Hierarchical Representation Learning for Robot Assistants

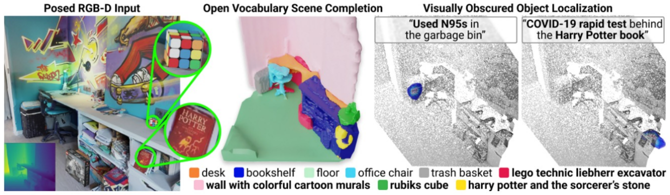
Columbia University: **Shuran Song, Carl Vondrick, Zhou Yu**

Project Goal: Just-in-time object delivery. From both sight and dialogue, our framework is able to anticipate what objects a person will need, localized deliver it at the right moment.

Border Impact: Expected results will enable machines to better understand and collaborate with people This framework has a wide range of applications such as allowing a robot pass the proper ingredients to the chef in kitchen, pass the right tool to the worker in factory, or pass the right equipment to doctors in an emergency room.

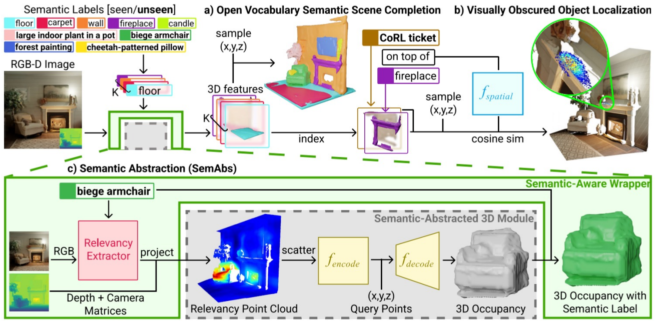
Open-vocabulary object localization and completion [1]

Goal: Localize objects in 3D and complete their 3D geometry from a open-set of vocabulary


Posed RGB-D Input
Open Vocabulary Scene Completion
Visually Obscured Object Localization
"Used N95s in the garbage bin"
"COVID-19 rapid test behind the Harry Potter book"
desk
bookshelf
floor
office chair
trash basket
lego technic liebherr excavator
wall with colorful cartoon murals
rubiks cube
harry potter and the sorcerer's stone

***Challenge**: How can we equip 2D Vision language model (e.g., CLIP) with new 3D capabilities, while maintaining their zero-shot robustness?*

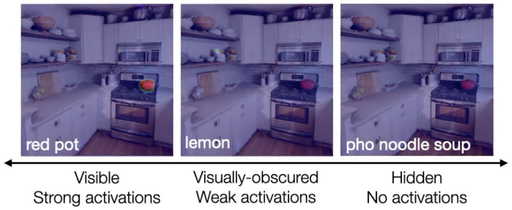
SemAbs Module for Completion & Localization



SemAbs is a reusable module for 3D scene understanding tasks. We train our the 3D module using data from a custom THOR simulator.

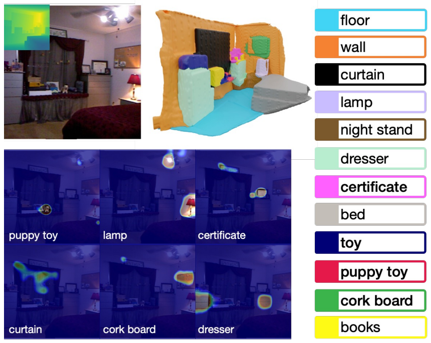
Key idea: Semantic abstracted 3D reasoning via relevancy

Relevancy activations ≈ **VLM's confidence** of whether and where object is in scene



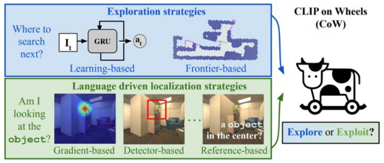
Semantic Aware	Semantic Abstracted
Complete the red pot	Complete *that object*
Inside the red pot	Inside *that* object

Zero-shot Sim2Real: Semantic Abstraction offloads visual-semantic reasoning challenges to CLIP. In doing so, its learned 3D spatial and geometric reasoning skills transfers sim2real in a zero-shot manner.

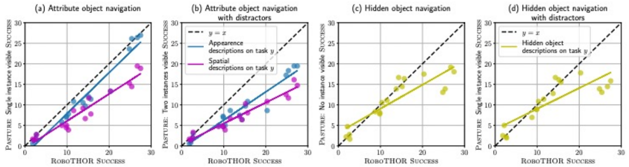


Language-Driven Zero-Shot Object Navigation [2]

We investigate a object navigation framework CLIP on Wheels (CoW), to adapt open-vocabulary models to this task without fine-tuning.



To evaluate L-ZSON, we introduce the **Pasture** benchmark, which considers finding uncommon objects, objects described by spatial and appearance attributes, and hidden objects described relative to visible objects.

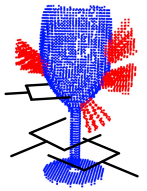


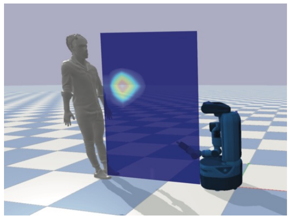
In general object navigation with descriptions is more challenging than the ROBOTHOR object navigation -- trend lines lying below the y = x line.

Robot-Human Handover [3]

Goal: Robot need to grasp novel object and successfully pass it to human subject. To do so the algorithm consider and optimized the following three factors:

1) Infer stable grasp on novel object
2) Infer human preferred grasp. Filter candidate robot grasps based on human preferred grasp to reduce overlap and avoid collision
3) Choose Object Transfer Point so that the human preferred grasp is accessible to user and improve human comfort using arm joint torque model





[1] Semantic Abstraction: Open-World 3D Scene Understanding from 2D Vision-Language Models. CoRL'22 (semantic-abstraction.cs.columbia.edu)
[2] CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation. CVPR'22 (cow.cs.columbia.edu)
[3] Robot-human handover (ongoing)

2023 FRR & NRI Principal Investigators' Meeting
May 2-3, 2023

Award ID#: 2132519