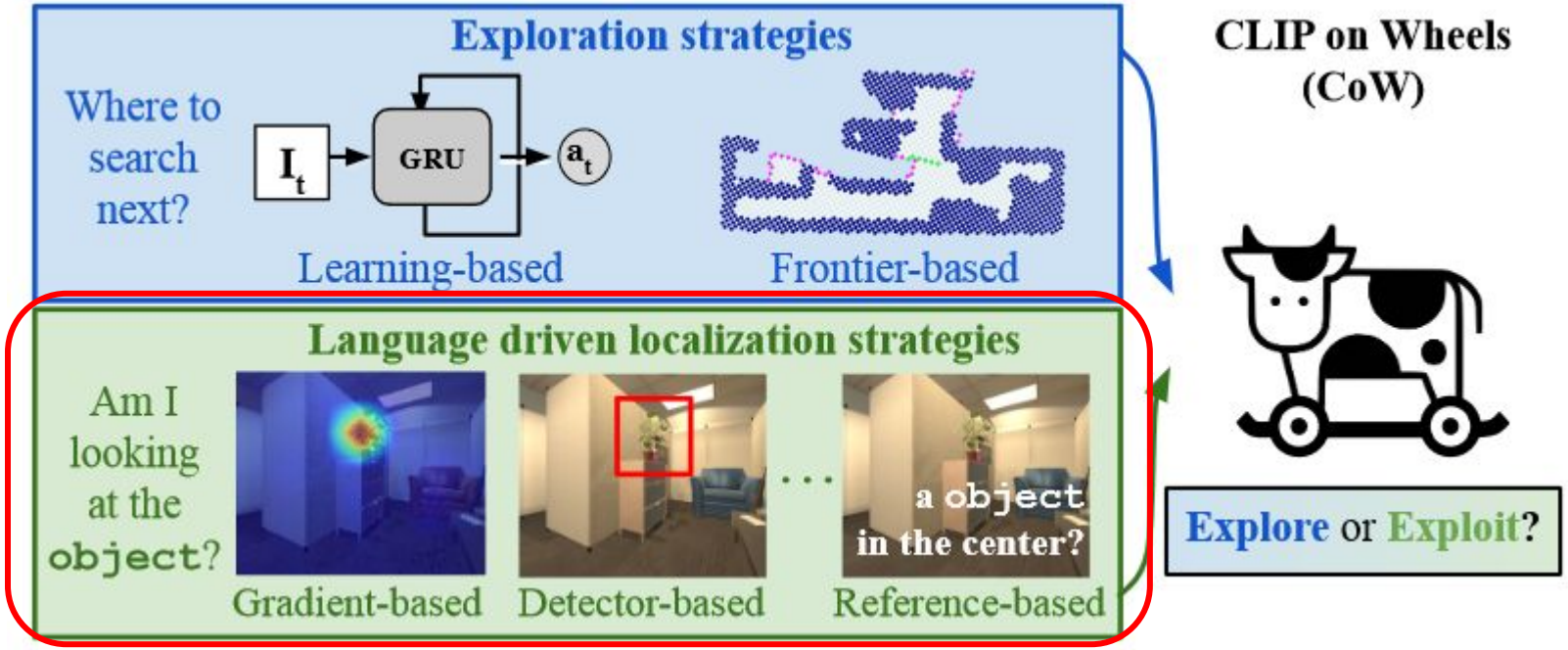
(c) Depth projected relevance (aggregated)



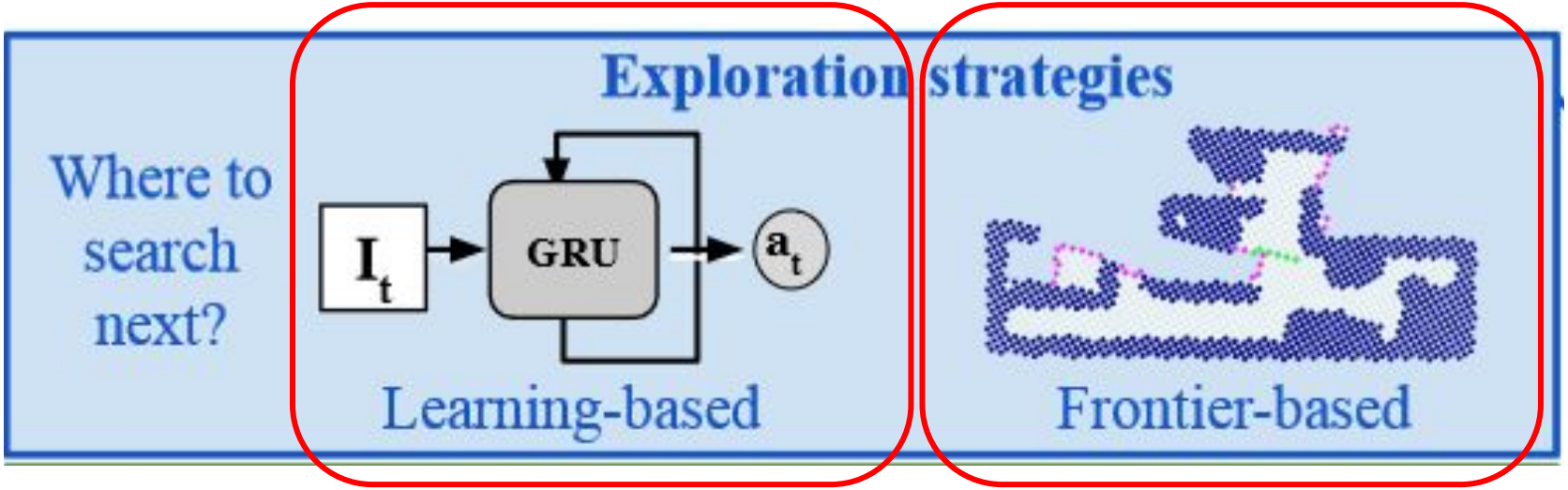
Methods (Baselines)



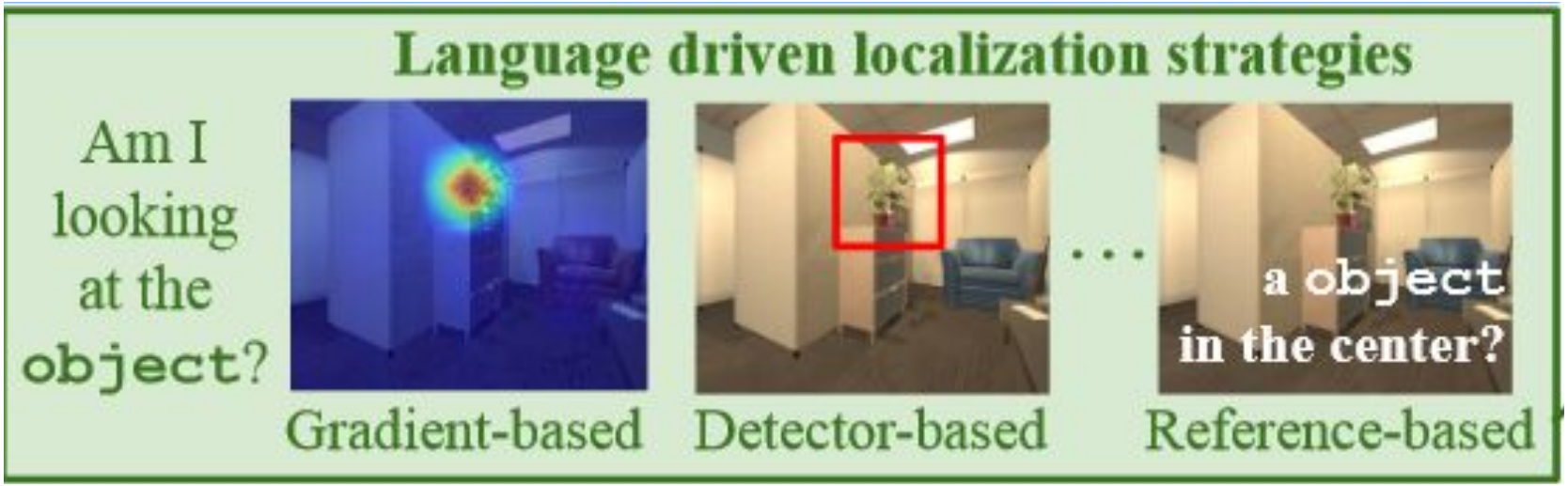
Methods (Baselines)



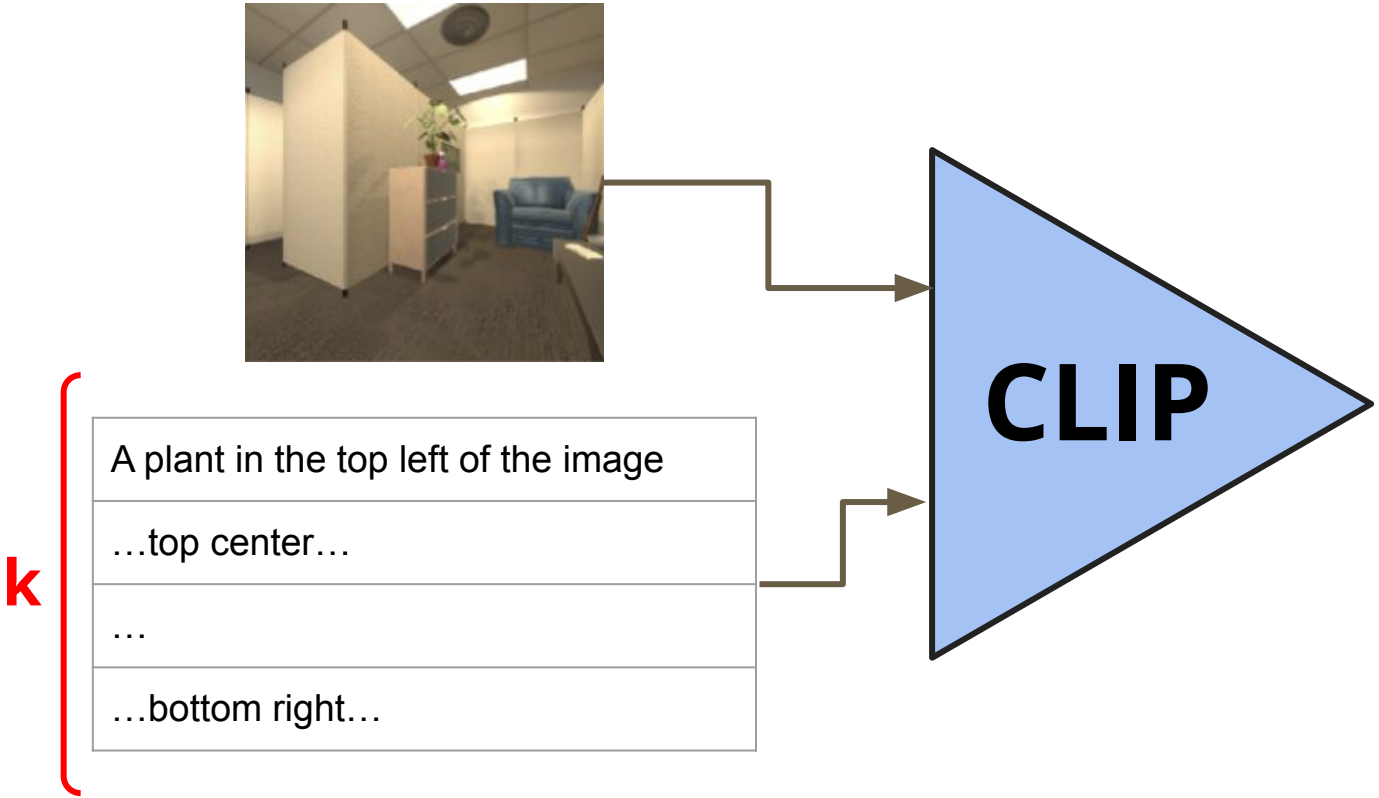
Methods (Baselines)



