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

Samir Yitzhak Gadre [1]

Mitchell Wortsman [2]

Gabriel Ilharco [2]

Ludwig Schmidt [2]

Shuran Song [1]

[1] Columbia University in the City of New York

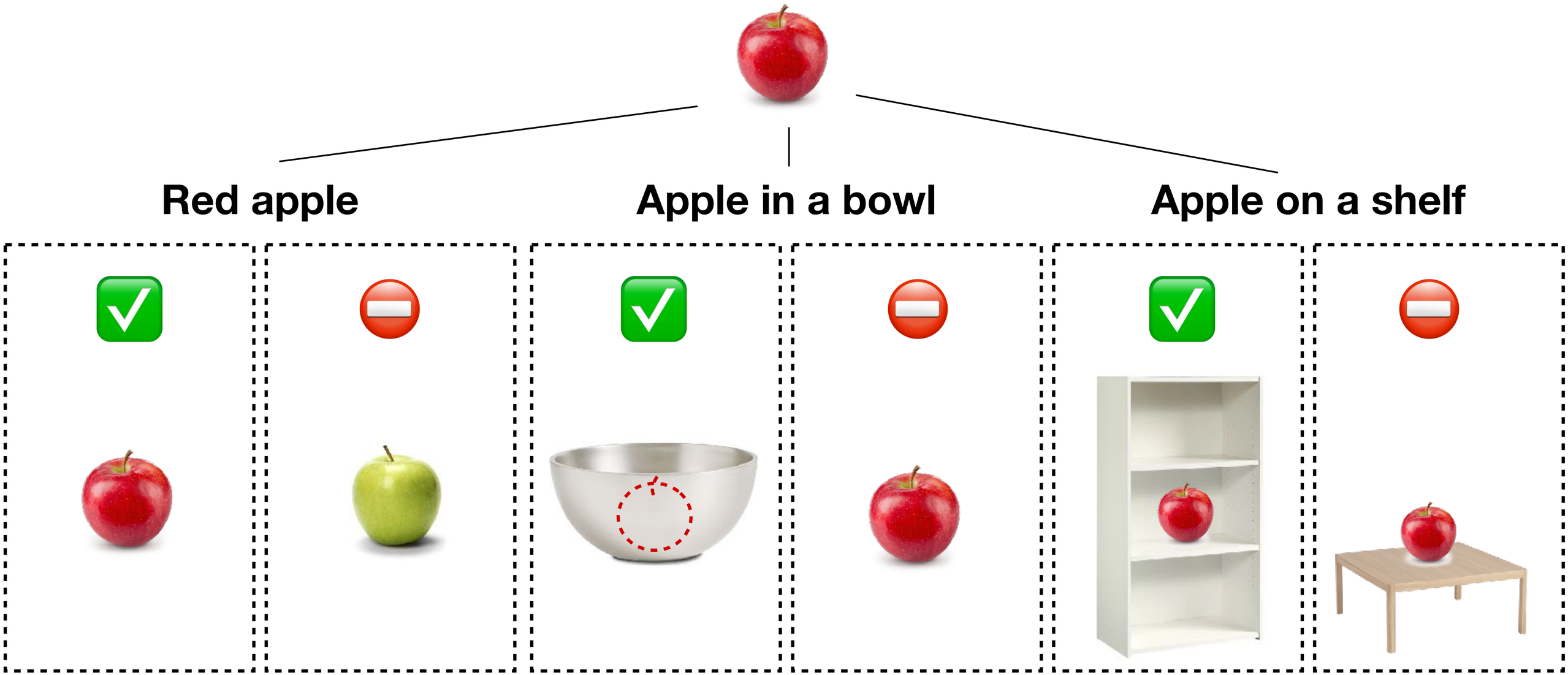
[2] University of Washington

Motivation: Zero-shot agents

- Want agents to find anything, even without additional training

- Move towards more general purpose A.I. systems

Motivation: Language-driven agents



Task

- Inputs:

Egocentric RGB + D

Language for the target object

"...apple..."

OR

"...apple on a table…"

OR

"...red apple…"

- Output:

Action: direction to move (or stop)

How would one do this task?

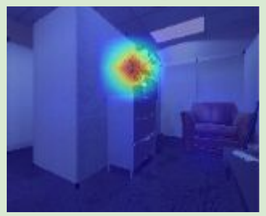
- Look around

- When you see what you are looking for, go to it!

CoW

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

Plug in an object localizer

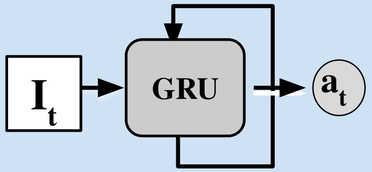


Gradient-based

Reference-based

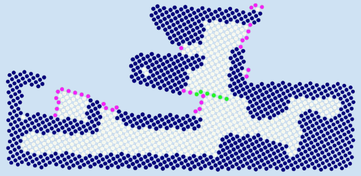
...

Detector-based



Learning-based

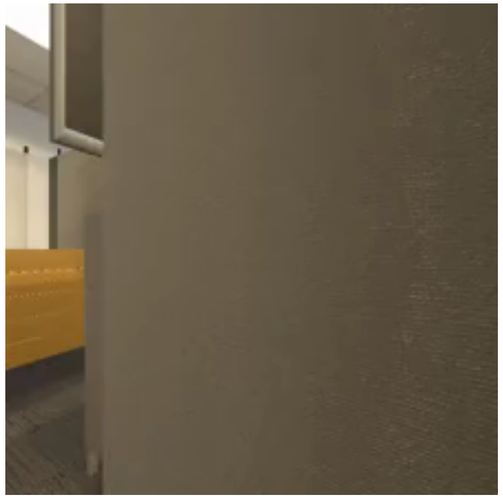
...



