# Abstraction as Inductive Bias:

## Open-World 3D Scene Understanding without Open-World 3D Data

Shuran Song
@Columbia University

CoRL 2022 Workshop Geometry, Physics, and Human Knowledge as Inductive Bias in Robot Learning

# Abstraction as Inductive Bias

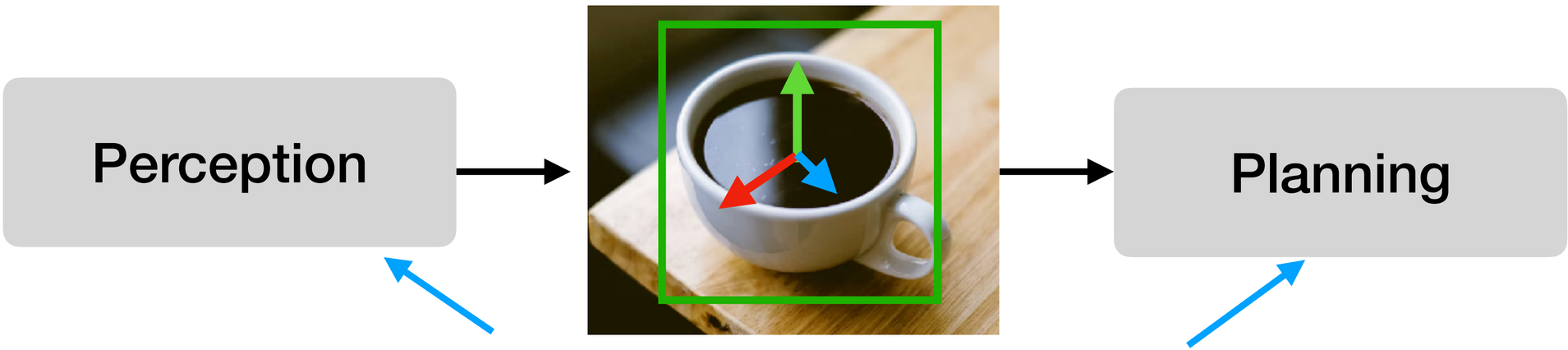
How to introduce **prior knowledge** into learning models?

The very first inductive bias we injected to the system
is the way we define or decompose the tasks (i.e., abstractions)

# Abstraction as Inductive Bias

How to introduce **prior knowledge** into learning models?

The very first inductive bias we injected to the system
is the way we define or decompose the tasks (i.e., abstractions)




Abstraction: bounding box or 6DoF pose

# Abstraction as Inductive Bias

## Limitations:

- Cannot adapt based on different end tasks
- Error propagation

# Abstraction as Inductive Bias

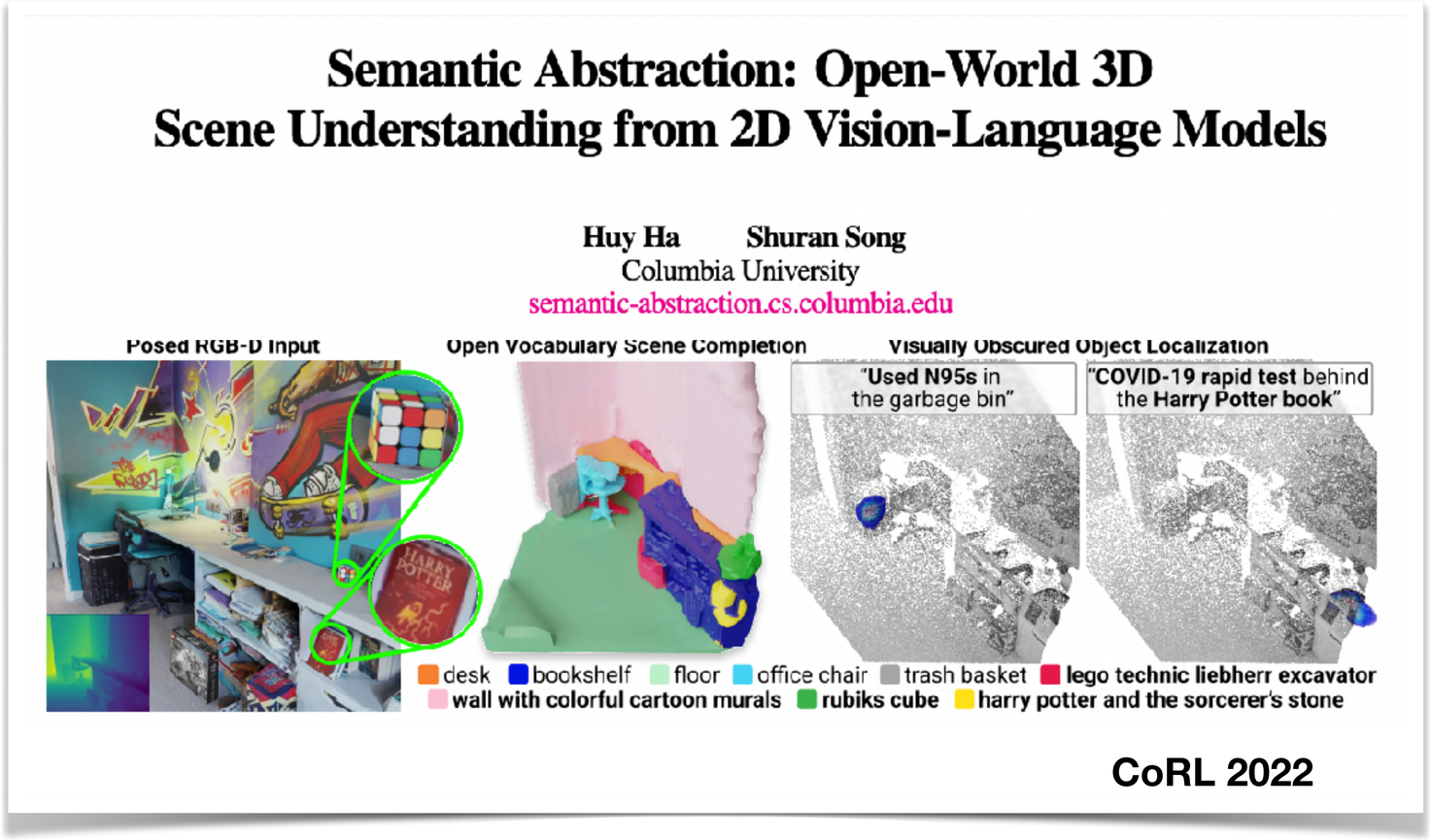
## Limitations:

- Cannot adapt based on different end tasks
- Error propagation

## Advantages:

- Decomposition -> Learn from less data
- Discover underlying structure -> Generalize beyond training data
- Abstraction as inductive bias is unavoidable.

# Abstraction as Inductive Bias




*"Abstraction as Inductive Bias" in the Age of Large Pre-trained Models*

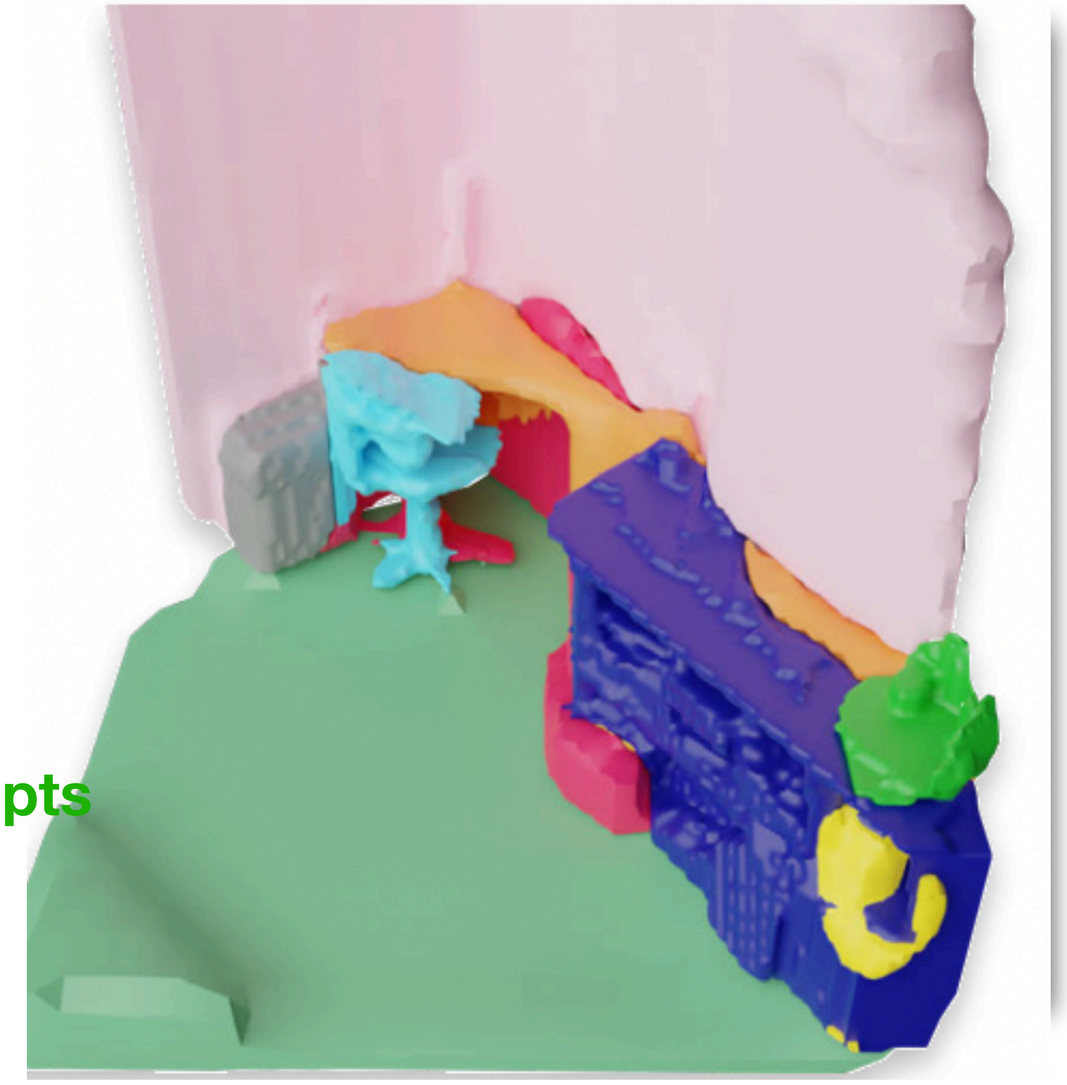
# Open World 3D Scene Understanding

Recognize a dynamic set of **semantic** categories and ground these semantics with **spatial** information in the **3D environment**.

