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

Red apple: ✅
Apple in a bowl: ✅
Apple on a shelf: ✅

Green apple: ⛔
Apple in a bowl: ⛔
Apple on a shelf: ⛔

Apple in a bowl: 🔄
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

Plug in a policy
Learning-based
Frontier-based

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

Target: plant!
Egocentric view  Object relevance

Target: plant!
Voxel projected object relevance map

Egocentric view

Object relevance

Target: plant!

explore
Egocentric view

Object relevance

Voxel projected object relevance map

Target: plant!
Egocentric view

Object relevance

Voxel projected object relevance map

Target: plant!
Egocentric view

Object relevance

Voxel projected object relevance map

Target: plant!
Pasture: Uncommon Objects

Parameters

- **“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…”

correct apple ✅
distractor apple ⛔
Pasture: Hidden objects

Hidden object task: “…mug under the bed…”
Results: Using attributes

(a) Attribute object navigation

![Graph showing success rates for attribute object navigation](image-url)
Results: Using attributes

(a) Attribute object navigation

![Graph showing the relationship between PASTURE: Single instance visible SUCCESS and RoboTHOR Success.]
Results: Using attributes

(a) Attribute object navigation

![Graph showing success rates for attribute object navigation vs. RoboTHOR success.](image)

- **Pasture**: Single instance visible
- **RoboTHOR**: Success

<table>
<thead>
<tr>
<th>Exp. Strategy</th>
<th>Pasture</th>
<th>RoboTHOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBE</td>
<td>13.1</td>
<td>20.9</td>
</tr>
<tr>
<td>GPT-3.5</td>
<td>9.8</td>
<td>20.4</td>
</tr>
</tbody>
</table>

We explore this hypothesis in Appx. Fig. 5.
Results: Incorporating priors

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

<table>
<thead>
<tr>
<th>ID</th>
<th>CoW breeds</th>
<th>PASTURE Uncom. SPL</th>
<th>ROBOTHOR SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Loc. Arch.</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>▲</td>
<td>OWL B/32</td>
<td>20.5 32.8</td>
<td>16.8 26.7</td>
</tr>
<tr>
<td>▲</td>
<td>OWL B/32</td>
<td>22.2 36.9</td>
<td>17.0 27.5</td>
</tr>
</tbody>
</table>

(+4.1) (+0.8)

Detector-based
Frontier-based

OWL ViT-B/32
ASTURE
REVERIE

Results: Incorporating priors

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

<table>
<thead>
<tr>
<th>ID</th>
<th>CoW breeds</th>
<th>PASTURE Uncom. SPL</th>
<th>ROBOTHOR SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Loc. Arch.</td>
<td>SR</td>
<td>SR</td>
</tr>
<tr>
<td>▲</td>
<td>OWL B/32</td>
<td>20.5 32.8</td>
<td>16.8 26.7</td>
</tr>
<tr>
<td>▲</td>
<td>OWL B/32</td>
<td>22.2 36.9</td>
<td>17.0 27.5</td>
</tr>
</tbody>
</table>

(+4.1) (+0.8)
Results: Comparison to prior art

<table>
<thead>
<tr>
<th>ID</th>
<th>CoW breeds</th>
<th>Loc.</th>
<th>Arch.</th>
<th>HABITAT (MP3D) SPL</th>
<th>SR</th>
<th>ROBOTHOR (subset) SPL</th>
<th>SR</th>
<th>ROBOTHOR (full) SPL</th>
<th>SR</th>
<th>Nav. training steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CLIP-Grad.</td>
<td>B/32</td>
<td>4.9</td>
<td>9.2</td>
<td>15.0</td>
<td>23.7</td>
<td>9.7</td>
<td>15.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>OWL</td>
<td>B/32</td>
<td>3.7</td>
<td>7.4</td>
<td>20.8</td>
<td>32.5</td>
<td>16.9</td>
<td>26.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>EmbCLIP-ZSON [38]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>14.0*</td>
<td>60M</td>
</tr>
<tr>
<td></td>
<td>SemanticNav-ZSON [46]</td>
<td>4.8</td>
<td>15.3</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>500M</td>
<td></td>
</tr>
</tbody>
</table>

*- The superior performance of SemanticNav-ZSON in all categories seen during training.

We would like to thank Jessie Chapman, et al. for valuable feedback. This work was supported in part by NSF Award. SYG is supported by a NSF Graduate Research Fellowship. Cheng Chi, Huy Ha, Zeyi Liu, Sachit Menon, and Sarah Pratt for their help.
Future Directions: Real World Mobile Manipulation

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
Mitchell Wortsman
Gabriel Ilharco
Ludwig Schmidt
Shuran Song

1 Columbia University
2 University of Washington