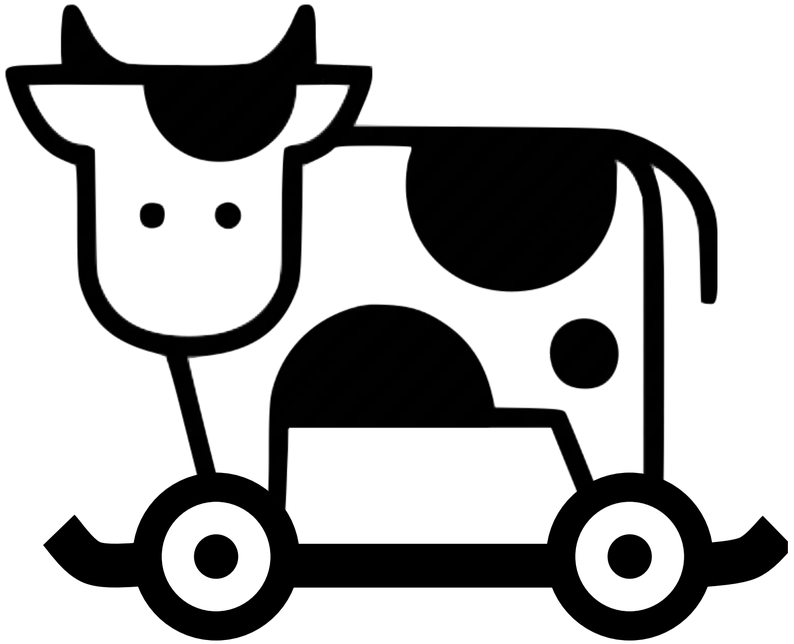
Frontier-based

Plug in a policy

Egocentric view

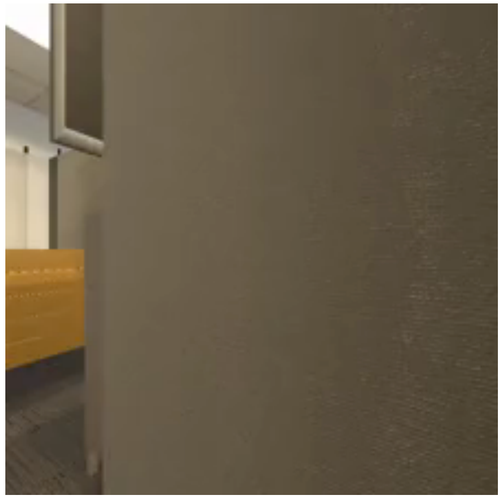


`explore`



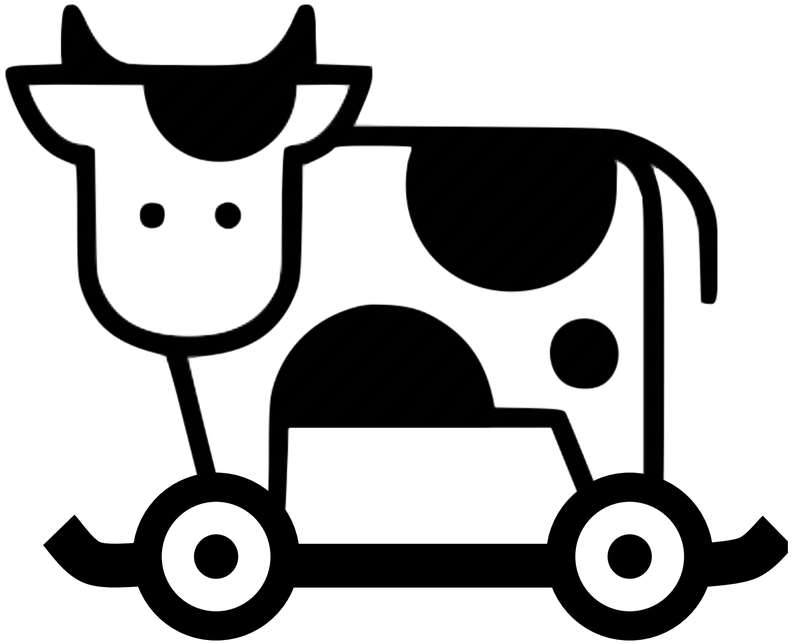
Target: plant!

Egocentric view



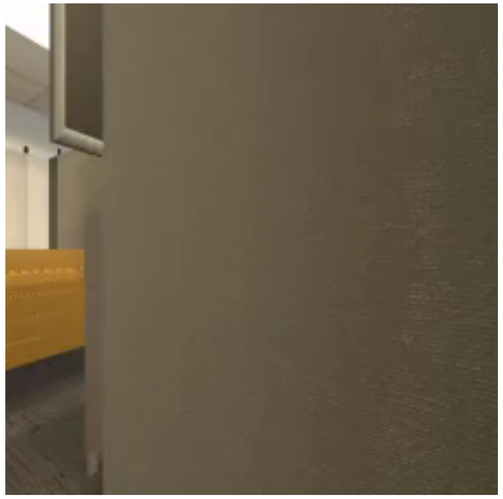
Object relevance

`explore`



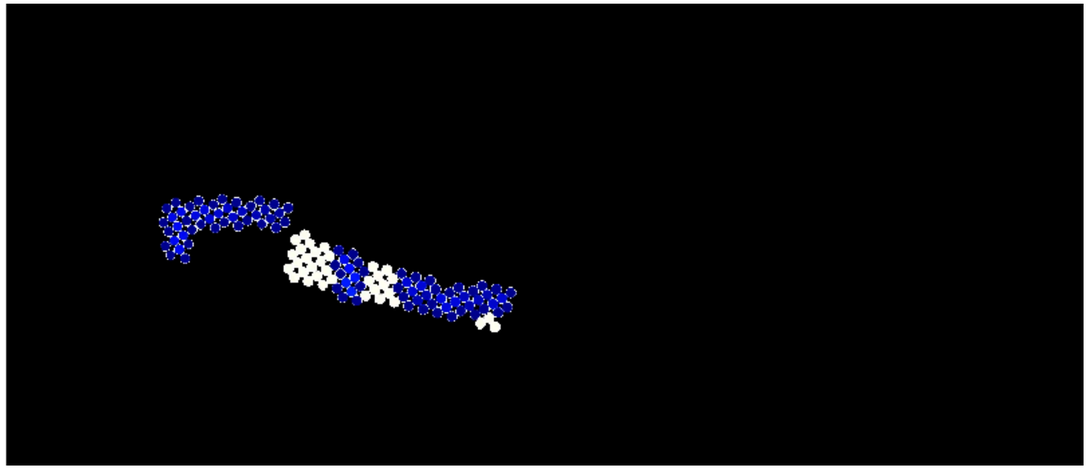
Target: plant!

