CoWs on Pasture:

Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation

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

Columbia University in the City of New York   University of Washington

1 Task

Give a natural language description of an object, with potentially many attributes, the task is to find the object.

2 CoW Baselines

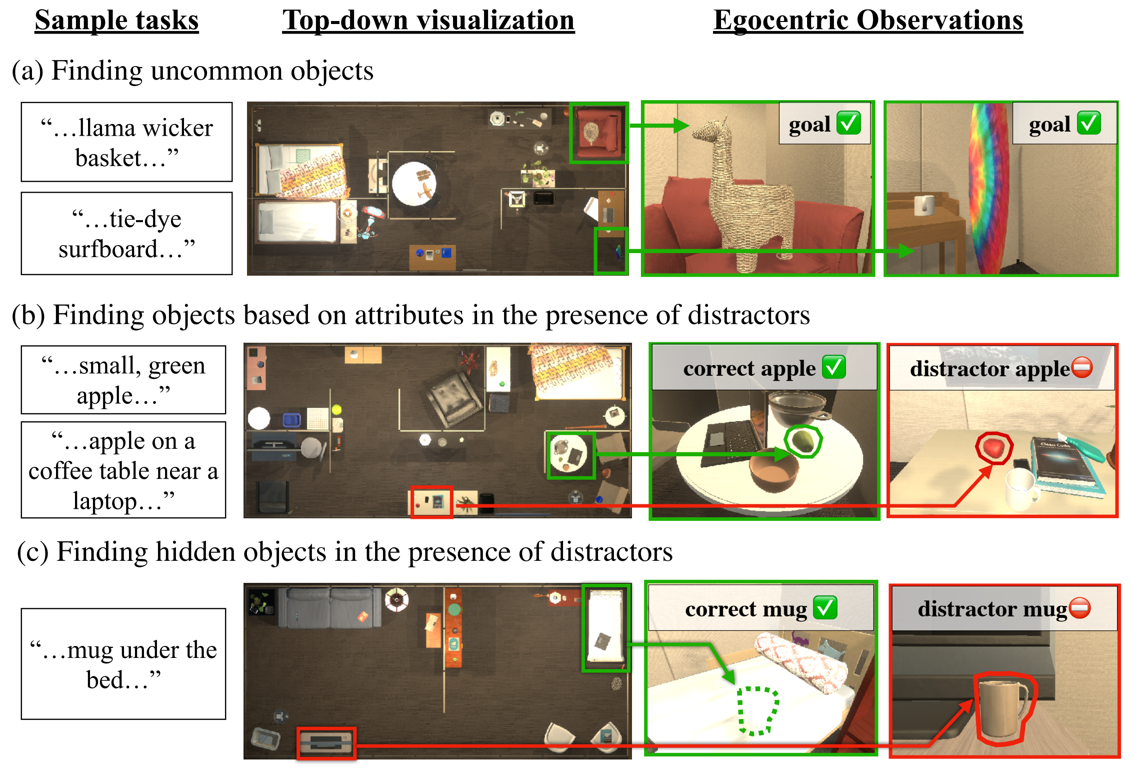
```
If object is in view:
    move to it
else:
    explore
```