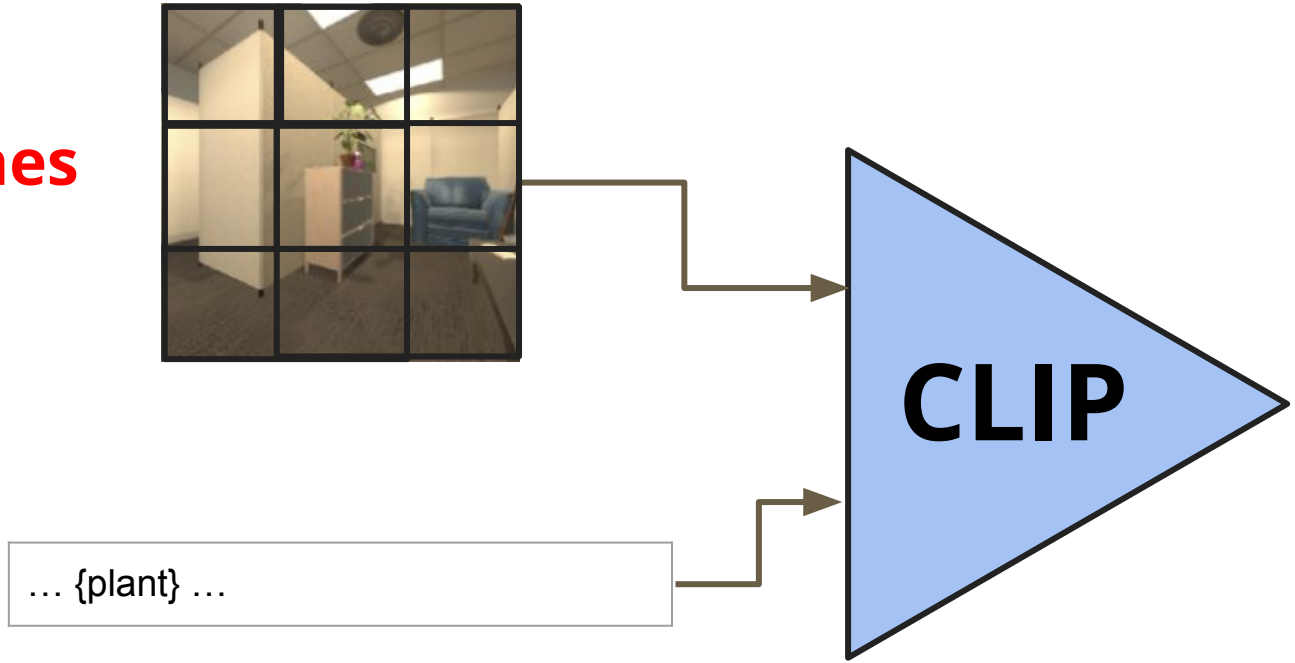
Methods (Baselines)



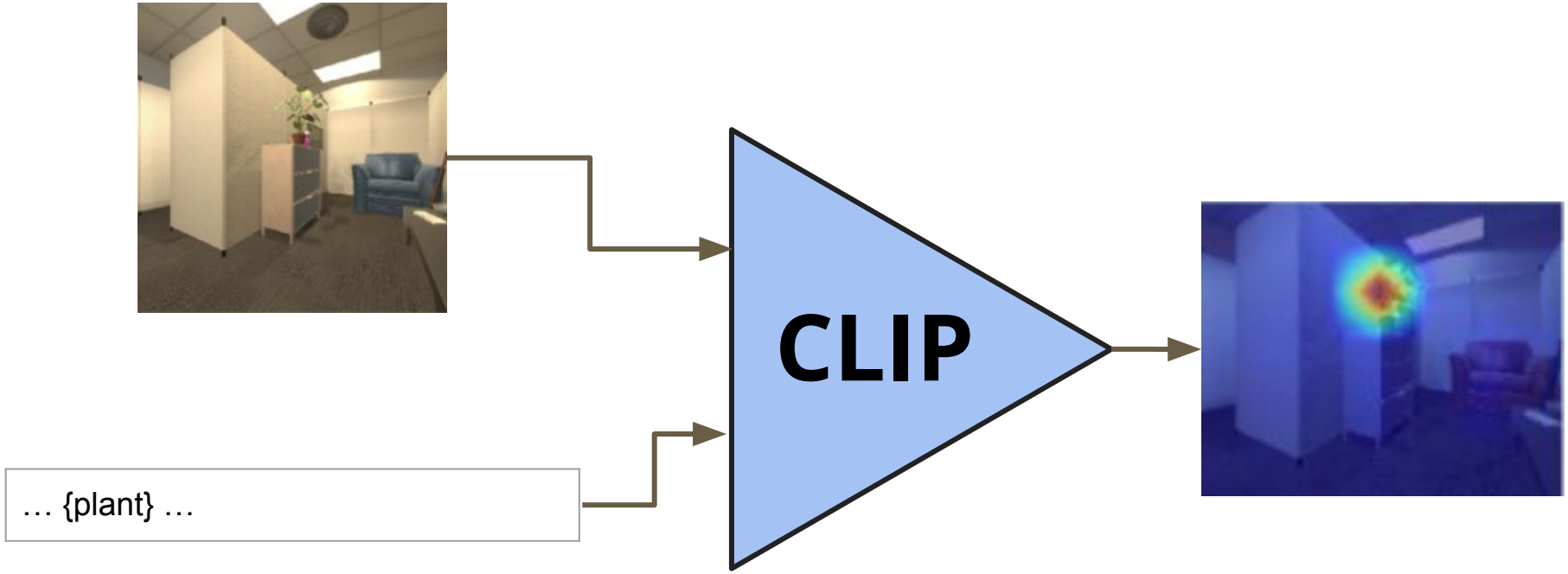
Methods (CLIP-Ref)



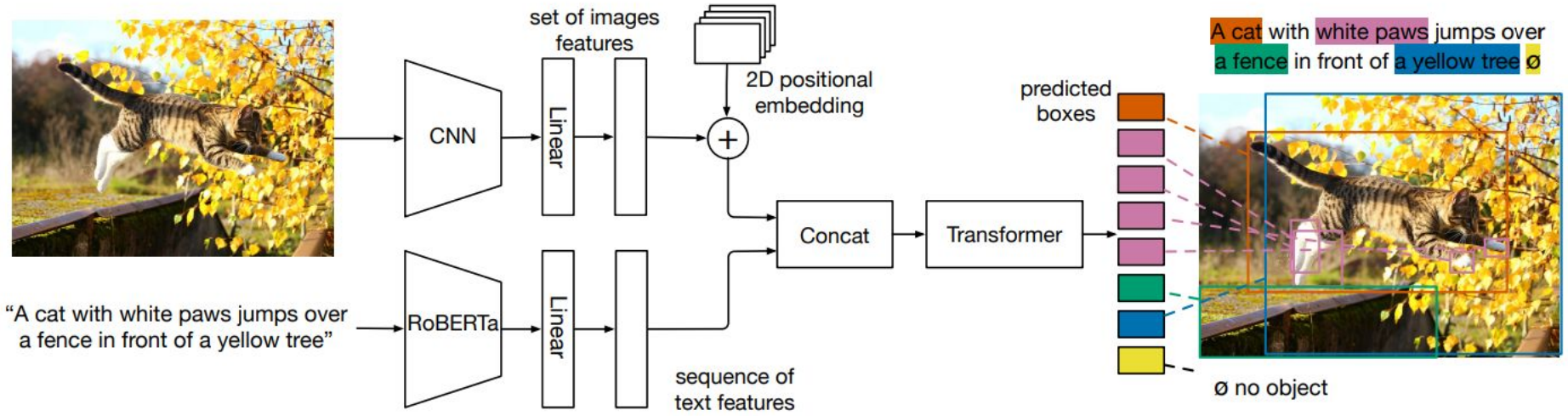
Methods (CLIP-Patch)



Methods (CLIP-Grad)