Egocentric view

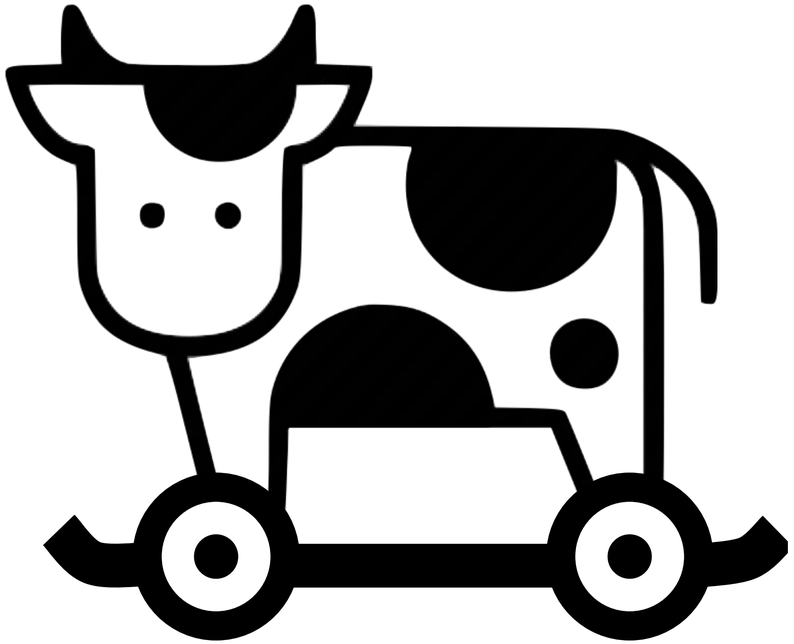


Object relevance

Voxel projected object relevance map

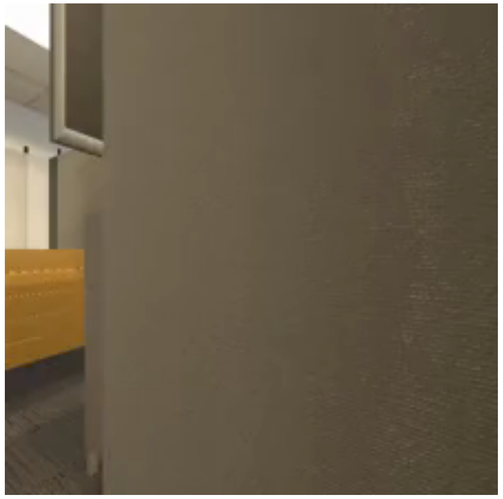


`explore`



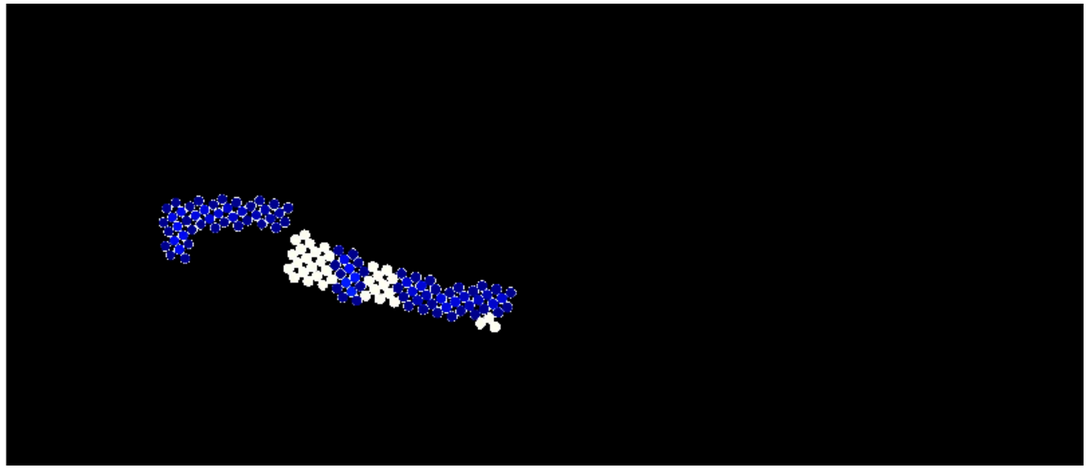
Target: plant!