"Find the **COVID-19 rapid test** behind **Harry Potter book**"

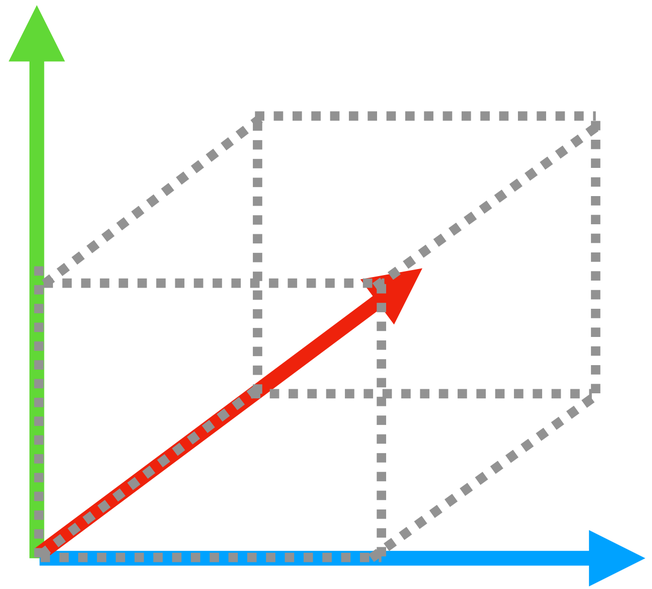
**Spatial Concepts** **Visual Semantics Concepts**

**Geometry Concepts** →



# Open World 3D Scene Understanding

**Visual Semantics Concepts** X **Geometry Concepts** X **Spatial Concepts**



Learning the joint distribution requires dataset encompass all the combinations

# Open World 3D Scene Understanding

**Visual Semantics Concepts** **Geometry Concepts** **Spatial Concepts**

Presents different properties
and can be learned in different manner and dataset

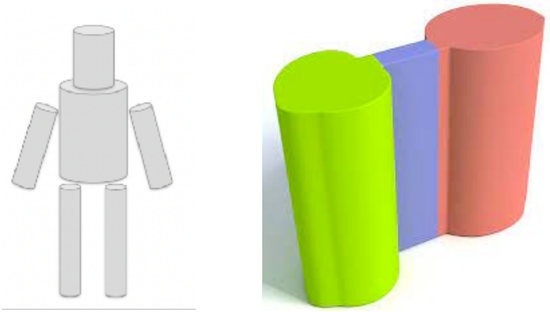
# Open World 3D Scene Understanding

## Visual Semantics Concepts

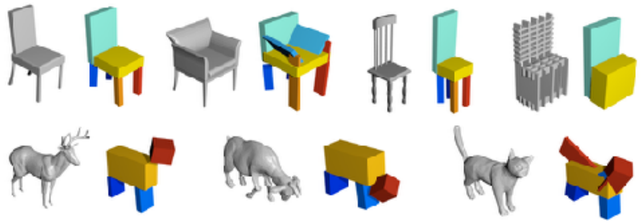
Large and non-decomposable


## Geometry Concepts

Large but decomposable



Generalized Cylinder Brook & Binford 1979



Learning Shape Abstractions, Tulsiani et al, 2017

## Spatial Concepts

Small and finite

Table 1. *Prepositions of English*

| | | |
|---|---|---|
| about | between | outside |
| above | betwixt | over |
| across | beyond | past |
| after | by | through |
| against | down | throughout |
| along | from | to |
| alongside | in | toward |
| amid(st) | inside | under |
| among(st) | into | underneath |
| around | near | up |
| at | nearby | upon |
| atop | off | via |
| behind | on | with |
| below | onto | within |
| beneath | opposite | without |
| beside | out | |
| *Compounds* | | |
| far from | on top of | |
| in back of | to the left of | |
| in between | to the right of | |
| in front of | to the side of | |
| in line with | | |
| *Intransitive prepositions* | | |
| afterward(s) | forward | right |
| apart | here | sideways |
| away | inward | south |
| back | left | there |
| backward | N-ward (e.g., homeward) | together |
| downstairs | north | upstairs |
| downward | outward | upward |
| east | | west |

Spatial Language, Landau & Jackendoff, 1993

# Open World 3D Scene Understanding

**Visual Semantics Concepts** >> **Geometry Concepts**
**Spatial Concepts**

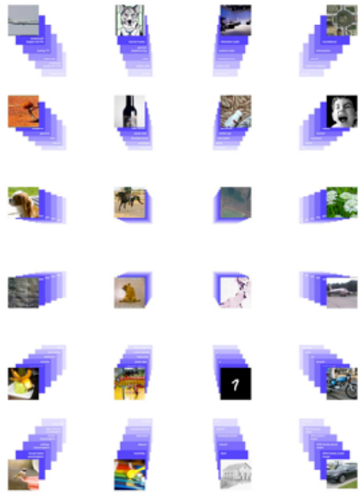
Requires exposure to
internet-scale datasets

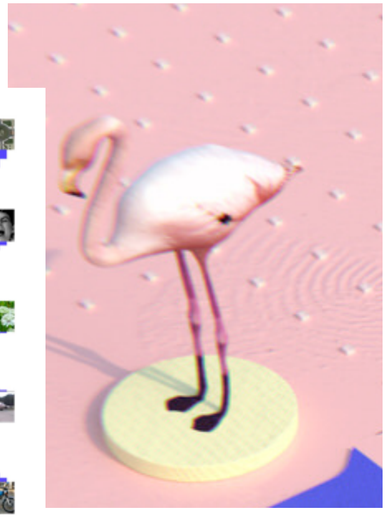
Tractable even with
limited synthetic datasets

# Open World 3D Scene Understanding

**Visual Semantics Concepts**

CLIP





Flamingo

**Requires exposure to internet-scale datasets**

**>>**

**Geometry Concepts**
**Spatial Concepts**

✓ Understands a large collection of visual-semantic concepts

✓ Robustness by learning from internet scale data.

# Open World 3D Scene Understanding

**Visual Semantics Concepts** >> **Geometry Concepts**
**Spatial Concepts**

CLIP

Flamingo

**How to add 3D reasoning skills?**

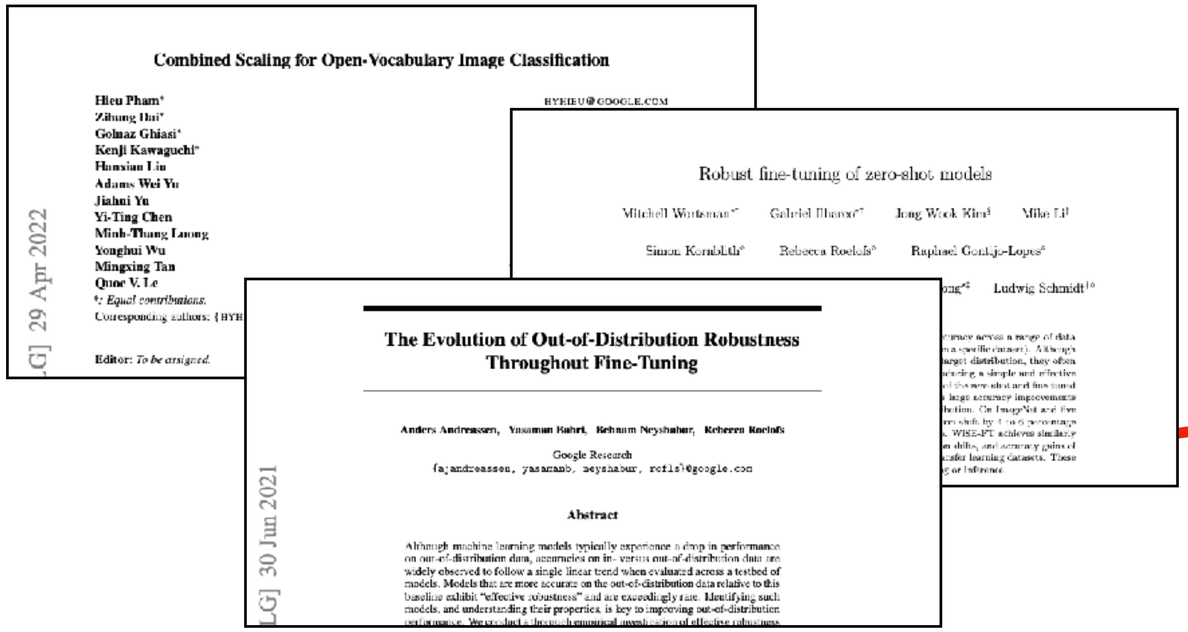
**Finetune with 3D data !?**

**Requires exposure to internet-scale datasets**

# Open World 3D Scene Understanding

**Visual Semantics Concepts** **>>** **Geometry Concepts**

**Spatial Concepts**




**How to add 3D reasoning skills?**

**Finetune with 3D data !?**

Naïvely finetuning VLMs on a limited and task-specific dataset results in reduced model's **robustness** and **generality.**

# Semantic Abstraction

**Visual Semantics Concepts**

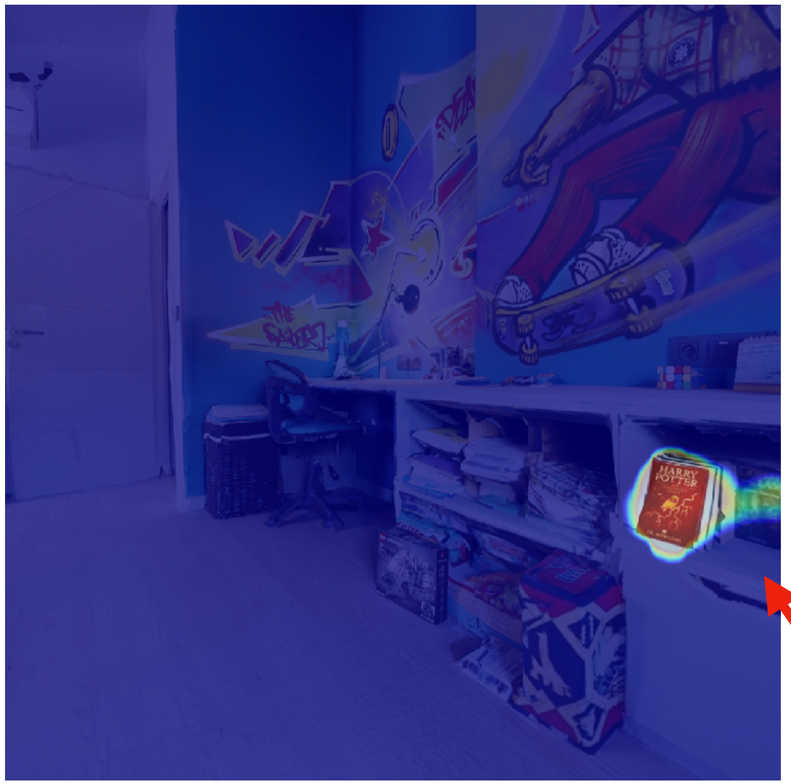
**Geometry Concepts**

**Spatial Concepts**

Goal: Reason about 3D concepts in a semantic-agnostic manner.

# Semantic Abstraction

## Visual Semantics Concepts




2D localization of the object

## Geometry Concepts
## Spatial Concepts

Goal: Reason about 3D concepts in a semantic-agnostic manner.