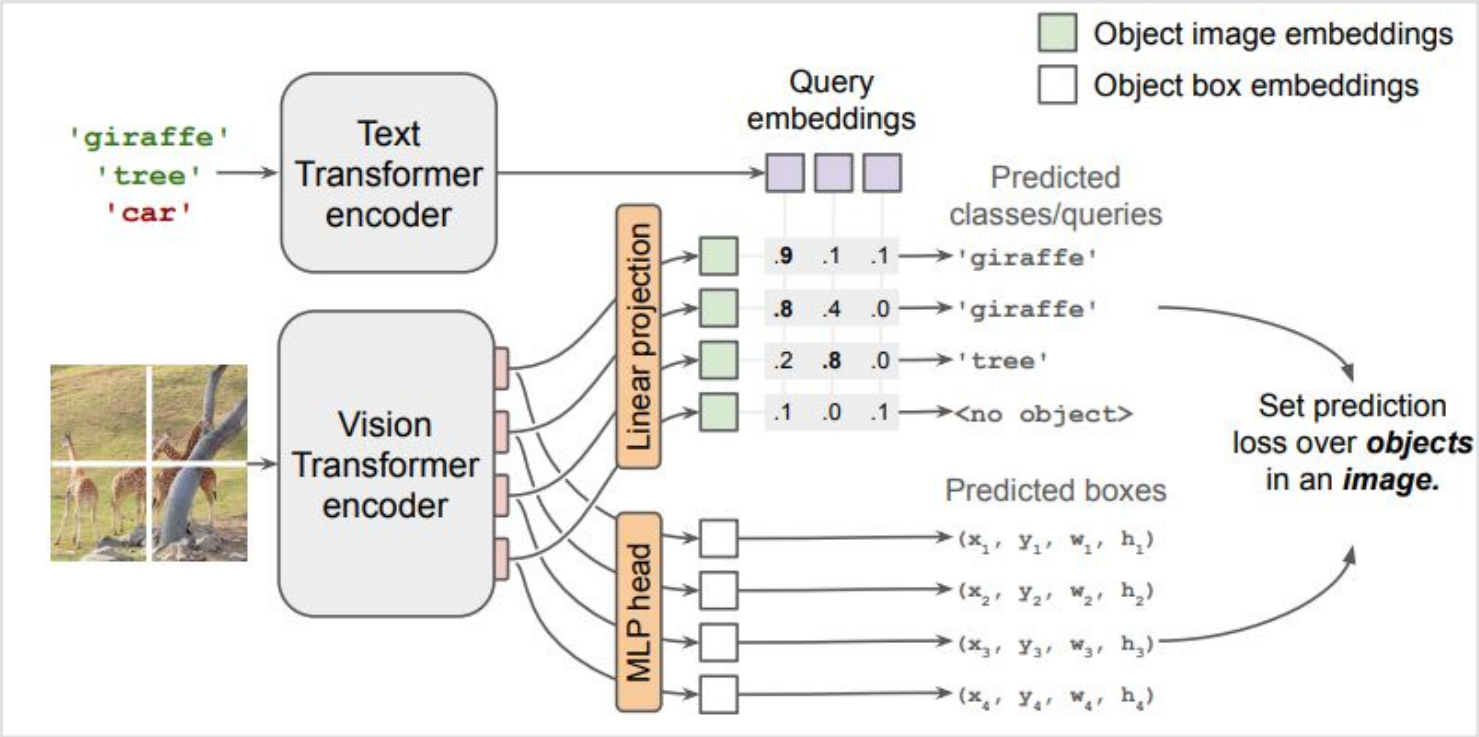


Methods (MDETR)



https://arxiv.org/pdf/2104.12763.pdf

Methods (OWL-ViT)

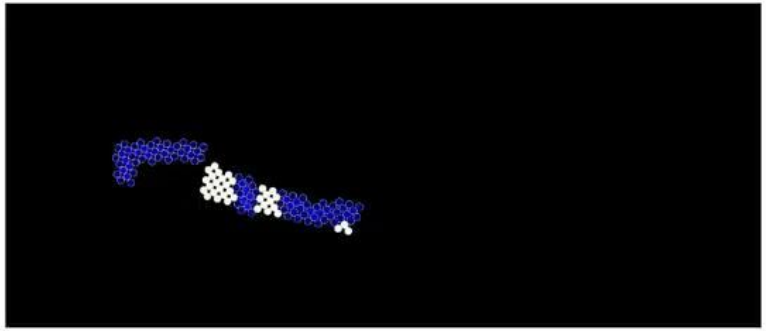


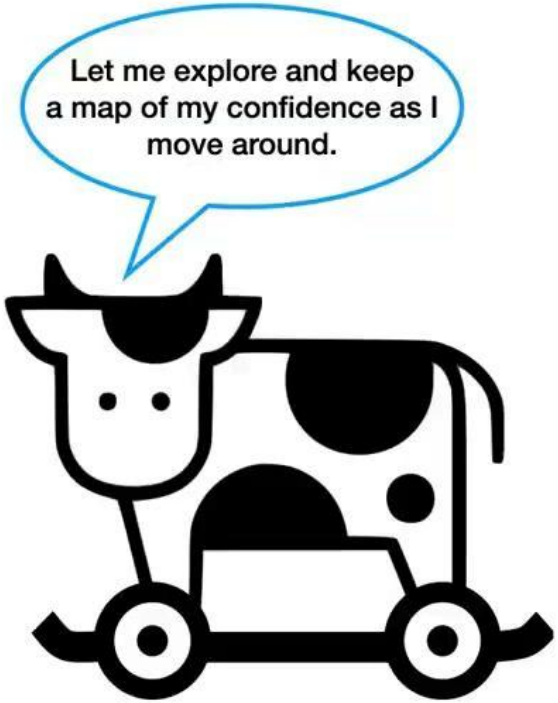
https://arxiv.org/pdf/2205.06230.pdf

Egocentric view

CLIP-based object relevance

Voxel projected object relevance map





Target: Plant!

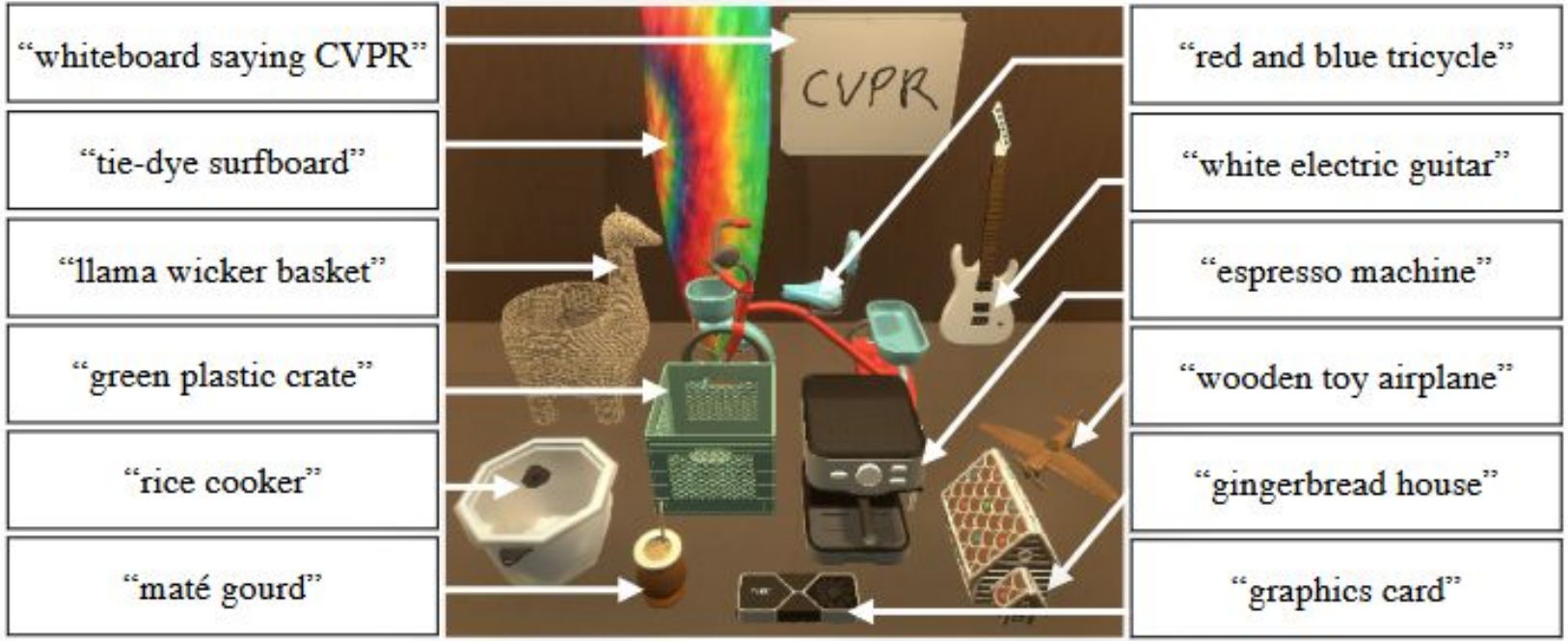
https://cow.cs.columbia.edu/

Methods (Benchmarks)

7 Tasks:

- Uncommon Objects
- Appearance Descriptions
- Appearance Descriptions w/Distractors
- Spatial Descriptions
- Spatial Descriptions w/Distractors
- Hidden Object Descriptions
- Hidden Object Descriptions w/Distractors

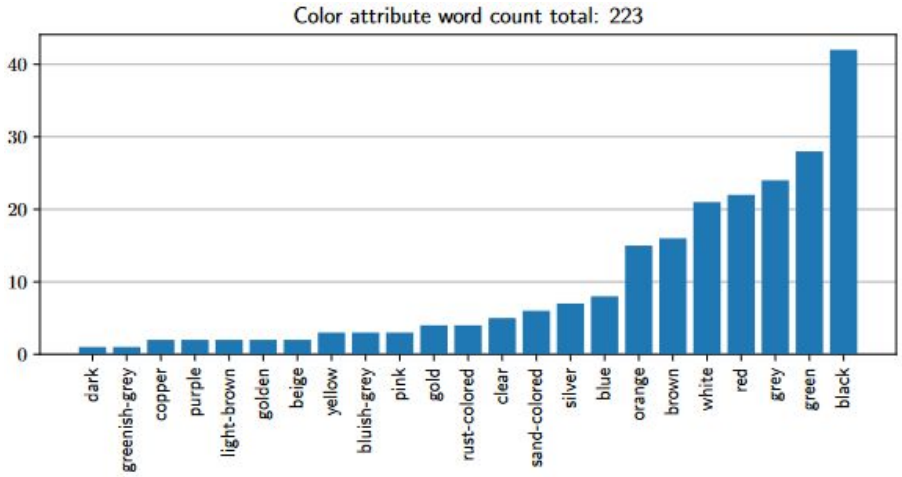
Methods (Uncommon Objects)