Egocentric view

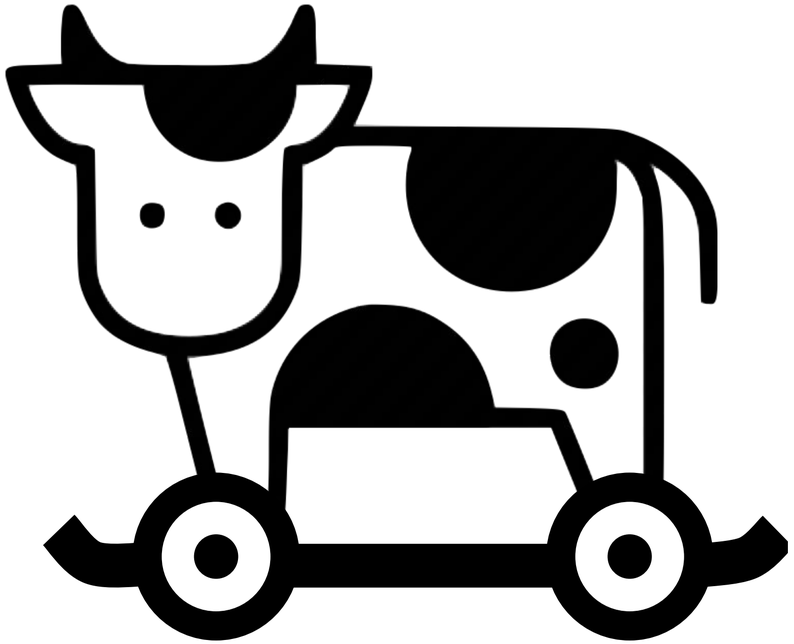


Object relevance

Voxel projected object relevance map



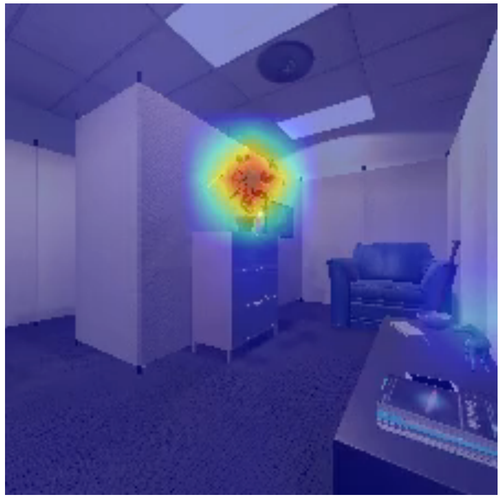
`explore`



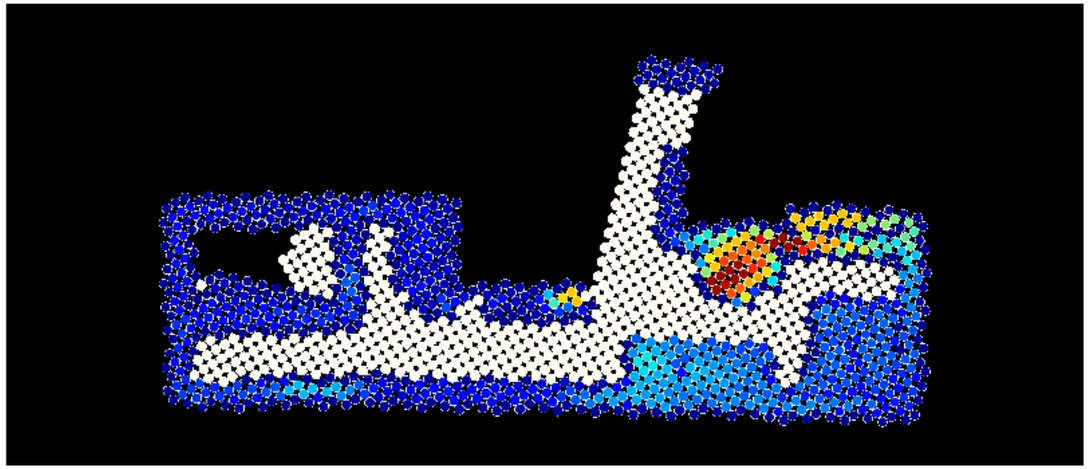
Target: plant!