**Learn different concepts for**

``behind the Harry Potter book”

``behind the trashcan”

…

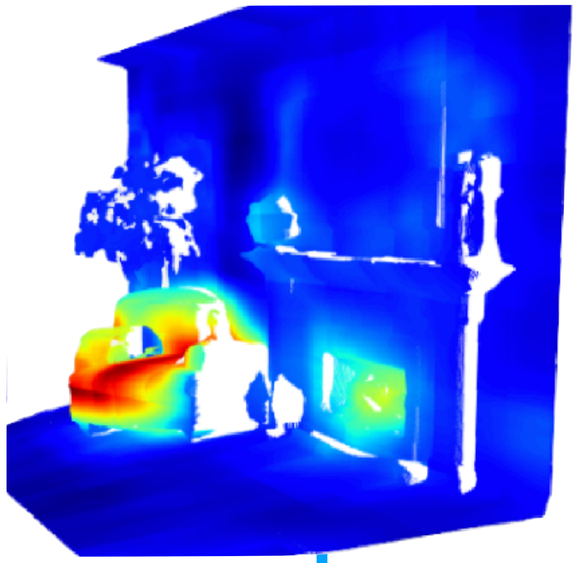
**Learn one semantic-agnostic concept of ``behind that object”**

# Semantic Abstraction

**Visual Semantics Concepts**

**Geometry Concepts**

**Spatial Concepts**



**Pixel-wise Relevancy Map:**
represents the probability of the object's locations.

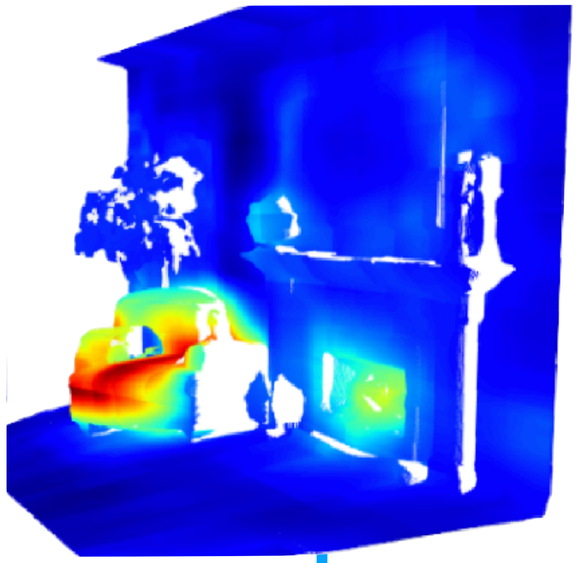
# Semantic Abstraction

## Visual Semantics Concepts

Text Input

Beige Arm Chair

RGB-D Input



**Frozen 2D VLM**
(Semantic-aware)

# Semantic Abstraction

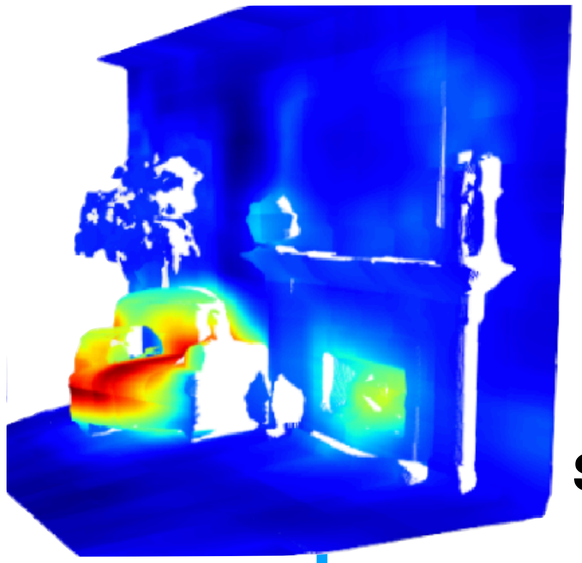
## Visual Semantics Concepts

Text Input

Beige Arm Chair

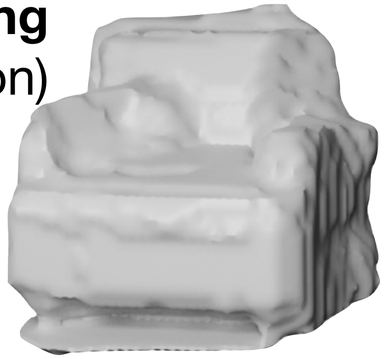
RGB-D Input

**Frozen 2D VLM**
(Semantic-aware)

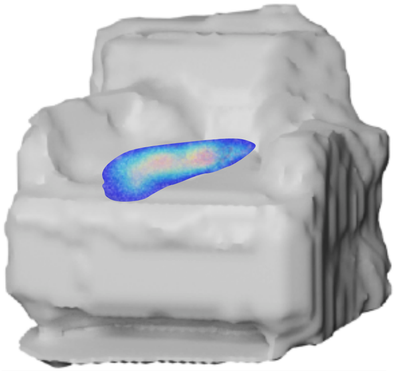


**Relevancy Map From 2D LVM**

**Geometry Reasoning**
(Shape Completion)



**Spatial Reasoning**
(Localization)



On-top-of

**Semantic-Agnostic**
Does not observe color or semantic label

# Relevancy Extraction

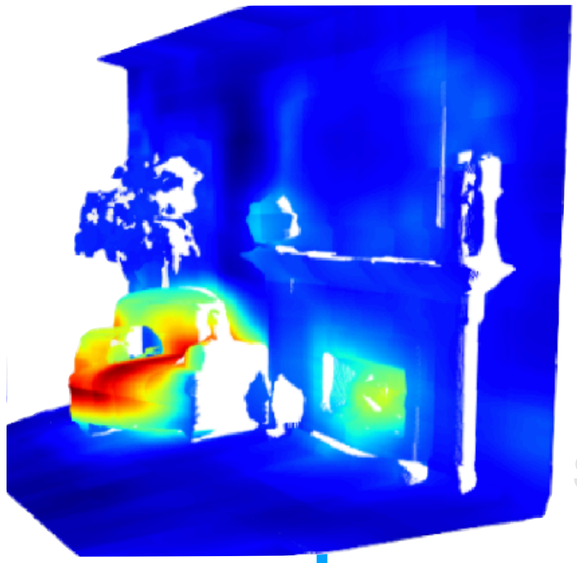
## Visual Semantics Concepts

Text Input

Beige Arm Chair

RGB-D Input

**Frozen 2D VLM**
(Semantic-aware)



**Relevancy Map From 2D LVM**

Geometry Reasoning
(Shape Completion)



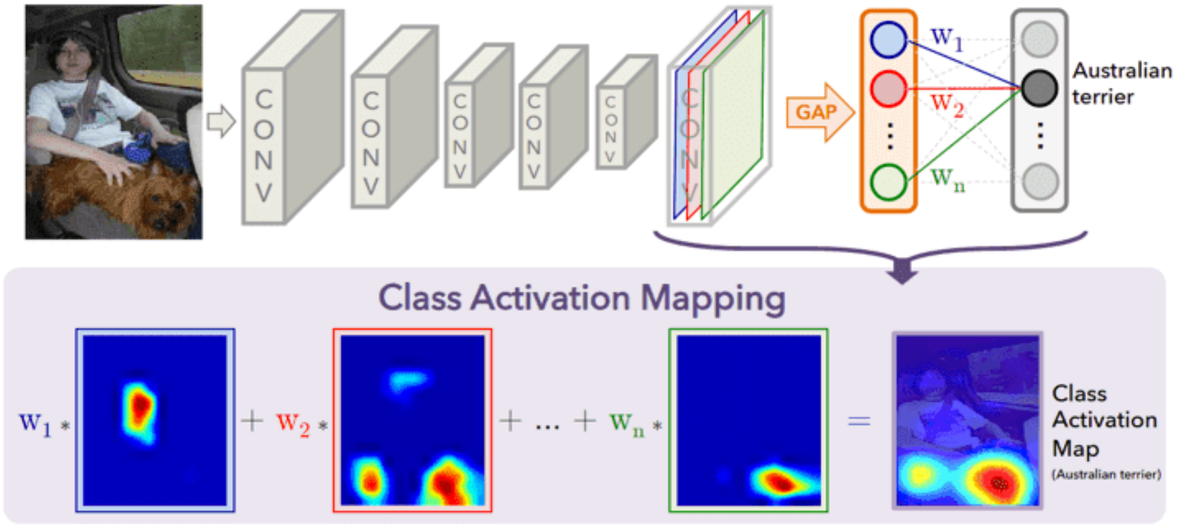
Spatial Reasoning
(Localization)