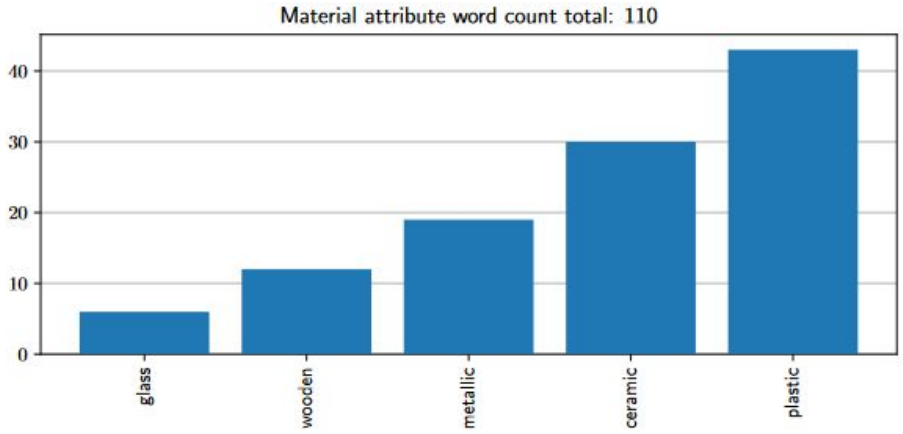


Methods (Appearance)

- “{Size}, {Color}, {Material}, {Object}”
- Ex.
  - “Small, red apple”
  - “Orange basketball”

Methods (Appearance)

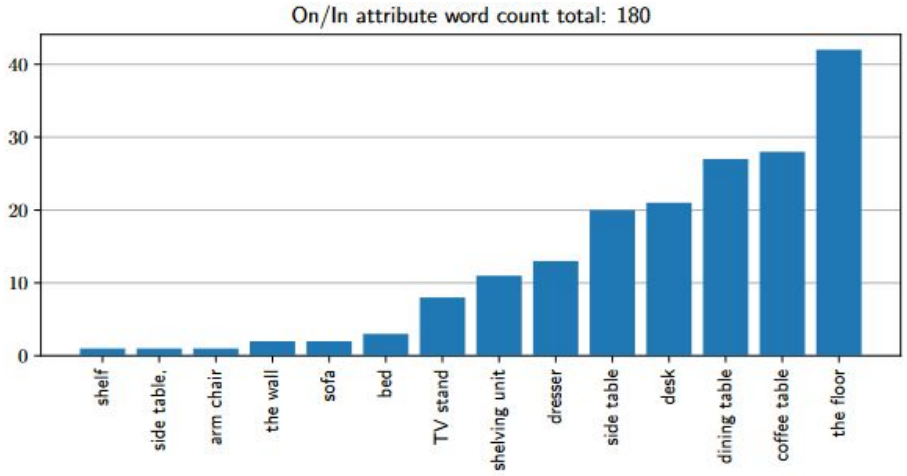


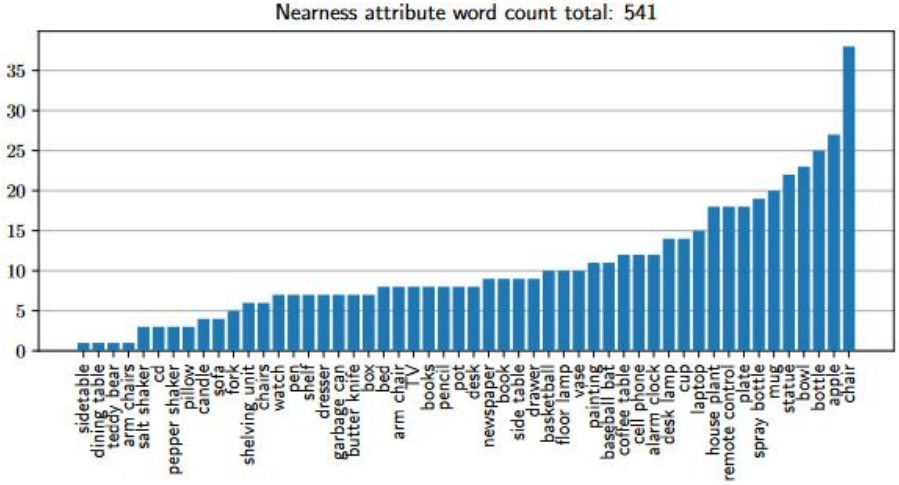


Methods (Spatial)

- “{Object} on top of {x}”
- “Near {y}, {z}”
- Ex.
  - “House plant on a dresser near a spray bottle”

Methods (Spatial)





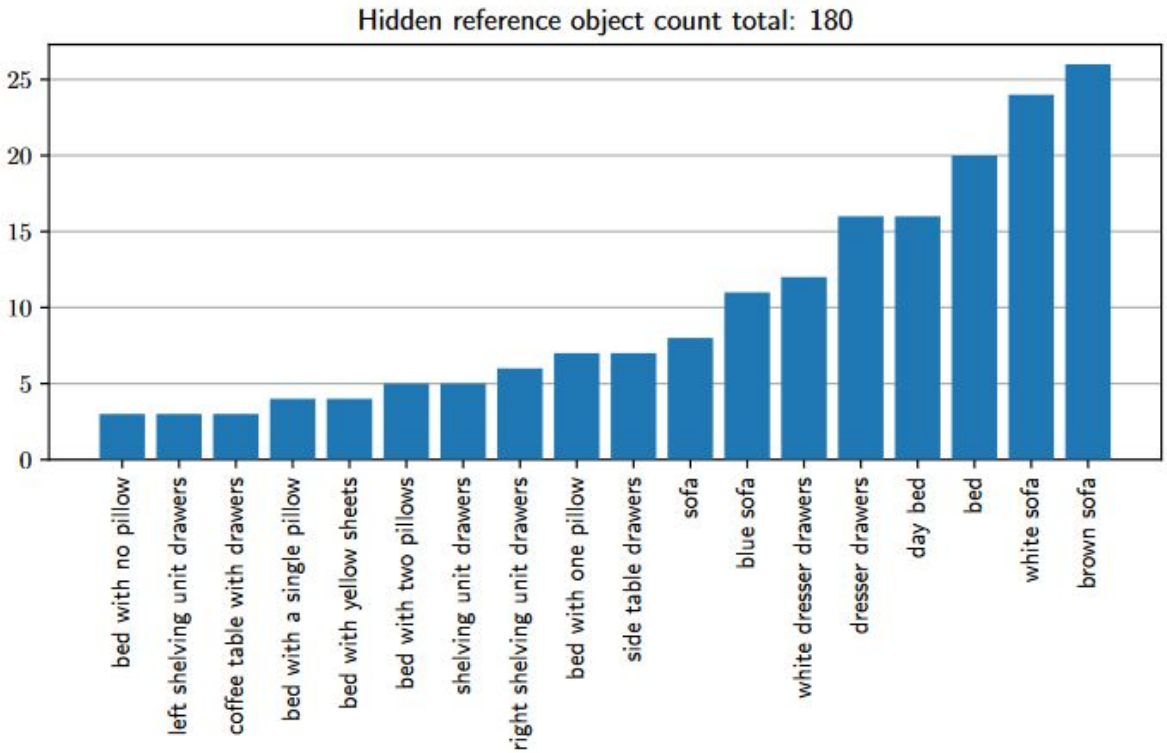
Methods (Distractors)

- Modified environment
- 2 distinct instances of each category
- Ex.
  - Both a red apple and green apple in environment for “red apple” target.

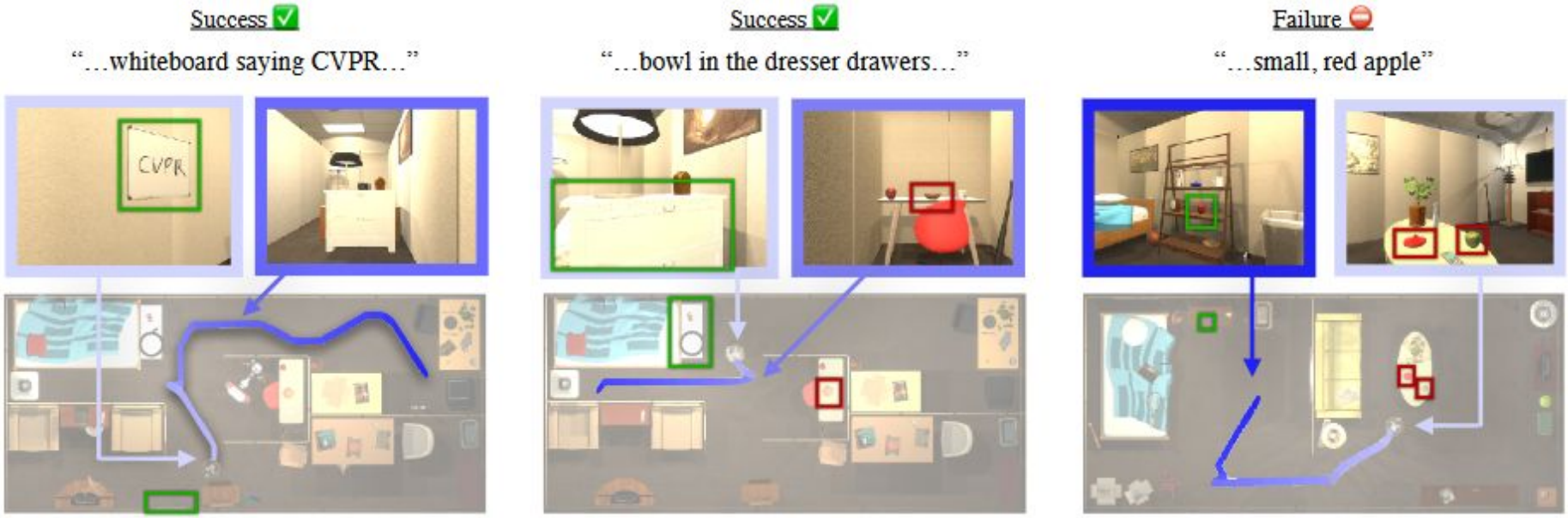
Methods (Hidden)

- “{Object} under/in {x}”
- Ex.
  - “Basketball in the dresser drawers”
  - “Vase under the sofa”
- Visible instances of {object} removed

Methods (Hidden)



Methods (Examples)



Experimental Setup

- Exploration:
  - Frontier-Based
- Object Localization
  - CLIP-Ref: k=9
  - CLIP-Patch: k=9
  - CLIP-Grad
  - MDETR
  - OWL-ViT
- ViT-B/32 vs. ViT-B/16
- Post-processing