Egocentric view

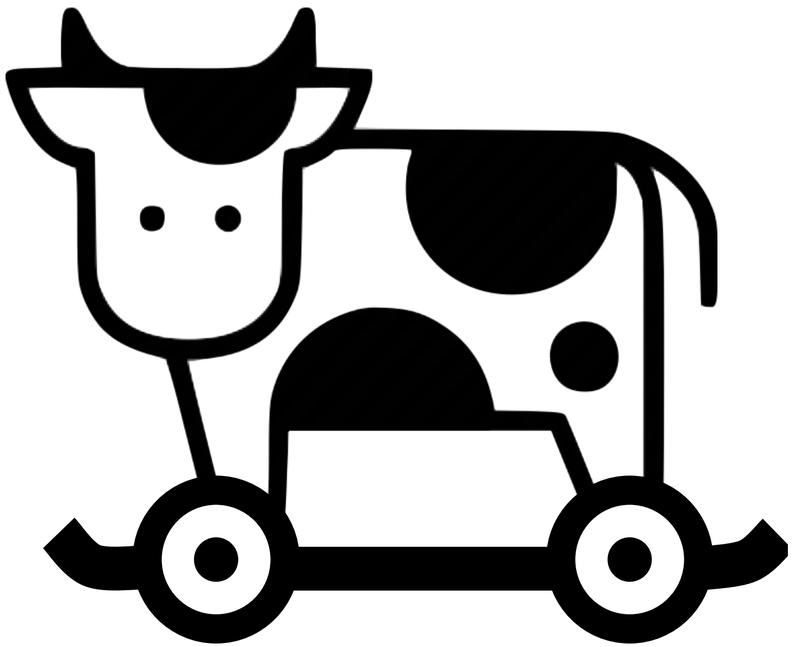
Object relevance



Voxel projected object relevance map



`object is in view`

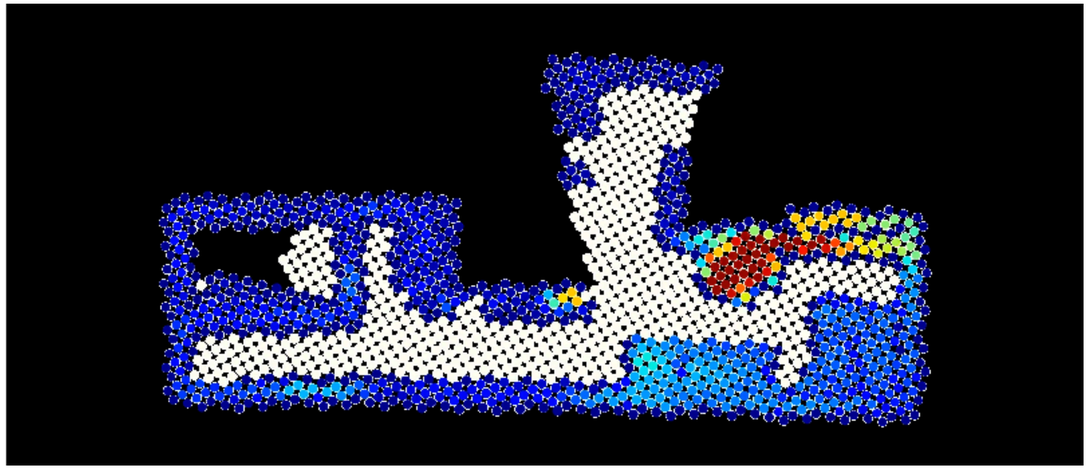


Target: plant!

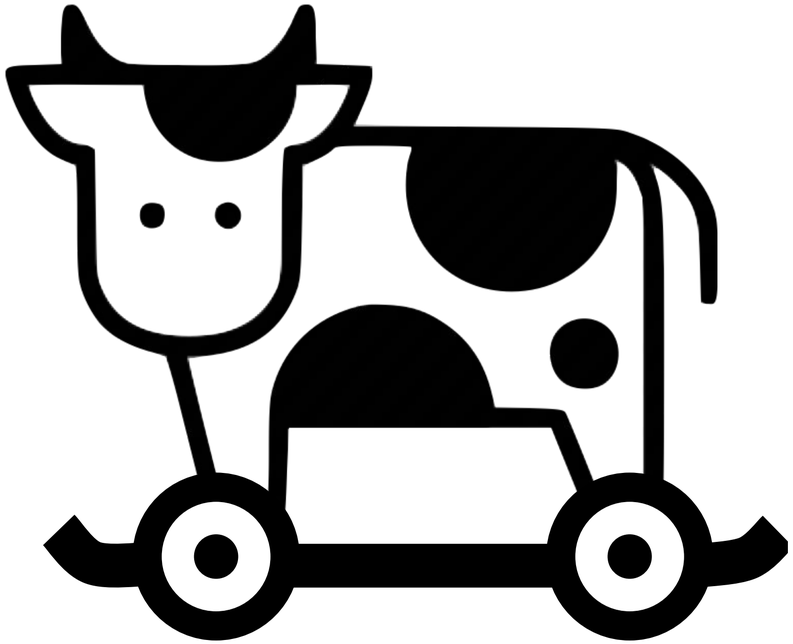
Egocentric view

Object relevance

Voxel projected object relevance map



`object is in view`



Target: plant!

Pasture: Uncommon Objects

“whiteboard saying CVPR”

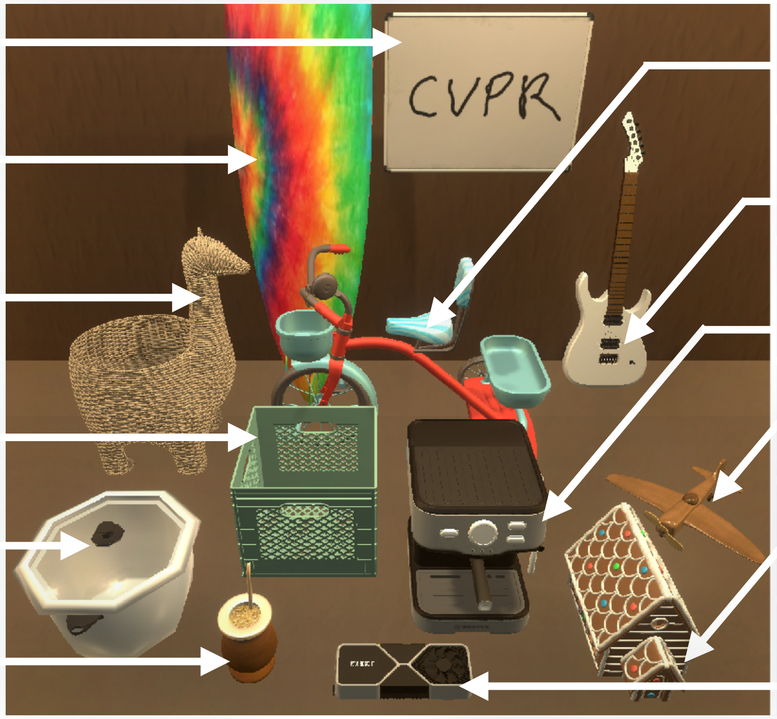
“tie-dye surfboard”

“llama wicker basket”

“green plastic crate”

“rice cooker”

“maté gourd”



“red and blue tricycle”

“white electric guitar”

“espresso machine”

“wooden toy airplane”

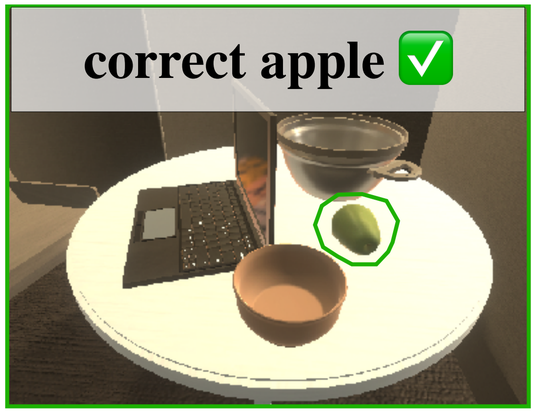
“gingerbread house”

“graphics card”

Pasture: Object Attributes

Appearance task: "…small, green apple…"