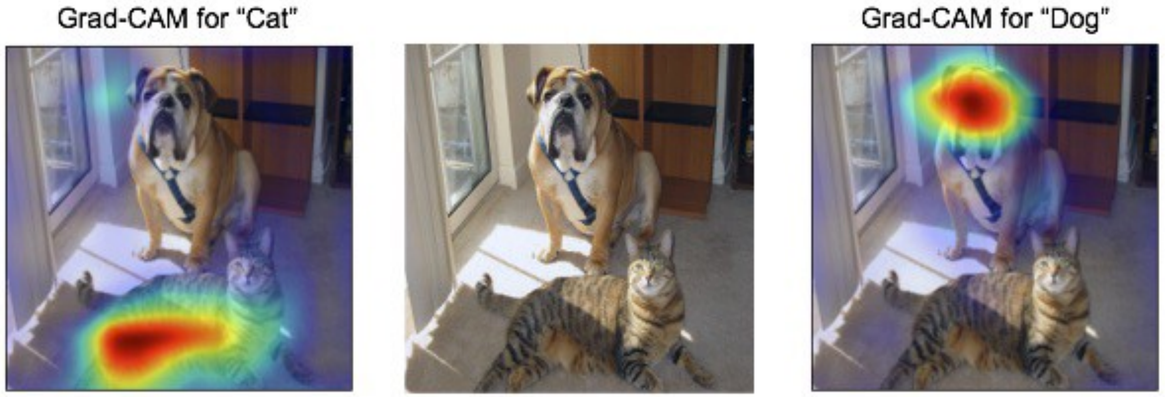


On-top-of

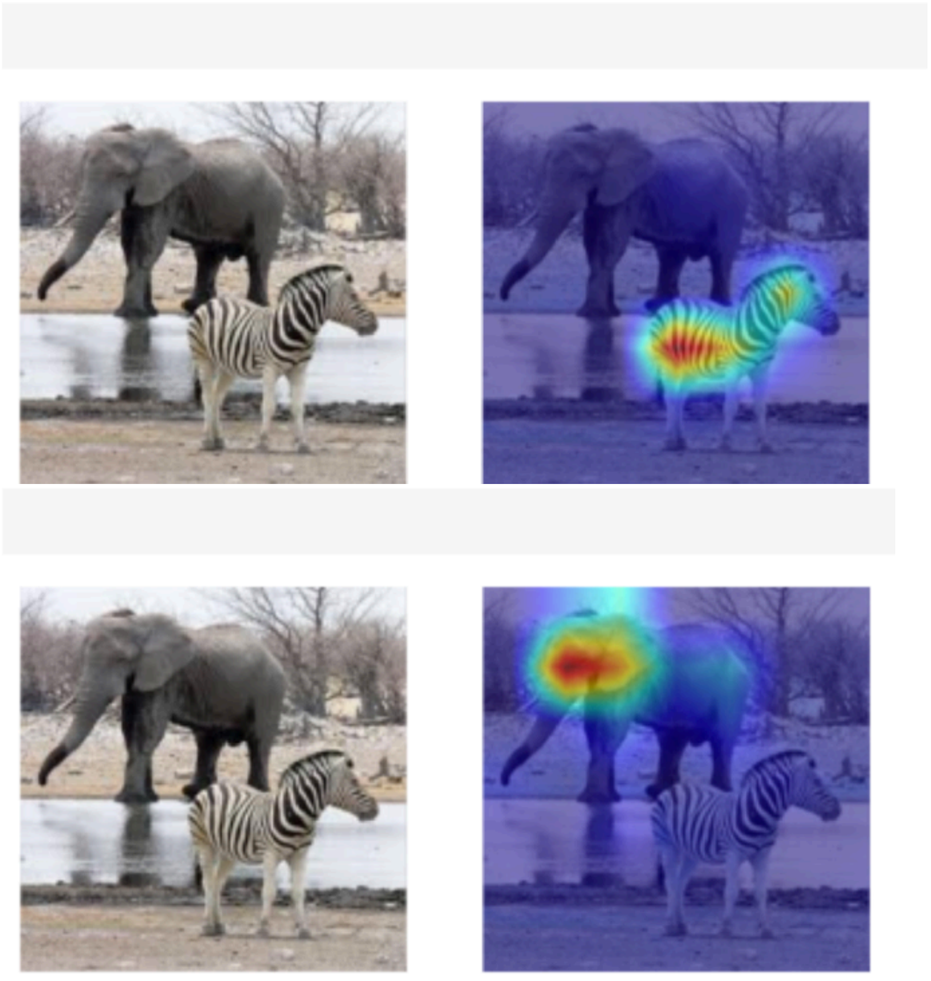
**Trained with limited 3D data**
(semantic-agnostic)

# Tools for Attention Extraction (GradCam)




**Grad-CAM.** Selvaraju et al.

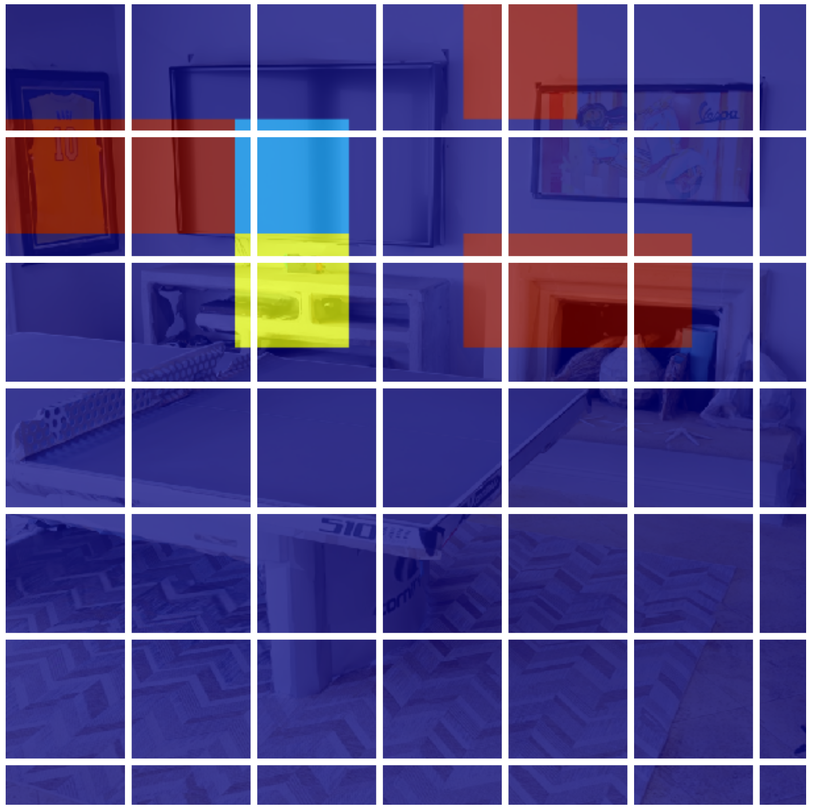


**Generic Attention-model Explainability.** Chefer et al.

# Noisy & Low-res relevancy maps

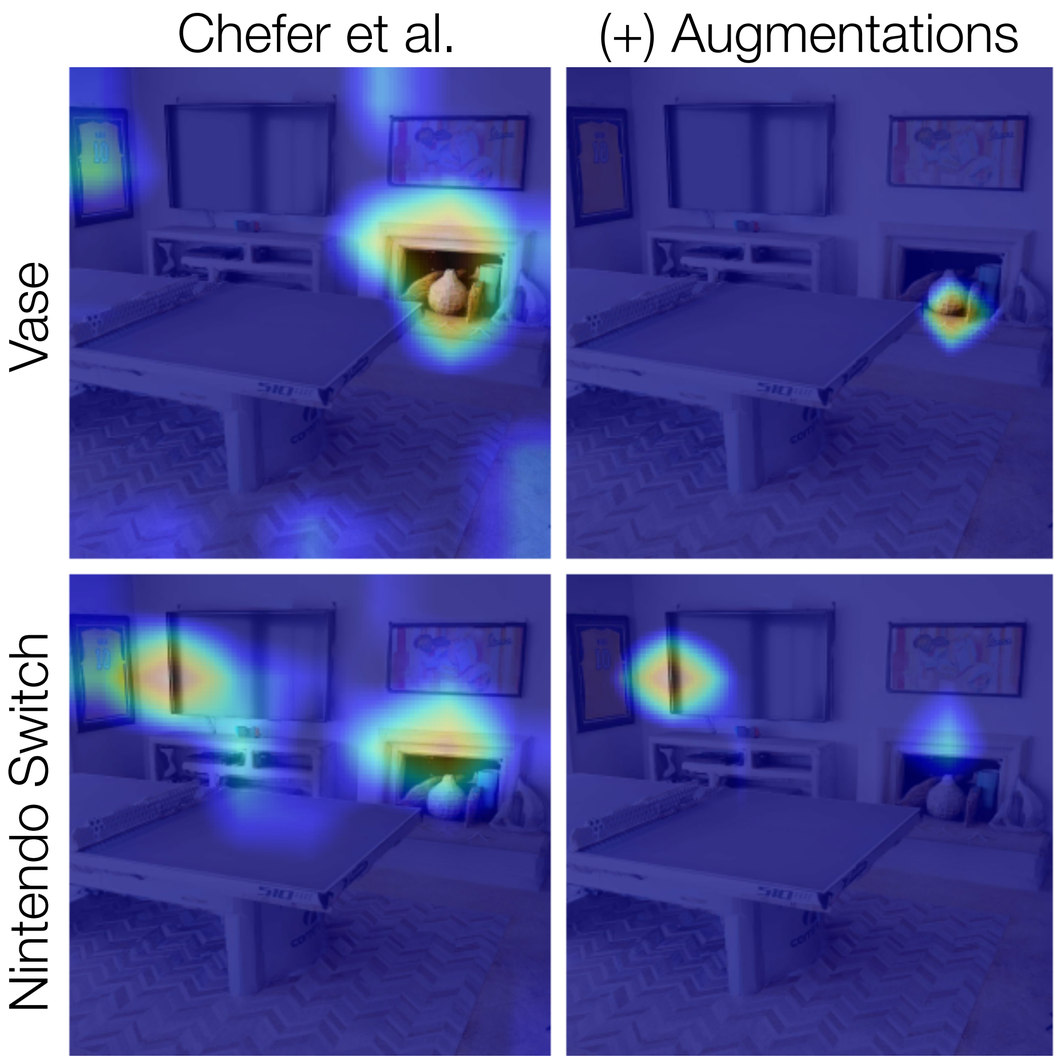
Nintendo Switch

Relevancy Extractor



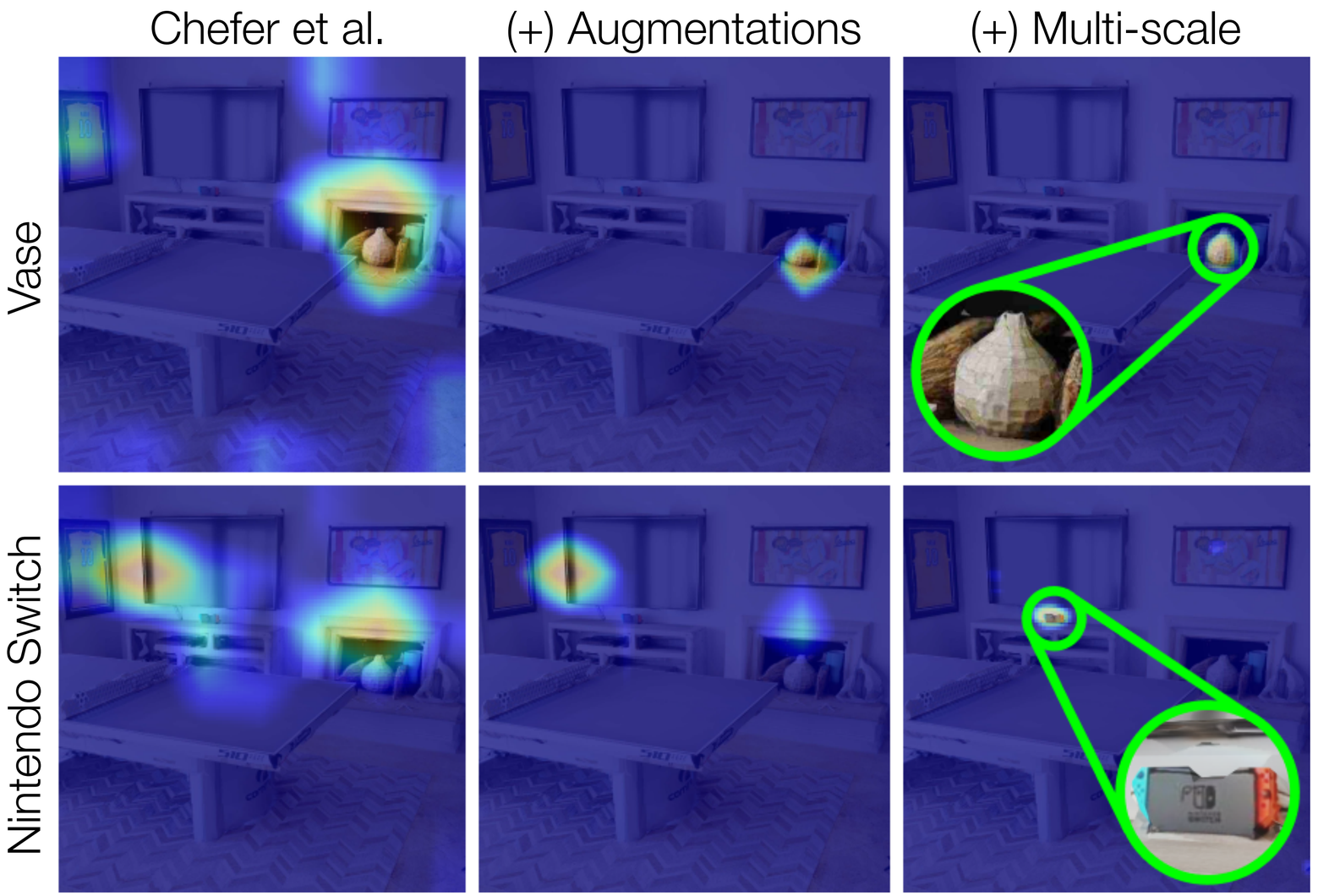
**Issues:** 1. Noisy & highlights irrelevant regions  2. Misses small objects