Results

CoW breeds				PASTURE							Avg.		ROBOTHOR	
				Uncom.	Appear.	Space	Appear. distract	Space distract	Hid.	Hid. distract				
ID	Localizer	Arch.	Post	SR	SR	SR	SR	SR	SR	SR	SPL	SR	SPL	SR
▲	CLIP-Ref.	B/32		2.8	1.4	1.4	0.8	1.4	4.7	5.0	1.2	2.5	1.6	2.2
△	CLIP-Ref.	B/32	✓	3.6	0.6	1.7	0.6	1.7	2.2	2.5	0.9	1.8	1.0	1.8
■	CLIP-Ref.	B/16		1.4	1.7	1.7	1.9	1.9	2.8	2.2	1.7	1.9	2.4	2.6
□	CLIP-Ref.	B/16	✓	1.4	2.8	2.8	3.1	3.3	1.7	1.9	1.7	2.4	2.1	2.7
▲	CLIP-Patch	B/32		10.6	9.7	6.7	6.4	6.4	16.7	16.7	7.5	10.4	9.0	14.3
△	CLIP-Patch	B/32	✓	18.1	13.3	13.3	8.6	10.8	17.5	**17.8**	9.0	14.2	10.6	20.3
■	CLIP-Patch	B/16		5.6	7.8	3.9	5.0	3.9	10.6	10.8	5.4	6.8	8.2	10.3
□	CLIP-Patch	B/16	✓	10.6	11.4	7.8	10.8	8.1	16.4	15.6	7.7	11.5	9.7	15.7
▲	CLIP-Grad.	B/32		13.6	10.6	9.2	7.5	7.2	13.9	12.8	8.3	10.7	9.6	13.8
△	CLIP-Grad.	B/32	✓	16.1	11.9	11.7	9.7	10.3	14.4	16.1	9.2	12.9	9.7	15.2
■	CLIP-Grad.	B/16		6.1	5.8	5.0	5.0	4.7	8.3	6.9	4.9	6.0	7.3	8.8
□	CLIP-Grad.	B/16	✓	8.1	10.8	8.6	8.6	6.7	11.1	11.4	6.7	9.3	8.6	11.6
◆	MDETR	B3		3.1	6.9	4.4	7.2	4.7	7.8	8.9	5.3	6.2	8.3	9.8
◇	MDETR	B3	✓	3.1	7.2	5.0	6.9	4.7	8.1	8.9	5.4	6.3	8.4	9.9
▲	OWL	B/32		23.1	26.1	14.4	18.3	11.7	13.9	13.1	11.1	17.2	16.6	25.4
△	OWL	B/32	✓	**32.8**	26.4	**19.4**	**19.4**	**16.1**	**19.2**	14.4	**12.6**	**21.1**	16.9	26.7
■	OWL	B/16		25.8	23.6	15.3	17.2	12.5	13.1	13.9	11.4	17.3	16.2	24.8
□	OWL	B/16	✓	31.9	**26.9**	18.9	**19.4**	14.7	18.1	15.8	**12.6**	20.8	**17.2**	**27.5**
ProcTHOR fine-tune (supervised) [18]				n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	27.4	66.4

Results

				PASTURE									ROBOTHOR	
	CoW breeds			Uncom.	Appear.	Space	Appear. distract	Space distract	Hid.	Hid. distract	Avg.			
ID	Localizer	Arch.	Post	SR	SR	SR	SR	SR	SR	SR	SPL	SR	SPL	SR
▲	CLIP-Ref.	B/32		2.8	1.4	1.4	0.8	1.4	4.7	5.0	1.2	2.5	1.6	2.2
△	CLIP-Ref.	B/32	✓	3.6	0.6	1.7	0.6	1.7	2.2	2.5	0.9	1.8	1.0	1.8
■	CLIP-Ref.	B/16		1.4	1.7	1.7	1.9	1.9	2.8	2.2	1.7	1.9	2.4	2.6
□	CLIP-Ref.	B/16	✓	1.4	2.8	2.8	3.1	3.3	1.7	1.9	1.7	2.4	2.1	2.7
▲	CLIP-Patch	B/32		10.6	9.7	6.7	6.4	6.4	16.7	16.7	7.5	10.4	9.0	14.3
△	CLIP-Patch	B/32	✓	18.1	13.3	13.3	8.6	10.8	17.5	**17.8**	9.0	14.2	10.6	20.3
■	CLIP-Patch	B/16		5.6	7.8	3.9	5.0	3.9	10.6	10.8	5.4	6.8	8.2	10.3
□	CLIP-Patch	B/16	✓	10.6	11.4	7.8	10.8	8.1	16.4	15.6	7.7	11.5	9.7	15.7
▲	CLIP-Grad.	B/32		13.6	10.6	9.2	7.5	7.2	13.9	12.8	8.3	10.7	9.6	13.8
△	CLIP-Grad.	B/32	✓	16.1	11.9	11.7	9.7	10.3	14.4	16.1	9.2	12.9	9.7	15.2
■	CLIP-Grad.	B/16		6.1	5.8	5.0	5.0	4.7	8.3	6.9	4.9	6.0	7.3	8.8
□	CLIP-Grad.	B/16	✓	8.1	10.8	8.6	8.6	6.7	11.1	11.4	6.7	9.3	8.6	11.6
◆	MDETR	B3		3.1	6.9	4.4	7.2	4.7	7.8	8.9	5.3	6.2	8.3	9.8
◇	MDETR	B3	✓	3.1	7.2	5.0	6.9	4.7	8.1	8.9	5.4	6.3	8.4	9.9
▲	OWL	B/32		23.1	26.1	14.4	18.3	11.7	13.9	13.1	11.1	17.2	16.6	25.4
△	OWL	B/32	✓	**32.8**	26.4	**19.4**	**19.4**	**16.1**	**19.2**	14.4	**12.6**	**21.1**	16.9	26.7
■	OWL	B/16		25.8	23.6	15.3	17.2	12.5	13.1	13.9	11.4	17.3	16.2	24.8
□	OWL	B/16	✓	31.9	**26.9**	18.9	**19.4**	14.7	18.1	15.8	**12.6**	20.8	**17.2**	**27.5**
ProcTHOR fine-tune (supervised) [18]				n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	27.4	66.4

Results

	CoW breeds			PASTURE									ROBOTHOR	
				Uncom.	Appear.	Space	Appear. distract	Space distract	Hid.	Hid. distract	Avg.			
ID	Localizer	Arch.	Post	SR	SR	SR	SR	SR	SR	SR	SPL	SR	SPL	SR
▲	CLIP-Ref.	B/32		2.8	1.4	1.4	0.8	1.4	4.7	5.0	1.2	2.5	1.6	2.2
△	CLIP-Ref.	B/32	✓	3.6	0.6	1.7	0.6	1.7	2.2	2.5	0.9	1.8	1.0	1.8
■	CLIP-Ref.	B/16		1.4	1.7	1.7	1.9	1.9	2.8	2.2	1.7	1.9	2.4	2.6
□	CLIP-Ref.	B/16	✓	1.4	2.8	2.8	3.1	3.3	1.7	1.9	1.7	2.4	2.1	2.7
▲	CLIP-Patch	B/32		10.6	9.7	6.7	6.4	6.4	16.7	16.7	7.5	10.4	9.0	14.3
△	CLIP-Patch	B/32	✓	18.1	13.3	13.3	8.6	10.8	17.5	**17.8**	9.0	14.2	10.6	20.3
■	CLIP-Patch	B/16		5.6	7.8	3.9	5.0	3.9	10.6	10.8	5.4	6.8	8.2	10.3
□	CLIP-Patch	B/16	✓	10.6	11.4	7.8	10.8	8.1	16.4	15.6	7.7	11.5	9.7	15.7
▲	CLIP-Grad.	B/32		13.6	10.6	9.2	7.5	7.2	13.9	12.8	8.3	10.7	9.6	13.8
△	CLIP-Grad.	B/32	✓	16.1	11.9	11.7	9.7	10.3	14.4	16.1	9.2	12.9	9.7	15.2
■	CLIP-Grad.	B/16		6.1	5.8	5.0	5.0	4.7	8.3	6.9	4.9	6.0	7.3	8.8
□	CLIP-Grad.	B/16	✓	8.1	10.8	8.6	8.6	6.7	11.1	11.4	6.7	9.3	8.6	11.6
◆	MDETR	B3		3.1	6.9	4.4	7.2	4.7	7.8	8.9	5.3	6.2	8.3	9.8
◇	MDETR	B3	✓	3.1	7.2	5.0	6.9	4.7	8.1	8.9	5.4	6.3	8.4	9.9
▲	OWL	B/32		23.1	26.1	14.4	18.3	11.7	13.9	13.1	11.1	17.2	16.6	25.4
△	OWL	B/32	✓	**32.8**	26.4	**19.4**	**19.4**	**16.1**	**19.2**	14.4	**12.6**	**21.1**	16.9	26.7
■	OWL	B/16		25.8	23.6	15.3	17.2	12.5	13.1	13.9	11.4	17.3	16.2	24.8
□	OWL	B/16	✓	31.9	**26.9**	18.9	**19.4**	14.7	18.1	15.8	**12.6**	20.8	**17.2**	**27.5**
ProcTHOR fine-tune (supervised) [18]				n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	27.4	66.4