Spatial task: "…apple on a coffee table near a laptop…"

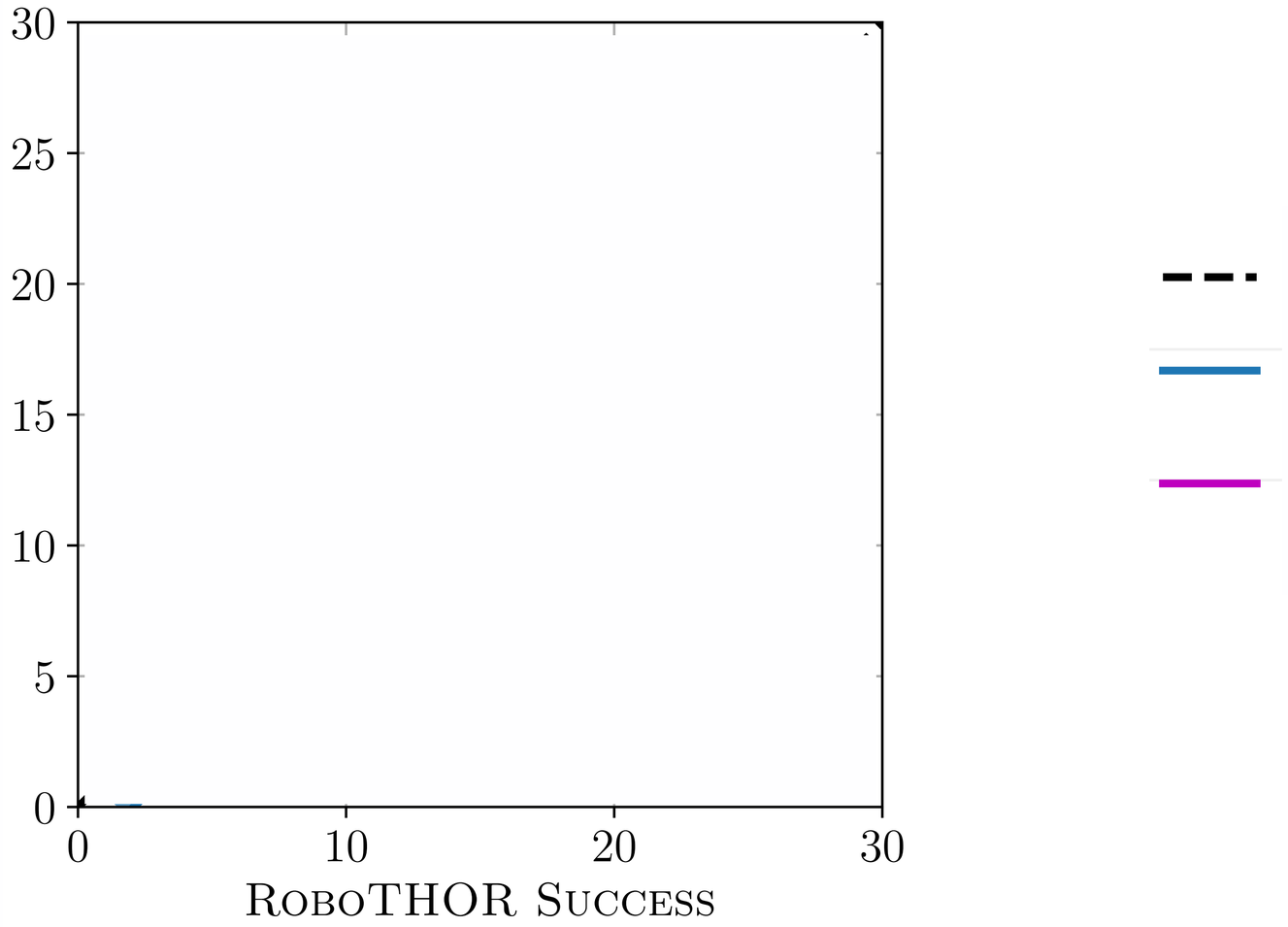


Pasture: Hidden objects

Hidden object task: "…mug under the bed…"