# Multi-scale Relevancy Extractor




1. Apply many image augmentations, and then average the resulting relevancy maps to reduce noise

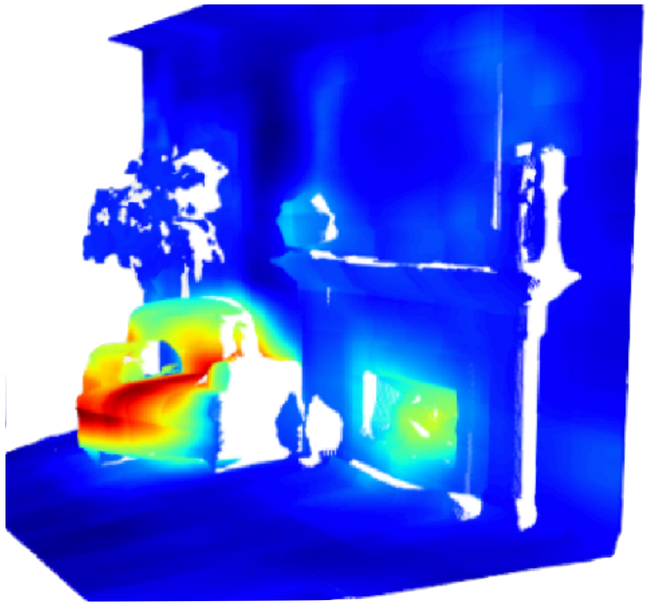
# Multi-scale Relevancy Extractor

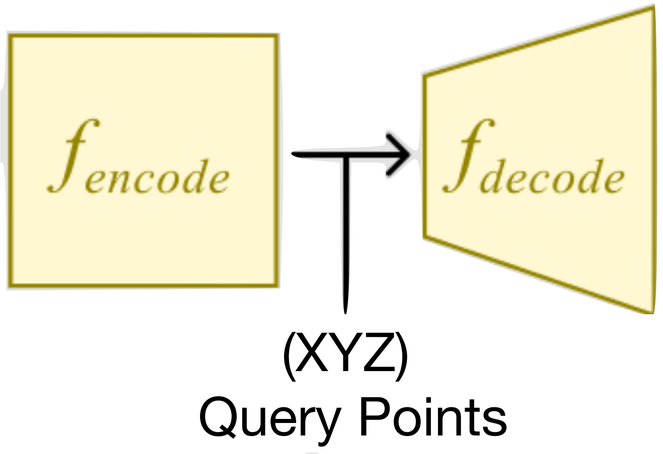



3) Extracting relevancy at a multiple scale.

# Semantic-Abstracted 3D Module

Relevancy value does not need to be perfect,
completion network is trained to correct small errors



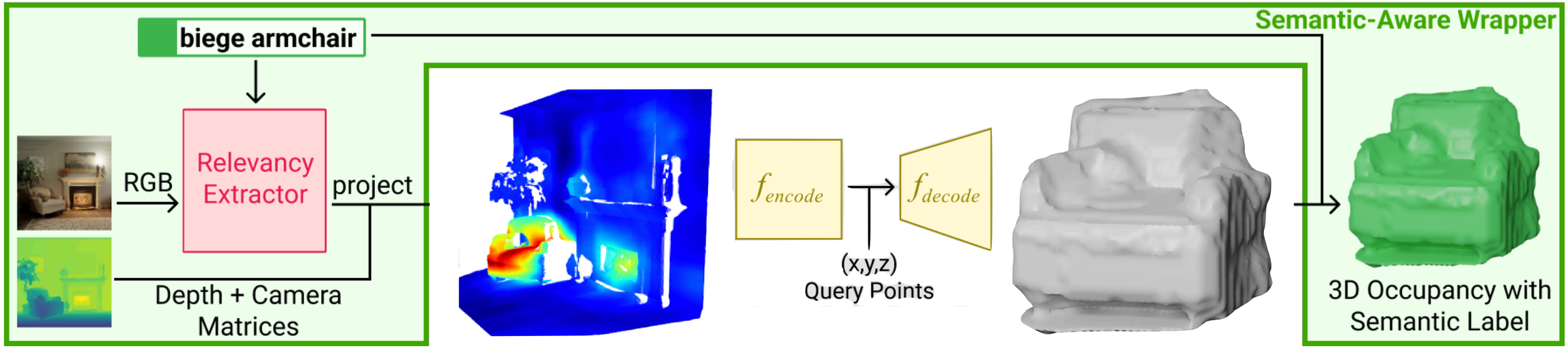



**Relevancy Map**

**3D Shape Completion**

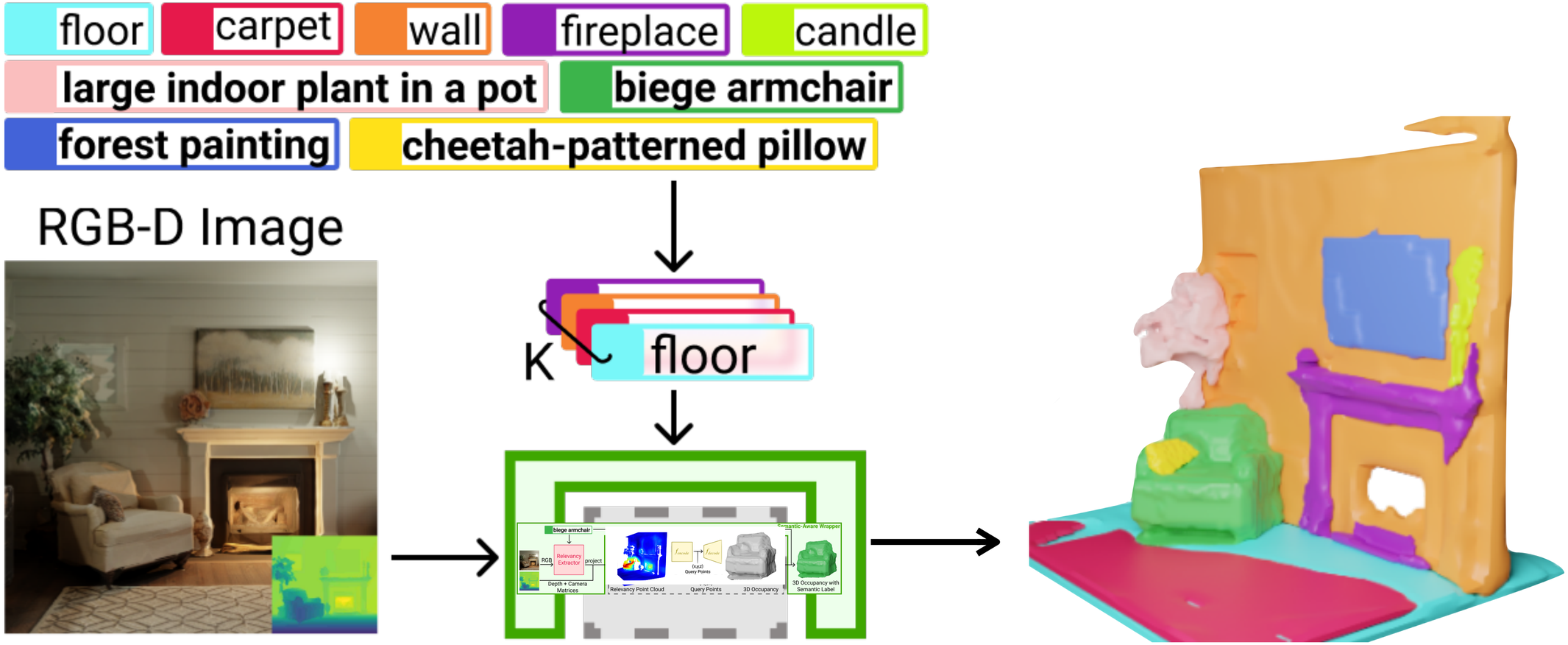
interpret it as a rough indicator to the 3D network about which object is referred and need to be complete in the scene.

# Semantic-Abstracted 3D Module




Together with the text description provided in the beginning,
we can get the completed shape for a particular object referenced in text.

# Open-Vocabulary Scene Completion




Note: If the input object is not visible or cannot be find the 2D VLM, the relevancy map produce low relevancy score for the whole image.

# Visually Obscured Object Localization

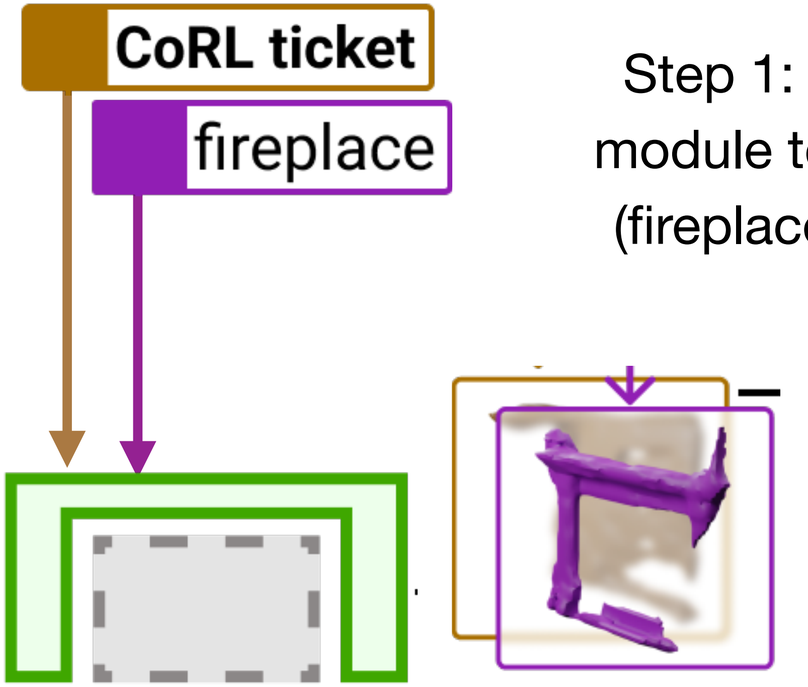
Provide additional spatial reference to help localize the object