Results

ID	CoW breeds: Localizer	Arch.	Post	PASTURE: Uncom. SR	Appear. SR	Space SR	Appear. distract SR	Space distract SR	Hid. SR	Hid. distract SR	Avg. SPL	Avg. SR	ROBOTHOR SPL	ROBOTHOR SR
▲	CLIP-Ref.	B/32		2.8	1.4	1.4	0.8	1.4	4.7	5.0	1.2	2.5	1.6	2.2
△	CLIP-Ref.	B/32	✓	3.6	0.6	1.7	0.6	1.7	2.2	2.5	0.9	1.8	1.0	1.8
■	CLIP-Ref.	B/16		1.4	1.7	1.7	1.9	1.9	2.8	2.2	1.7	1.9	2.4	2.6
□	CLIP-Ref.	B/16	✓	1.4	2.8	2.8	3.1	3.3	1.7	1.9	1.7	2.4	2.1	2.7
▲	CLIP-Patch	B/32		10.6	9.7	6.7	6.4	6.4	16.7	16.7	7.5	10.4	9.0	14.3
△	CLIP-Patch	B/32	✓	18.1	13.3	13.3	8.6	10.8	17.5	**17.8**	9.0	14.2	10.6	20.3
■	CLIP-Patch	B/16		5.6	7.8	3.9	5.0	3.9	10.6	10.8	5.4	6.8	8.2	10.3
□	CLIP-Patch	B/16	✓	10.6	11.4	7.8	10.8	8.1	16.4	15.6	7.7	11.5	9.7	15.7
▲	CLIP-Grad.	B/32		13.6	10.6	9.2	7.5	7.2	13.9	12.8	8.3	10.7	9.6	13.8
△	CLIP-Grad.	B/32	✓	16.1	11.9	11.7	9.7	10.3	14.4	16.1	9.2	12.9	9.7	15.2
■	CLIP-Grad.	B/16		6.1	5.8	5.0	5.0	4.7	8.3	6.9	4.9	6.0	7.3	8.8
□	CLIP-Grad.	B/16	✓	8.1	10.8	8.6	8.6	6.7	11.1	11.4	6.7	9.3	8.6	11.6
◆	MDETR	B3		3.1	6.9	4.4	7.2	4.7	7.8	8.9	5.3	6.2	8.3	9.8
◇	MDETR	B3	✓	3.1	7.2	5.0	6.9	4.7	8.1	8.9	5.4	6.3	8.4	9.9
▲	OWL	B/32		23.1	26.1	14.4	18.3	11.7	13.9	13.1	11.1	17.2	16.6	25.4
△	OWL	B/32	✓	**32.8**	26.4	**19.4**	**19.4**	**16.1**	**19.2**	14.4	**12.6**	**21.1**	16.9	26.7
■	OWL	B/16		25.8	23.6	15.3	17.2	12.5	13.1	13.9	11.4	17.3	16.2	24.8
□	OWL	B/16	✓	31.9	**26.9**	18.9	**19.4**	14.7	18.1	15.8	**12.6**	20.8	**17.2**	**27.5**
ProcTHOR fine-tune (supervised) [18]				n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	27.4	66.4

Results

CoW breeds				Uncom.	Appear.	Space	PASTURE		Hid.	Hid.	Avg.		RoboTHOR	
							Appear. distract	Space distract		distract				
ID	Localizer	Arch.	Post	SR	SR	SR	SR	SR	SR	SR	SPL	SR	SPL	SR
▲	CLIP-Ref.	B/32		2.8	1.4	1.4	0.8	1.4	4.7	5.0	1.2	2.5	1.6	2.2
△	CLIP-Ref.	B/32	✓	3.6	0.6	1.7	0.6	1.7	2.2	2.5	0.9	1.8	1.0	1.8
■	CLIP-Ref.	B/16		1.4	1.7	1.7	1.9	1.9	2.8	2.2	1.7	1.9	2.4	2.6
□	CLIP-Ref.	B/16	✓	1.4	2.8	2.8	3.1	3.3	1.7	1.9	1.7	2.4	2.1	2.7
▲	CLIP-Patch	B/32		10.6	9.7	6.7	6.4	6.4	16.7	16.7	7.5	10.4	9.0	14.3
△	CLIP-Patch	B/32	✓	18.1	13.3	13.3	8.6	10.8	17.5	**17.8**	9.0	14.2	10.6	20.3
■	CLIP-Patch	B/16		5.6	7.8	3.9	5.0	3.9	10.6	10.8	5.4	6.8	8.2	10.3
□	CLIP-Patch	B/16	✓	10.6	11.4	7.8	10.8	8.1	16.4	15.6	7.7	11.5	9.7	15.7
▲	CLIP-Grad.	B/32		13.6	10.6	9.2	7.5	7.2	13.9	12.8	8.3	10.7	9.6	13.8
△	CLIP-Grad.	B/32	✓	16.1	11.9	11.7	9.7	10.3	14.4	16.1	9.2	12.9	9.7	15.2
■	CLIP-Grad.	B/16		6.1	5.8	5.0	5.0	4.7	8.3	6.9	4.9	6.0	7.3	8.8
□	CLIP-Grad.	B/16	✓	8.1	10.8	8.6	8.6	6.7	11.1	11.4	6.7	9.3	8.6	11.6
◆	MDETR	B3		3.1	6.9	4.4	7.2	4.7	7.8	8.9	5.3	6.2	8.3	9.8
◇	MDETR	B3	✓	3.1	7.2	5.0	6.9	4.7	8.1	8.9	5.4	6.3	8.4	9.9
▲	OWL	B/32		23.1	26.1	14.4	18.3	11.7	13.9	13.1	11.1	17.2	16.6	25.4
△	OWL	B/32	✓	**32.8**	26.4	**19.4**	**19.4**	**16.1**	**19.2**	14.4	**12.6**	**21.1**	16.9	26.7
■	OWL	B/16		25.8	23.6	15.3	17.2	12.5	13.1	13.9	11.4	17.3	16.2	24.8
□	OWL	B/16	✓	31.9	**26.9**	18.9	**19.4**	14.7	18.1	15.8	**12.6**	20.8	**17.2**	**27.5**
ProcTHOR fine-tune (supervised) [18]				n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	27.4	66.4

Results

category	PASTURE Uncom. △ SR	SPL	△ SR	SPL
GingerbreadHouse	20.0	14.3	**26.7**	**18.6**
EspressoMachine	10.0	7.7	**46.7**	**24.6**
Crate	23.3	18.2	**40.0**	**27.0**
ElectricGuitar	16.7	10.0	**46.7**	**30.8**
RiceCooker	3.3	2.9	**20.0**	**11.6**
LlamaWickerBasket	16.7	12.6	**30.0**	**24.5**
Whiteboard	**63.3**	**43.2**	30.0	18.7
Surfboard	26.7	20.6	**60.0**	**38.9**
Tricycle	10.0	9.0	**53.3**	**31.7**
GraphicsCard	3.3	2.1	**13.3**	**6.0**
Mate	**0.0**	**0.0**	**0.0**	**0.0**
ToyAirplane	0.0	0.0	**26.7**	**13.7**

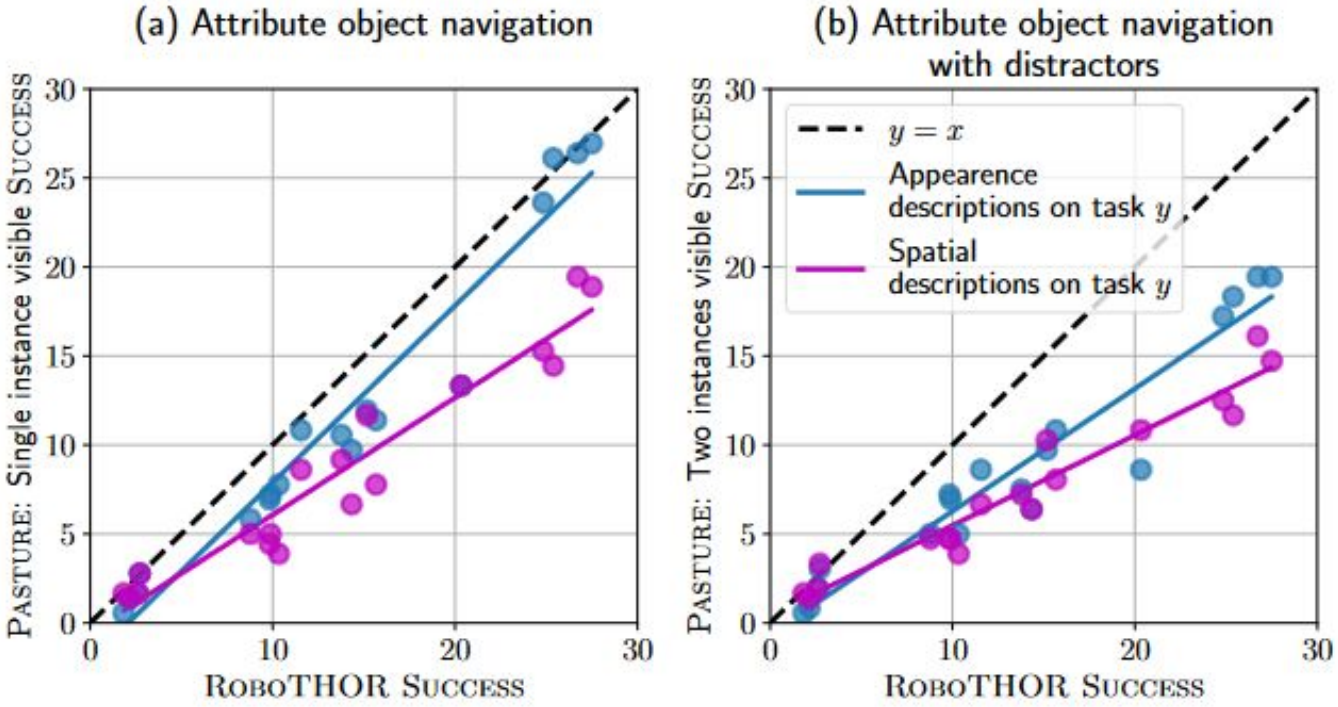
Results

category	PASTURE Appear. △ SR	PASTURE Appear. △ SPL	PASTURE Appear. △ SR	PASTURE Appear. △ SPL	PASTURE Space △ SR	PASTURE Space △ SPL	PASTURE Space △ SR	PASTURE Space △ SPL
ALARMCLOCK	6.7	4.0	**23.3**	**10.7**	3.3	3.3	**10.0**	**3.8**
APPLE	6.7	5.7	**36.7**	**17.0**	**10.0**	**8.4**	3.3	1.0
BASEBALLBAT	0.0	0.0	**3.3**	**1.2**	3.3	2.5	**6.7**	**2.7**
BASKETBALL	6.7	2.8	**36.7**	**24.1**	10.0	5.6	**36.7**	**26.8**
BOWL	3.3	0.5	**13.3**	**5.9**	10.0	5.6	**16.7**	**6.9**
GARBAGECAN	26.7	20.2	**50.0**	**31.5**	30.0	23.0	**40.0**	**23.2**
HOUSEPLANT	20.0	16.9	**30.0**	**20.2**	13.3	10.8	**40.0**	**21.9**
LAPTOP	13.3	10.6	**20.0**	**11.5**	13.3	9.6	**20.0**	**13.7**
MUG	10.0	7.5	**46.7**	**27.4**	10.0	**7.5**	**13.3**	5.4
SPRAYBOTTLE	16.7	13.6	**33.3**	**19.2**	**16.7**	**15.8**	**16.7**	6.8
TELEVISION	10.0	**10.0**	**13.3**	8.9	6.7	6.4	**20.0**	**9.9**
VASE	**23.3**	**17.5**	10.0	9.1	**13.3**	**9.8**	10.0	5.4