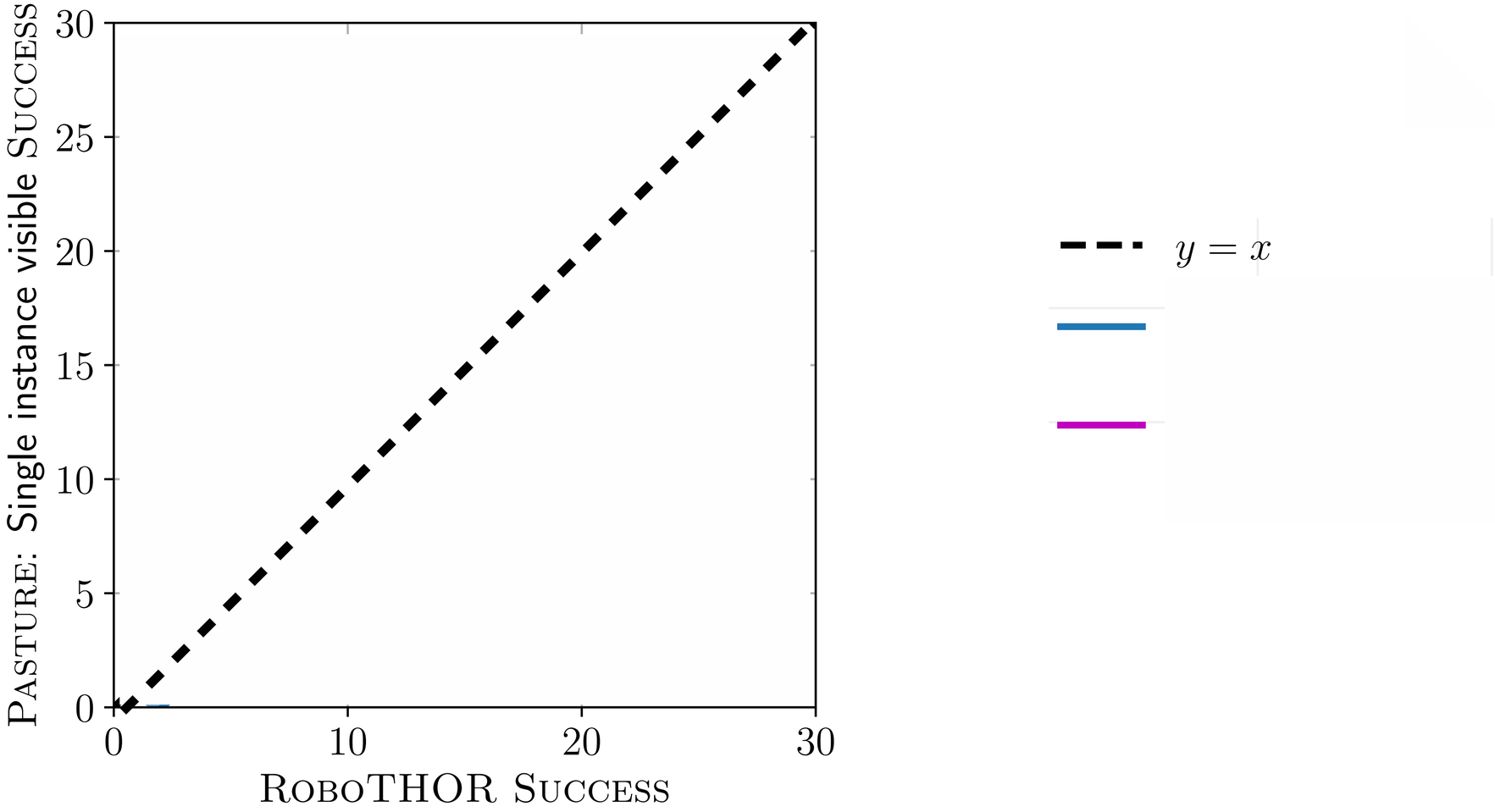
Results: Using attributes

(a) Attribute object navigation



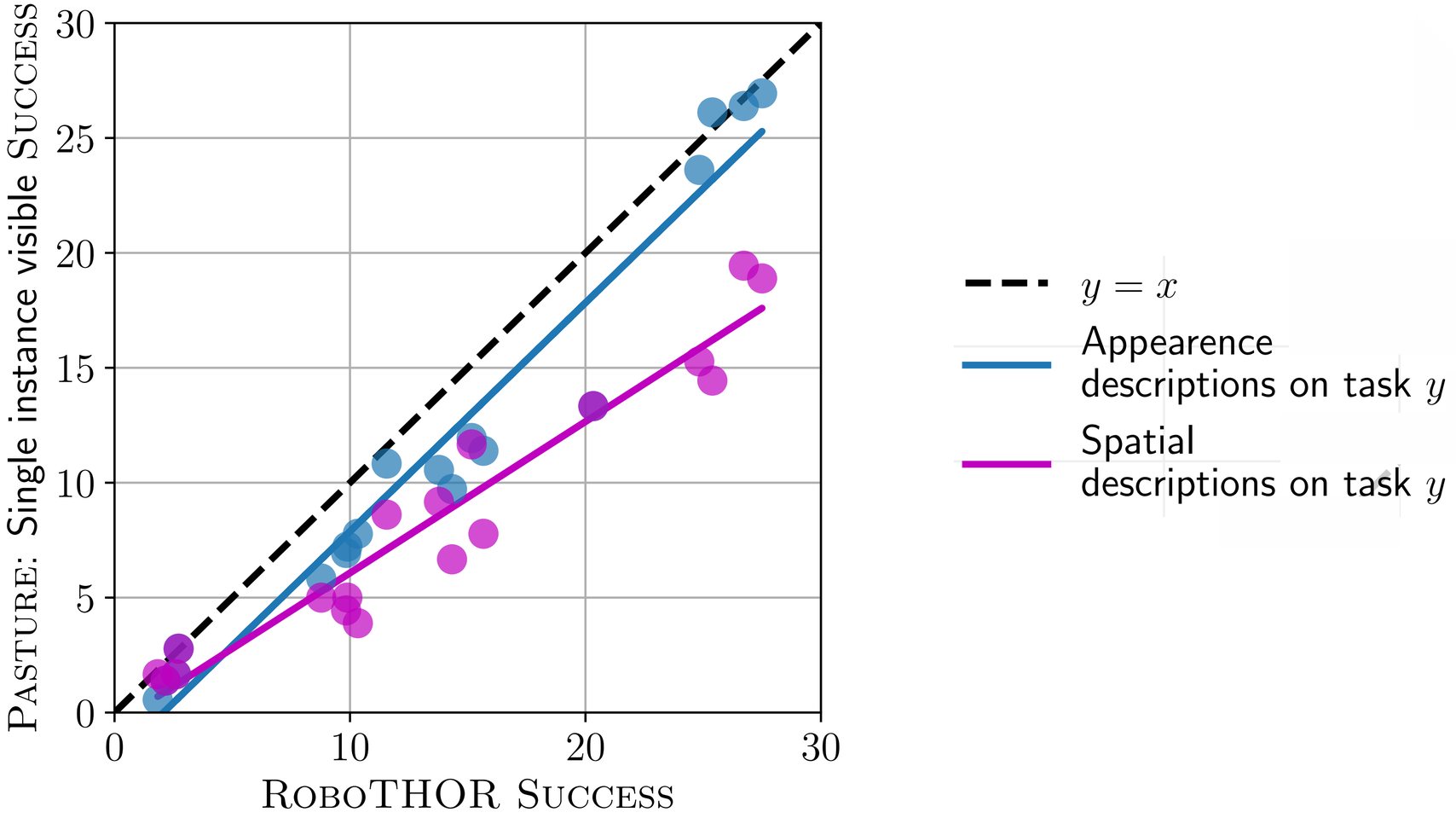
Results: Using attributes

(a) Attribute object navigation



Results: Using attributes

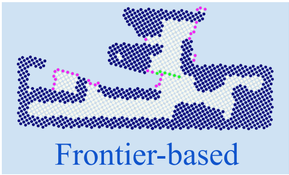
(a) Attribute object navigation



Results: Incorporating priors

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

+ GPT Priors

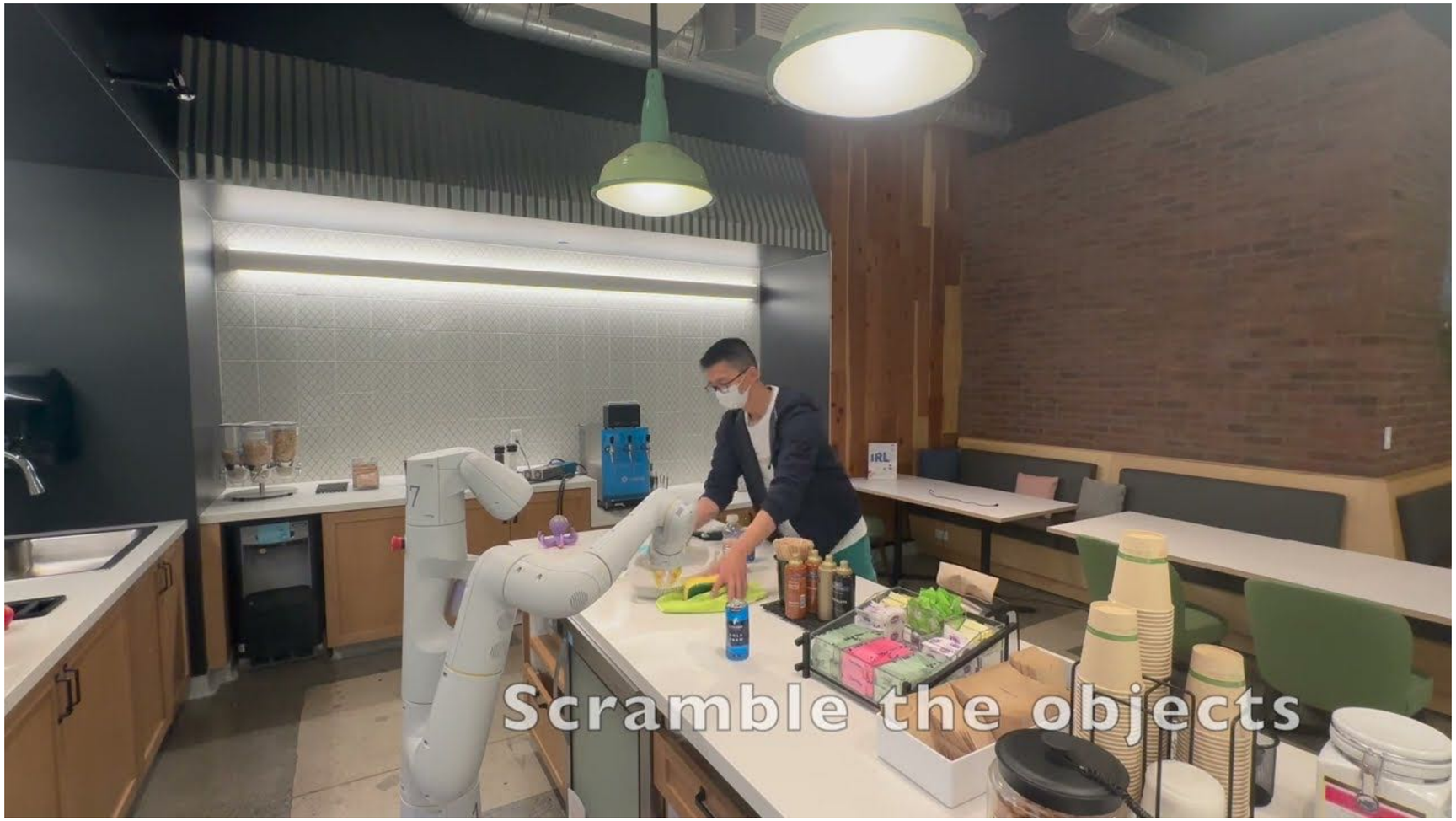


ID	CoW breeds Loc.	Arch.	Obj. Prior	PASTURE Uncom. SPL	SR	ROBOTHOR SPL	SR
▲	OWL	B/32	None	20.5	32.8	16.8	26.7
▲	OWL	B/32	GPT-3.5	**22.2**	**36.9**	**17.0**	**27.5**
					(+4.1)		**(+0.8)**

Results: Comparison to prior art

ID	CoW breeds Loc.	Arch.	HABITAT (MP3D) SPL	HABITAT (MP3D) SR	ROBOTHOR (subset) SPL	ROBOTHOR (subset) SR	ROBOTHOR (full) SPL	ROBOTHOR (full) 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 [38]			–	–	–	8.1	–	14.0*	60M
SemanticNav-ZSON [46]			4.8	**15.3**	–	–	–	–	500M

Future Directions: Real World Mobile Manipulation



Stone et al. Open-World Object Manipulation using Pre-Trained Vision-Language Model. 2023.

Key Takeaways

- Baselines, even if they are heuristic or naive, are incredibly important to contextualize the performance of learned methods
- Zero-shot object navigation is an important problem to work on, current methods are still in their infancy

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

Samir Yitzhak Gadre [1]

Mitchell Wortsman [2]

Gabriel Ilharco [2]

Ludwig Schmidt [2]

Shuran Song [1]

[1] Columbia University in the City of New York

[2] University of Washington