# Visually Obscured Object Localization

Provide additional spatial reference to help localize the object

CoRL ticket on top of the fire place

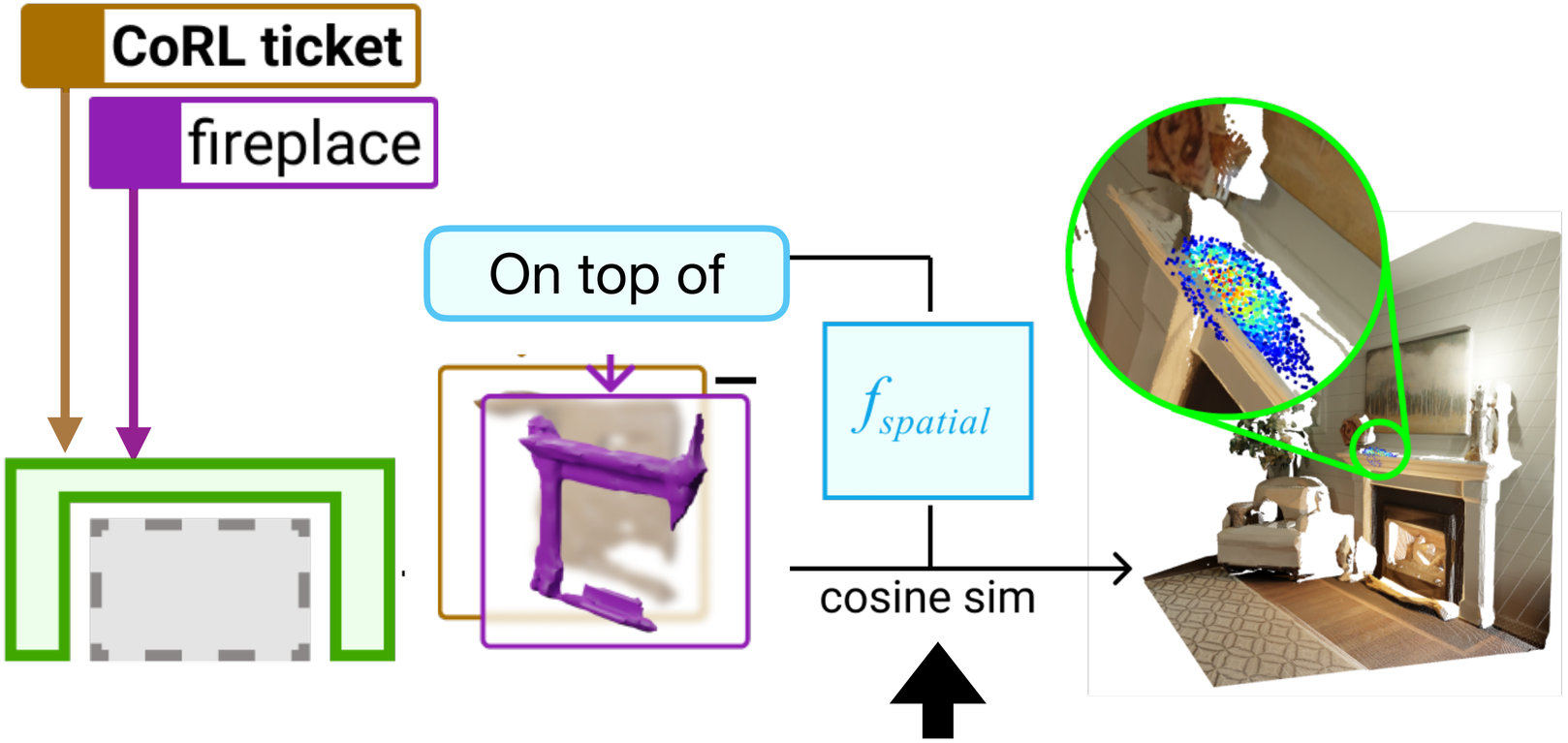
# Visually Obscured Object Localization




Step 1: Use the same shape completion module to find and complete the reference (fireplace) and target object (CoRL ticket)

# Visually Obscured Object Localization

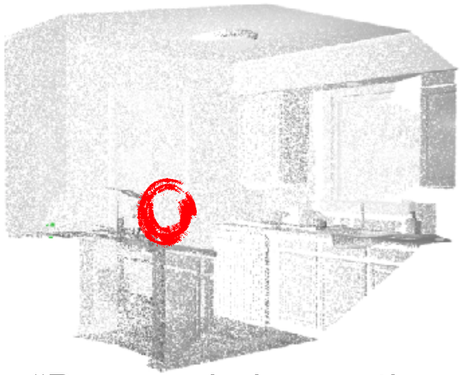
Step 2: Localize target object based on spatial relations (e.g., on top of)



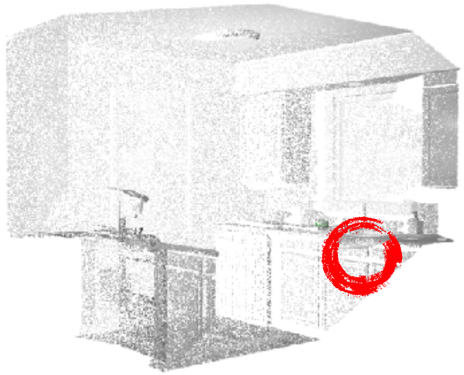

Spatial embedding is learned for each spatial relation (six in total)

# Groundtruth from Simulation (AI2-THOR)

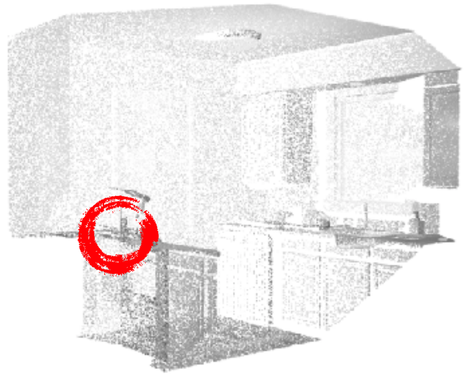
Dense 3D label for both the shape completion task and the spatial localization task.



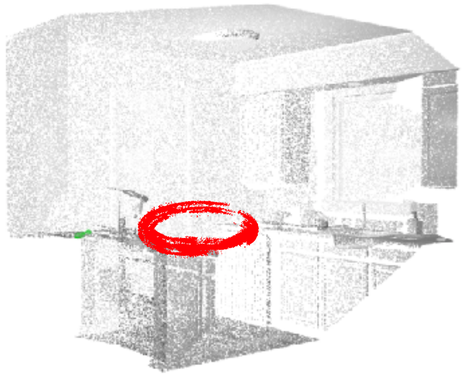
“Pepper shaker on the left of the burner”



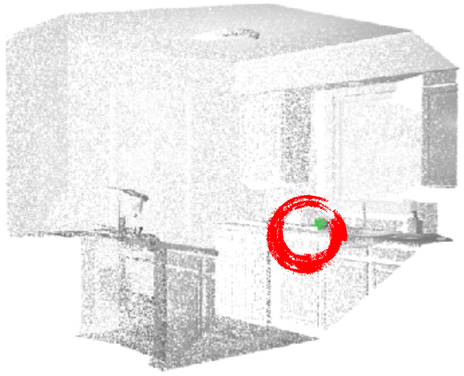
“Potato on the right of the lettuce”



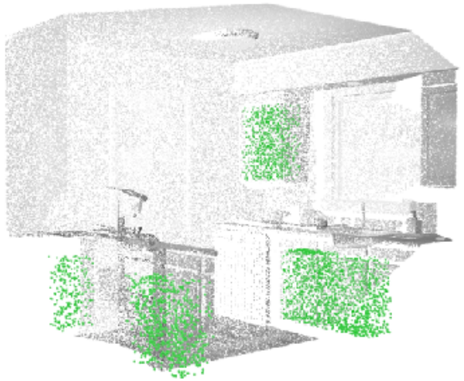
“Salt shaker in front of the pan”



“Knife behind the stove”



“Tomato on the sink”



“Banana in the cabinet”

**AI2-THOR: An Interactive 3D Environment for Visual AI**, Kolve et al.

# Open-World Evaluation

We hope to inherent the generality from the 2D VLMs

Open-world generalization

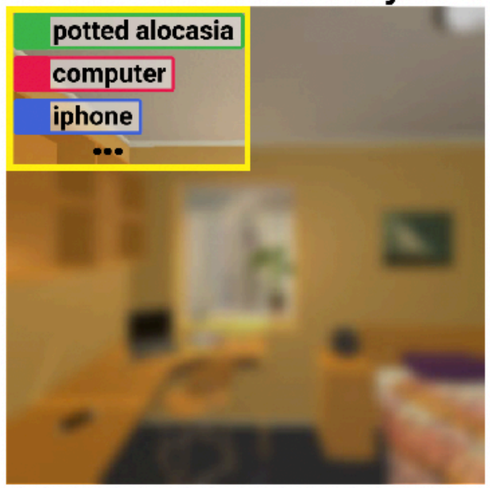
Visual Semantics

Spatial

Geometry


Novel Domain (Real)

Novel vocabulary




(+) OOD Nouns

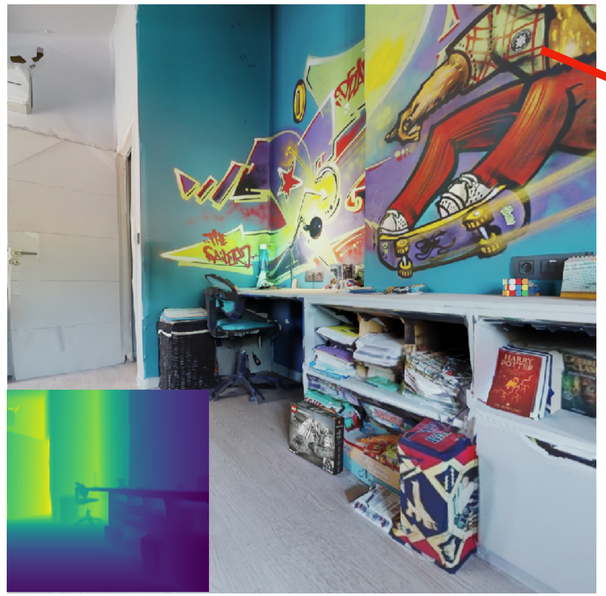
Novel semantic classes


(+) OOD 3D Geometry

# Qualitative Results on Matterport3D dataset

**ARKitScenes - A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data**, Baruch et al

Input RGB-D



OOD visual
(e.g., painted wall)