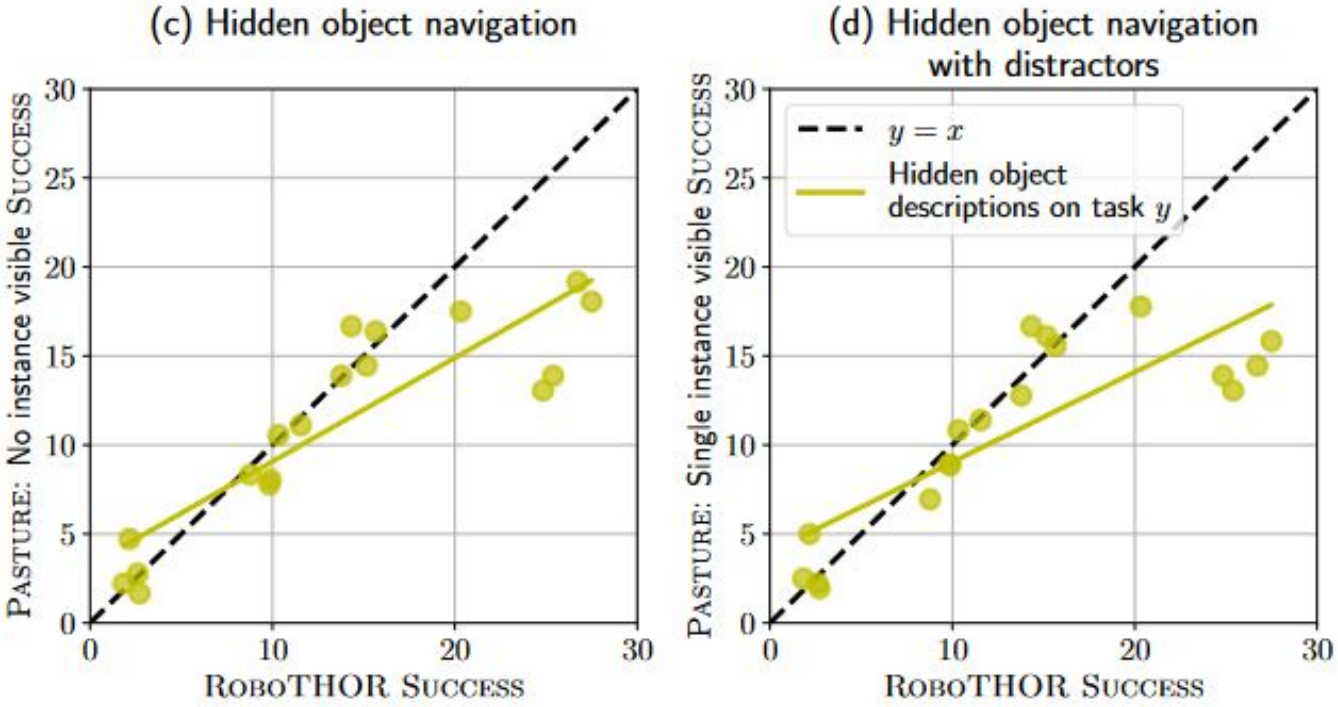
>

category	PASTURE Appear. distract △ SR	PASTURE Appear. distract △ SPL	PASTURE Appear. distract △ SR	PASTURE Appear. distract △ SPL	PASTURE Space distract △ SR	PASTURE Space distract △ SPL	PASTURE Space distract △ SR	PASTURE Space distract △ SPL
ALARMCLOCK	3.3	3.0	**13.3**	**6.5**	**6.7**	**6.3**	**6.7**	2.8
APPLE	**10.0**	6.4	**10.0**	**7.4**	3.3	3.3	**10.0**	**4.0**
BASEBALLBAT	0.0	0.0	**13.3**	**4.5**	0.0	0.0	**10.0**	**8.1**
BASKETBALL	6.7	3.3	**20.0**	**12.6**	**16.7**	9.4	**16.7**	**9.7**
BOWL	3.3	3.2	**16.7**	**8.5**	10.0	8.6	**23.3**	**12.6**
GARBAGECAN	26.7	19.9	**30.0**	**21.6**	13.3	10.5	**26.7**	**18.2**
HOUSEPLANT	10.0	6.0	**16.7**	**11.0**	13.3	10.8	**23.3**	**13.7**
LAPTOP	16.7	**13.6**	**23.3**	11.9	**16.7**	**11.9**	**16.7**	11.8
MUG	6.7	5.1	**26.7**	**17.8**	**10.0**	**7.8**	6.7	2.7
SPRAYBOTTLE	13.3	12.4	**26.7**	**15.2**	16.7	**15.4**	**20.0**	8.0
TELEVISION	6.7	6.6	**26.7**	**15.5**	6.7	3.3	**20.0**	**12.8**
VASE	**13.3**	**10.2**	10.0	8.6	10.0	6.5	**13.3**	**8.4**

Results


(a) Attribute object navigation
(b) Attribute object navigation with distractors
PASTURE: Single instance visible SUCCESS
PASTURE: Two instances visible SUCCESS
ROBOTHOR SUCCESS
0
5
10
15
20
25
30
$y = x$
Appearence descriptions on task y
Spatial descriptions on task y

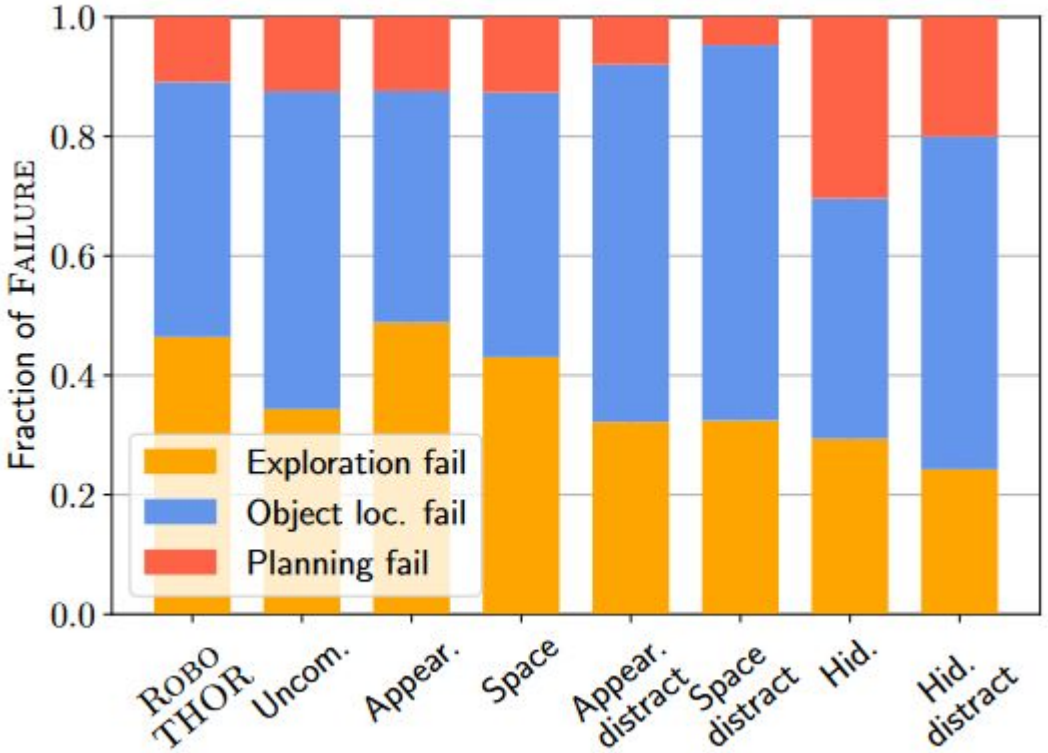
Results


(c) Hidden object navigation
(d) Hidden object navigation with distractors
PASTURE: No instance visible SUCCESS
PASTURE: Single instance visible SUCCESS
ROBOTHOR SUCCESS
ROBOTHOR SUCCESS
0
5
10
15
20
25
30
$y = x$
Hidden object descriptions on task y

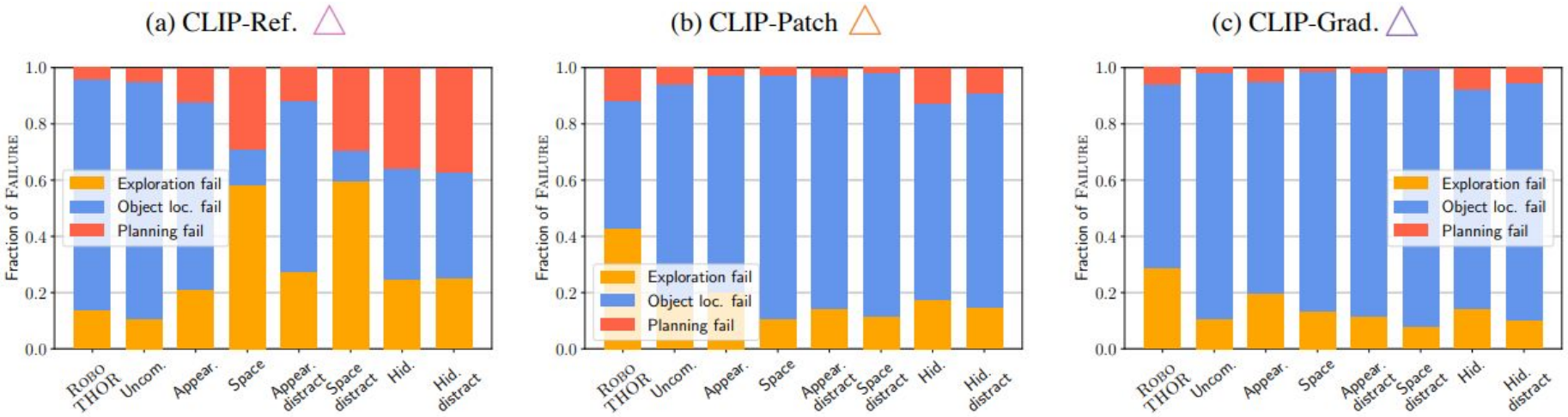
Results

ID	CoW breeds Loc.	Arch.	Post	HABITAT (MP3D) SPL	SR	ROBOTHOR (subset) SPL	SR	ROBOTHOR (full) SPL	SR	Nav. training steps
△	CLIP-Grad.	B/32	✓	**4.9**	9.2	15.0	23.7	9.7	15.2	**0**
△	OWL	B/32	✓	3.7	7.4	**20.8**	**32.5**	**16.9**	**26.7**	**0**
EmbCLIP-ZSON [37]				–	–	–	8.1	–	14.0*	60M
SemanticNav-ZSON [44]				4.8	**15.3**	–	–	–	–	500M