Plug in an object localizer

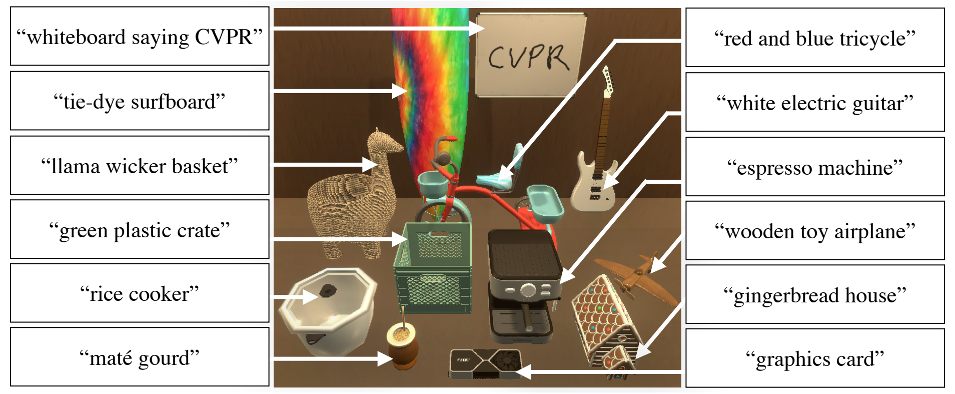
Plug in a policy

3 Pasture Benchmark

Sample tasks

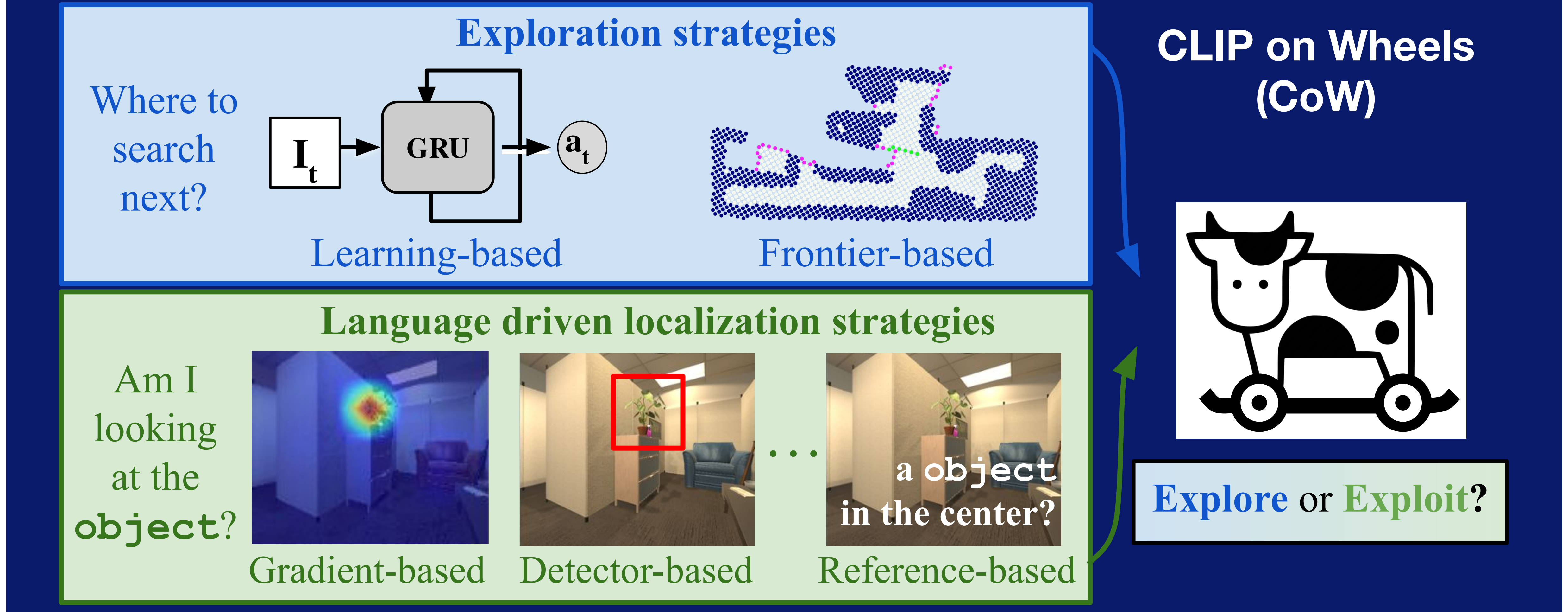


Uncommon object navigation targets

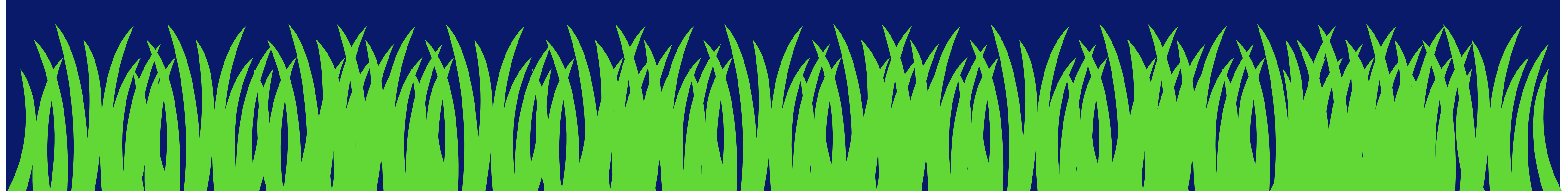


We turn CLIP models into **zero-shot object navigators**

without additional training



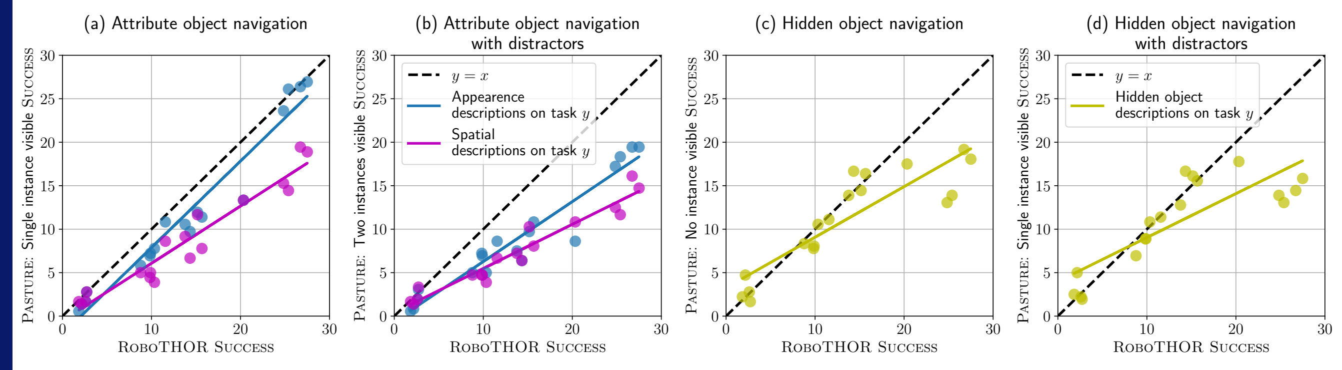
Take a **CLIP** model, put it **on Wheels**, get a **CoW**!



Paper: Code:

4 Attributes, Occlusion & Distractors

CoW baselines struggle to take advantage of spatial and appearance object attributes. Distractor objects hurt performance less when finding hidden objects.



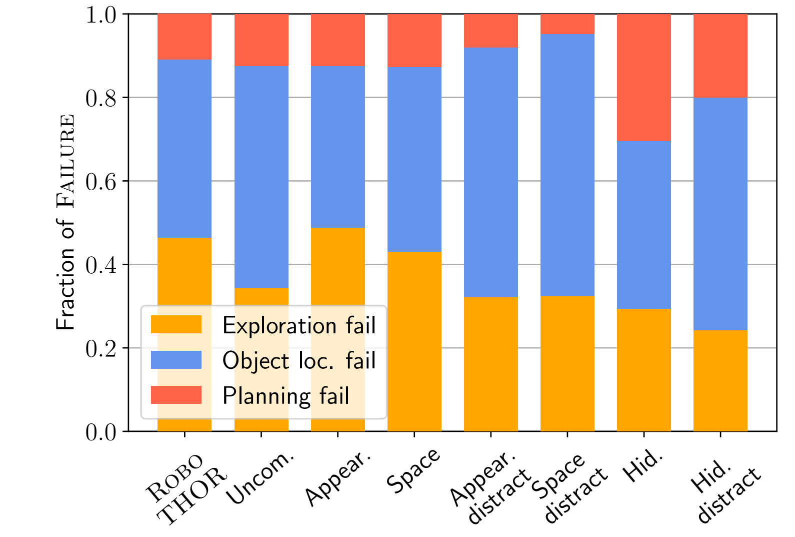
5 LLM Priors

Incorporating object and scene priors helps.

	CoW breeds			Pasture Uncom.		RoboTHOR	
ID	Loc.	Arch.	Obj. Prior	SPL	SR	SPL	SR
▲	OWL	B/32	None	20.5	32.8	16.8	26.7
▲	OWL	B/32	GPT-3.5	**22.2**	**36.9**	**17.0**	**27.5**

6 Failure Analysis

CoWs may improve with better perception and exploration strategies.



7 Comparison to Prior Art

CoWs beat baselines that train for millions of steps.

	CoW breeds		Habitat (MP3D)		RoboTHOR (subset)		RoboTHOR (full)		Nav. training steps
ID	Loc.	Arch.	SPL	SR	SPL	SR	SPL	SR	
▲	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 [38]			–	–	–	8.1	–	14.0*	60M
SemanticNav-ZSON [46]			4.8	**15.3**	–	–	–	–	500M