Scene Completion

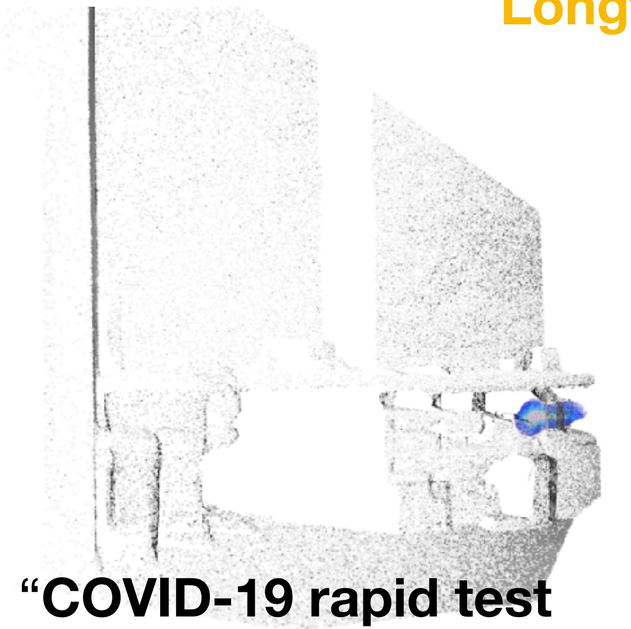
Small

Longtail

Semantic Labels

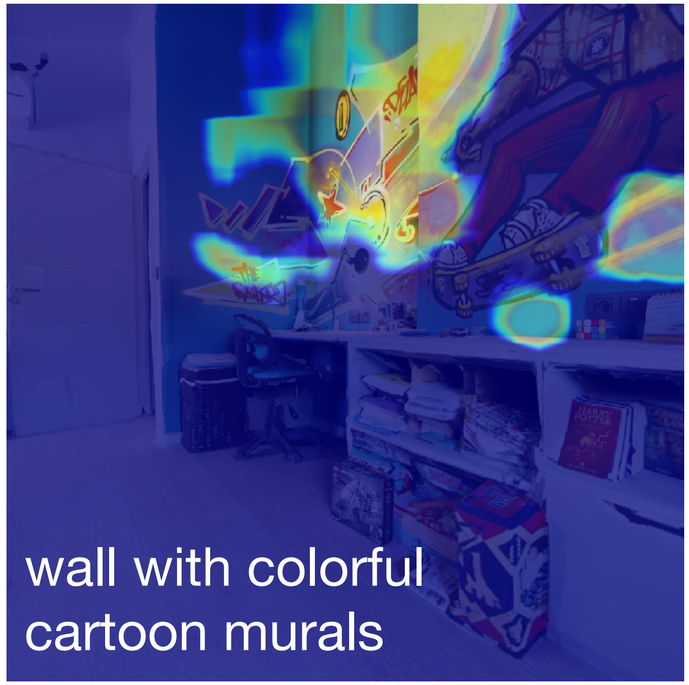
- office chair
- floor
- desk
- trash basket
- book shelf
- **rubiks cube**
- **harry potter book**
- **lego technic excavator**
- **wall with colorful cartoon murals**

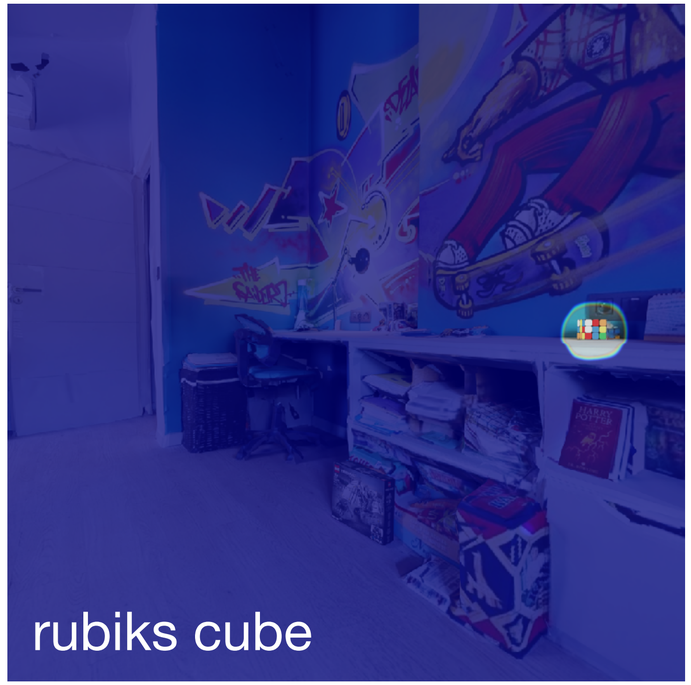
“**used N95s**
in the garbage bin”

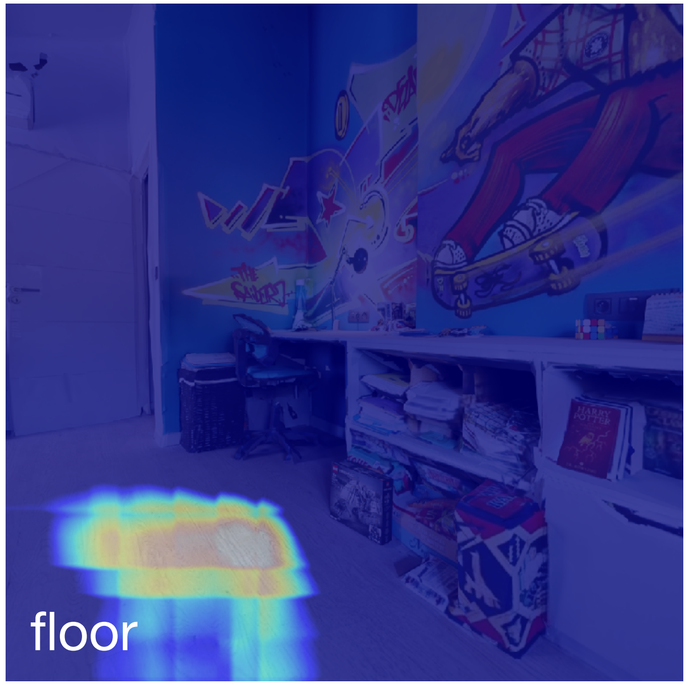


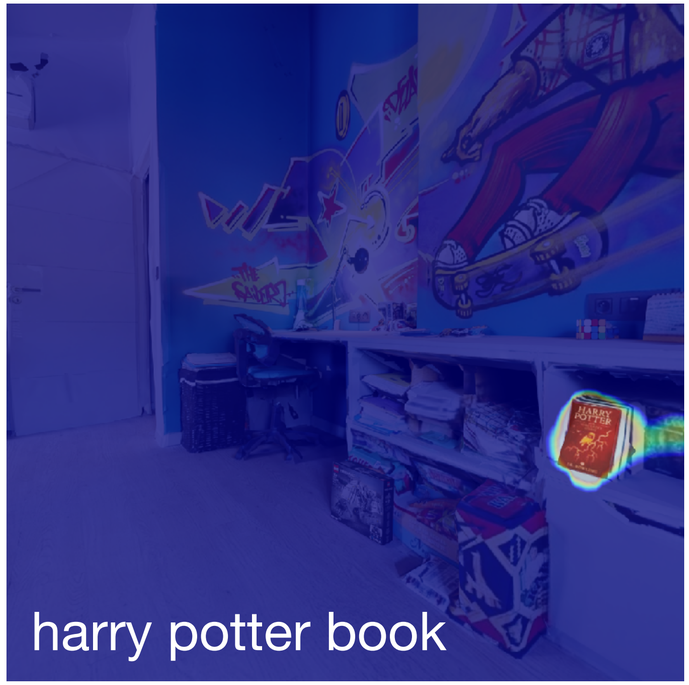
“**COVID-19 rapid test**
behind **Harry Potter book**”

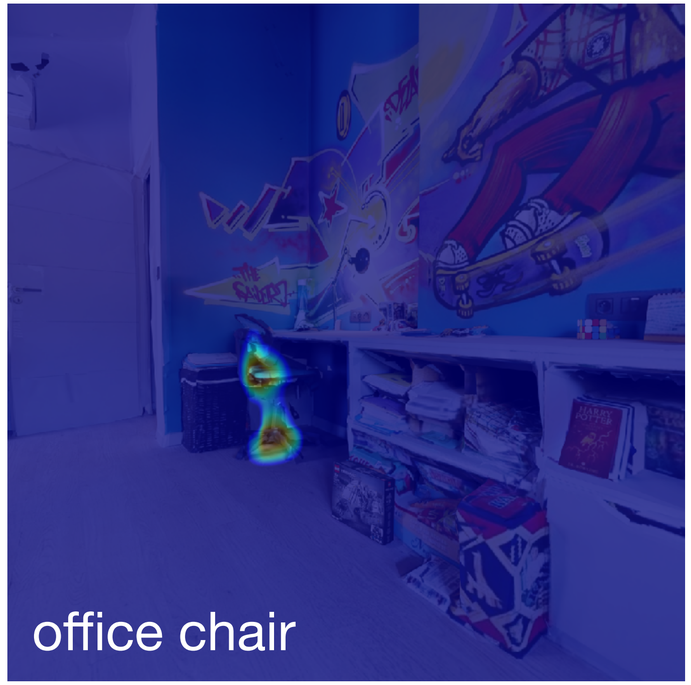

COVID-19 rapid test


wall with colorful
cartoon murals


rubiks cube


floor


harry potter book

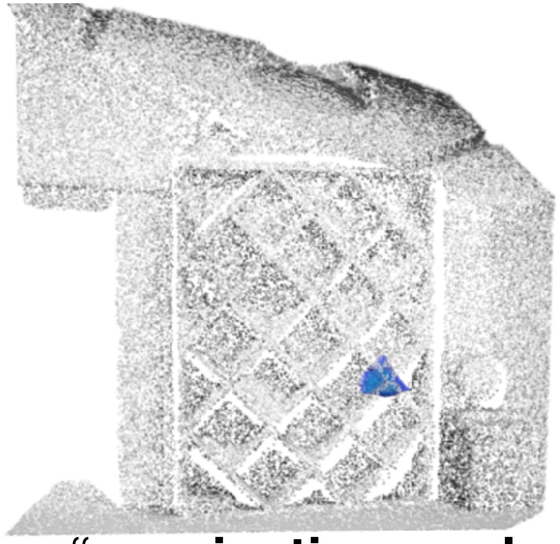

office chair

Input RGB-D

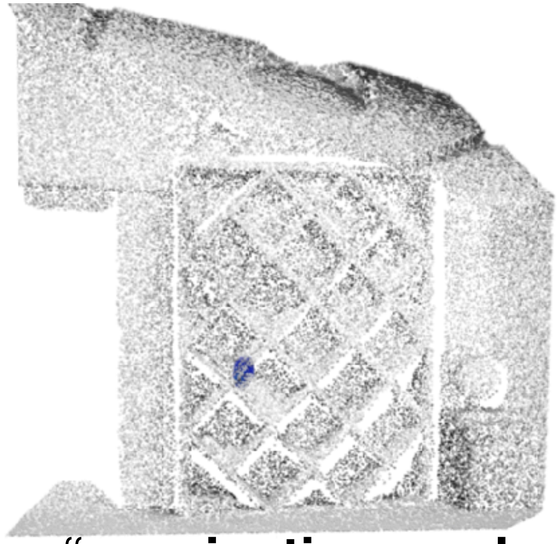
Scene Completion

Semantic Labels

- floor
- wall
- ceiling
- light
- book shelf
- house plant
- **globe**
- **black aluminum file cabinet**

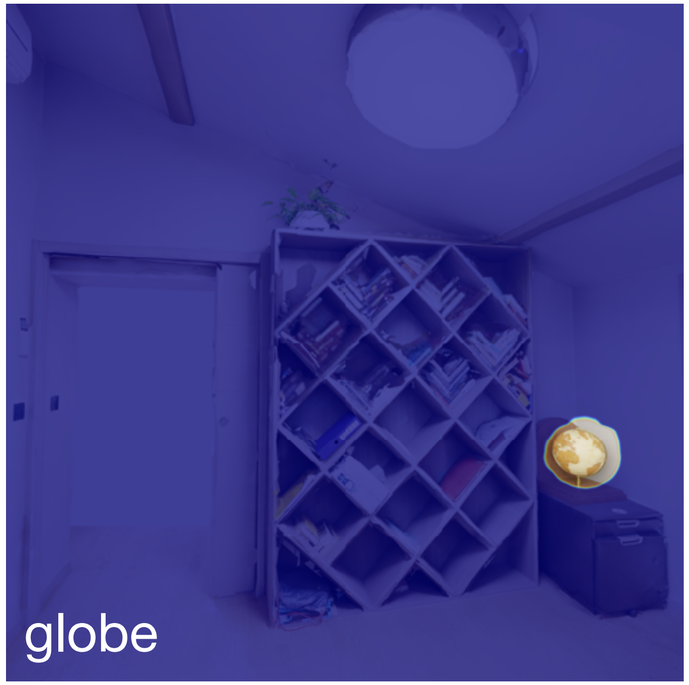


“**vaccination card** in the **red folder**”