Failure Analysis



Failure Analysis



Ablations

ID	CoW breeds Loc.	Arch.	Post	Exp. Strategy	PASTURE (Avg.) SPL	SR	ROBOTHOR SPL	SR
△	OWL	B/32	✓	ROBOTHOR learn.	10.2	17.3	13.1	20.9
△	OWL	B/32	✓	HABITAT learn.	8.6	19.4	9.8	20.4
△	OWL	B/32	✓	FBE	**12.6**	**21.1**	**16.9**	**26.7**

Ablations

ID	Loc.	CoW breeds Arch.	Post	Exp. Strategy	RoboTHOR SPL	SR
△	CLIP-Grad.	B/32	✓	photo	9.5	**15.2**
△	CLIP-Grad.	B/32	✓	prompt ens.	**9.7**	**15.2**

Observations (Uncommon)

RoboTHOR Objects

Pasture Uncommon Objects

Strengths

- Thorough Ablations
- Truly Zero-Shot ObjectNav (given the Frontier-Based Exploration)
- Unique and Varied ObjectNav subtasks
- Modular: Exploration/Encoders

Weaknesses

- Point estimates. What about confidence intervals? More Trials? Low SR makes confident comparison difficult
- Expensive to get hyperparameter tuning images
- FBE not as generic as learned exploration (How to use FBE for Room Rearrangement, PointNav, etc?).

Future Work

- Would we get more benefit from combining appearance and spatial descriptors? Would this reduce the number of false positives or false negatives?
- Different exploration heuristics for different embodied tasks.
- Ablations with k != 9? Could those work with CLIP-Ref?
- Varied Embodied Agents & Continuous Action Spaces

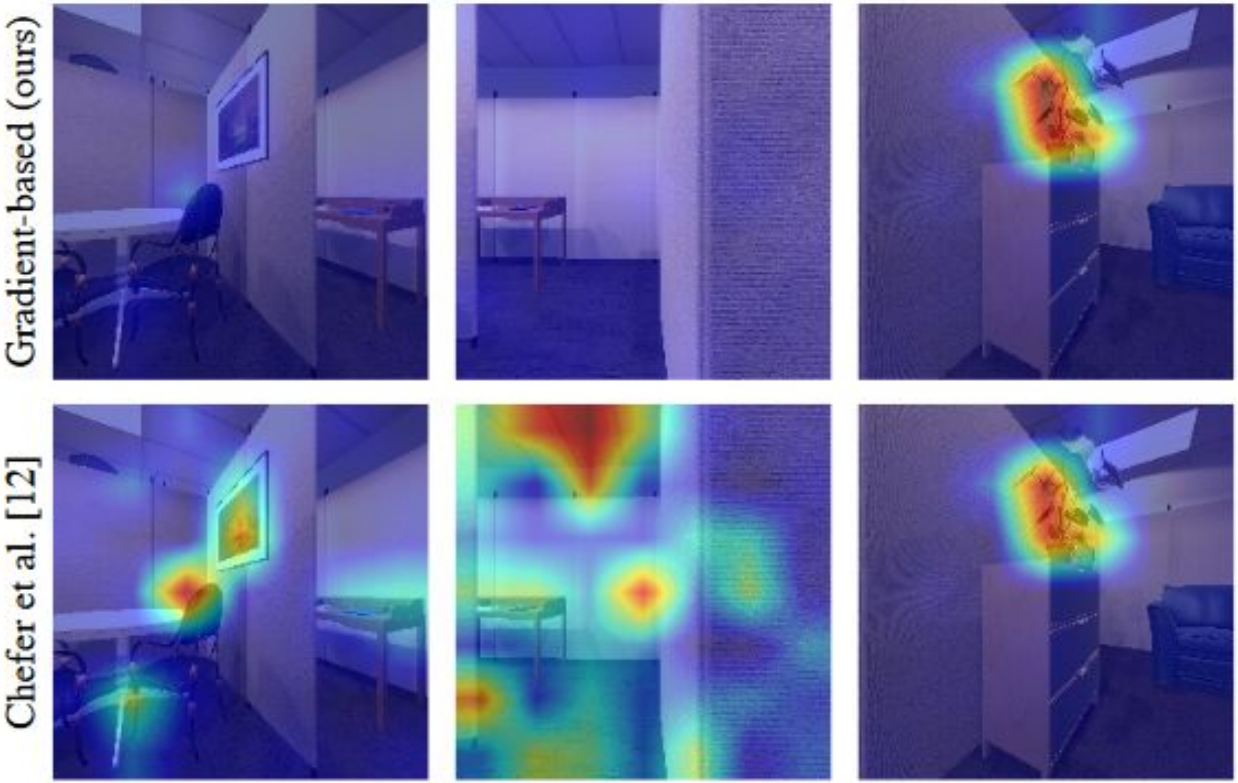
Discussion

- Why does higher compute lead to worse performance?
- Uncommon ObjectNav details. Whiteboard results may bring up questions.
- Ideas for zero-shot hyperparameter selection? Test-time/dynamic thresholds?

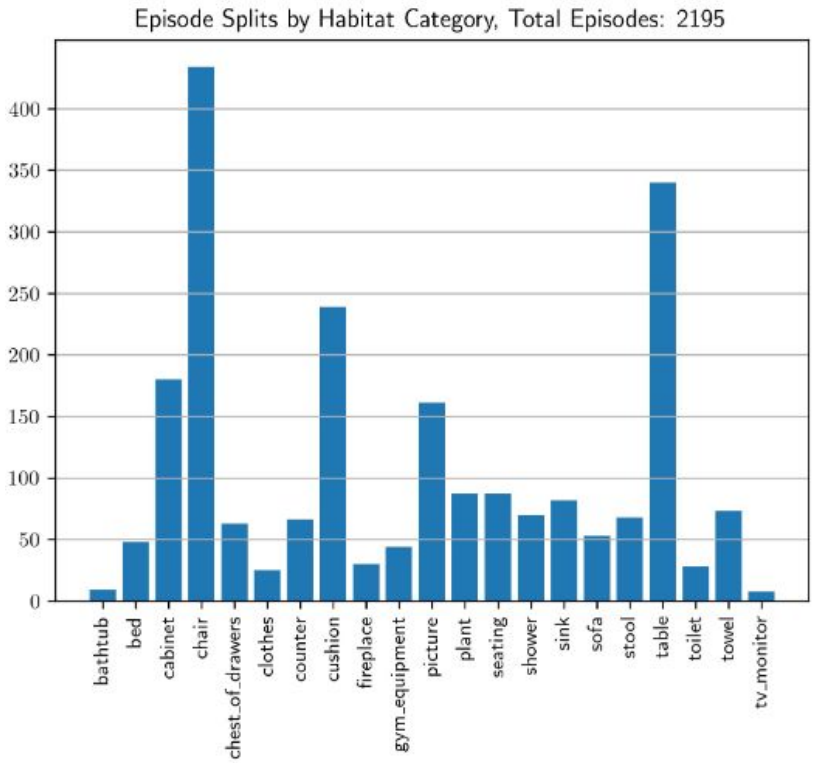
Other

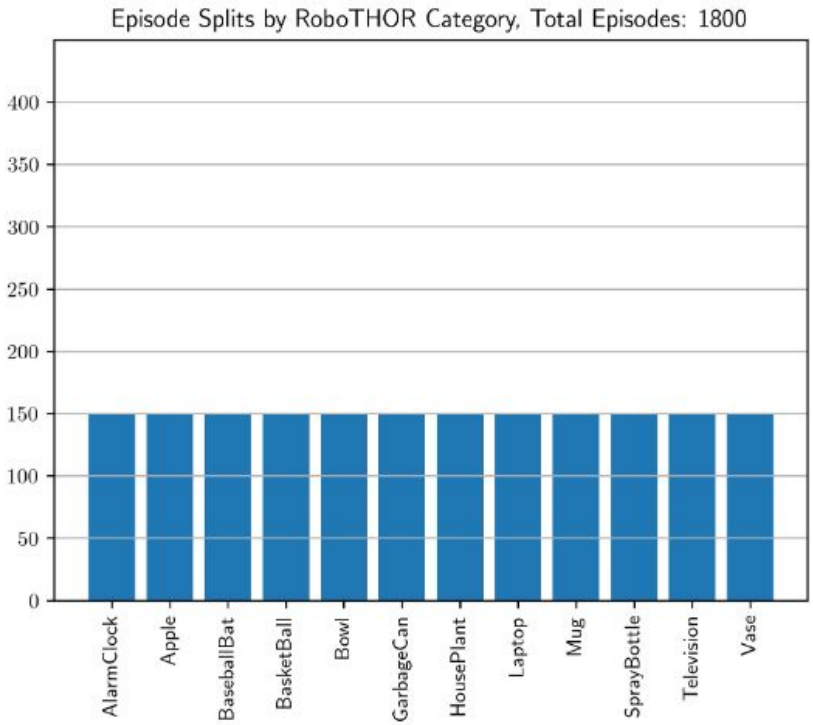
IDs	Localizer	Arch.	HABITAT	ROBOTHOR and PASTURE
▲, △	CLIP-Ref.	B/32	–	0.25
■, □	CLIP-Ref.	B/16	–	0.125
▲, △	CLIP-Patch	B/32	–	0.875
■, □	CLIP-Patch	B/16	–	0.75
▲, △	CLIP-Grad.	B/32	0.375	0.625
■, □	CLIP-Grad.	B/16	–	0.375
◆, ◇	MDETR	B3	–	0.95
▲, △	OWL	B/32	0.2	0.125
■, □	OWL	B/16	–	0.125

Other



Other


Episode Splits by Habitat Category, Total Episodes: 2195
0
50
100
150
200
250
300
350
400
bathtub
bed
cabinet
chair
chest_of_drawers
clothes
counter
cushion
fireplace
gym_equipment
picture
plant
seating
shower
sink
sofa
stool
table
toilet
towel
tv_monitor


Episode Splits by RoboTHOR Category, Total Episodes: 1800
0
50
100
150
200
250
300
350
400
AlarmClock
Apple
BaseballBat
BasketBall
Bowl
GarbageCan
HousePlant
Laptop
Mug
SprayBottle
Television
Vase