“**vaccination card** in the **blue folder**”


house plant


globe


Low relevancy value
vaccination card


red folder


blue folder


book shelf

# Qualitative Results on Apple ARKitScenes dataset

**Different environments, Different camera …**
**Direct test without fine-tuning**

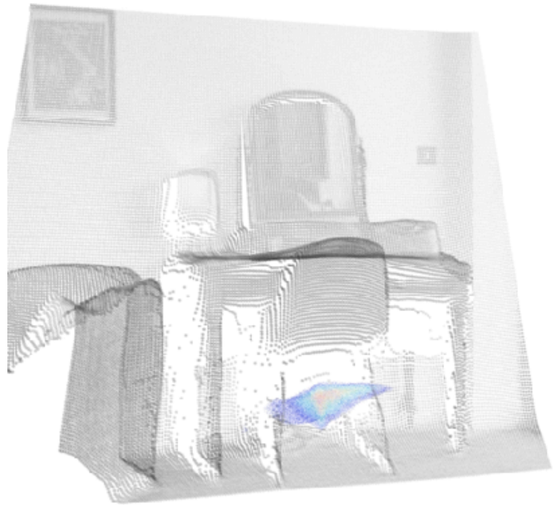
**ARKitScenes - A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data**, Baruch et al

Input RGB-D

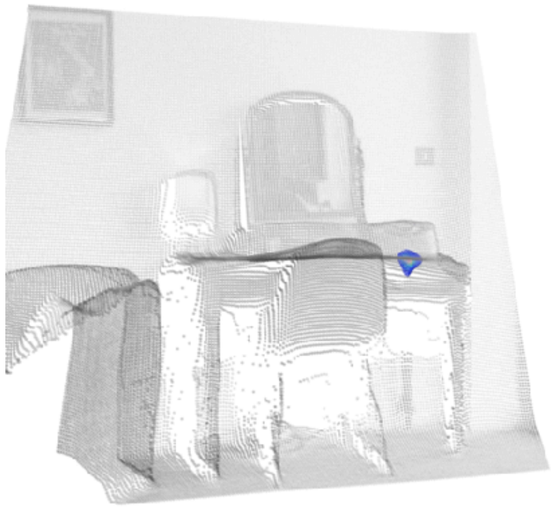
Scene Completion

Semantic Labels

- lamp
- wall
- table
- light switch
- mirror
- carpet
- **pink make up bag**
- **upholstered chair in faux leather**
- **woven chair**



**“Hair dryer with its wires tangled** behind **chair legs”**

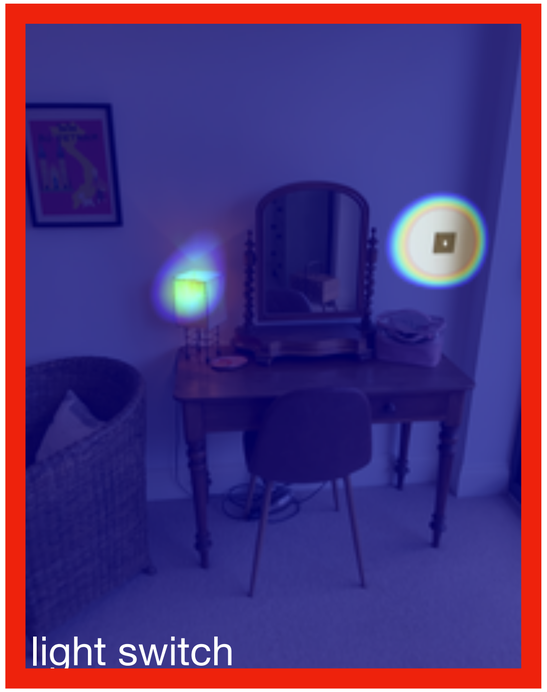


**“sunscreen bottle** in **pink make up bag”**

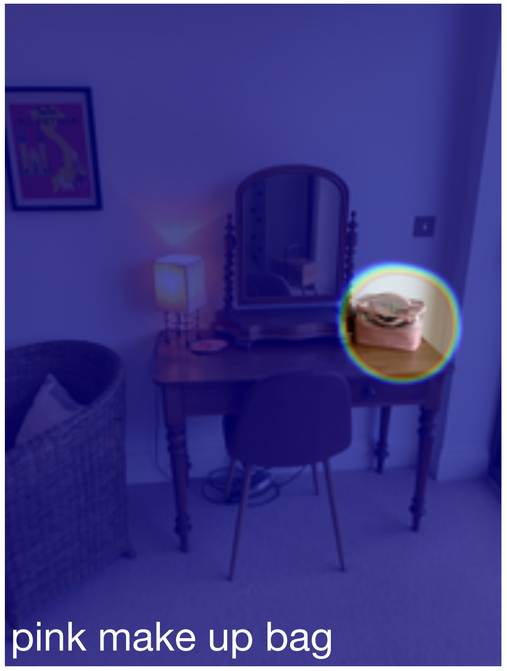
**Can handle long descriptions**


upholstered chair in faux leather


woven chair


light switch


mirror

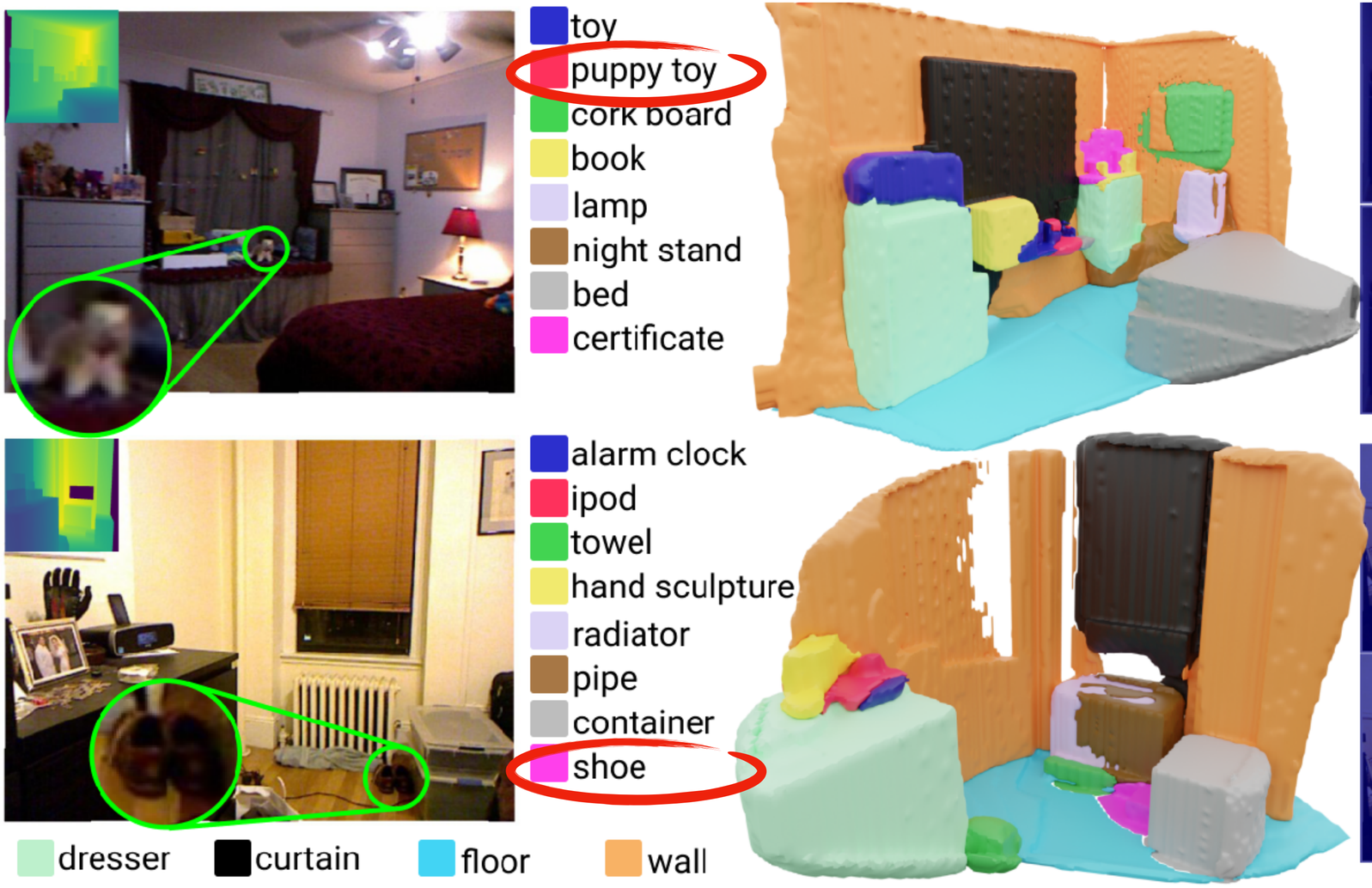

pink make up bag


table

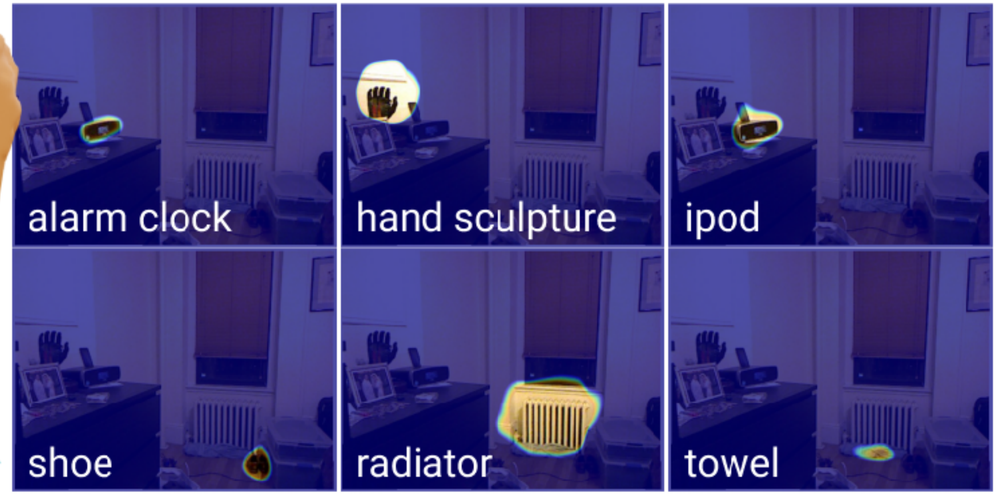

lamp


carpet

# Results on NYUv2 for all 894 classes

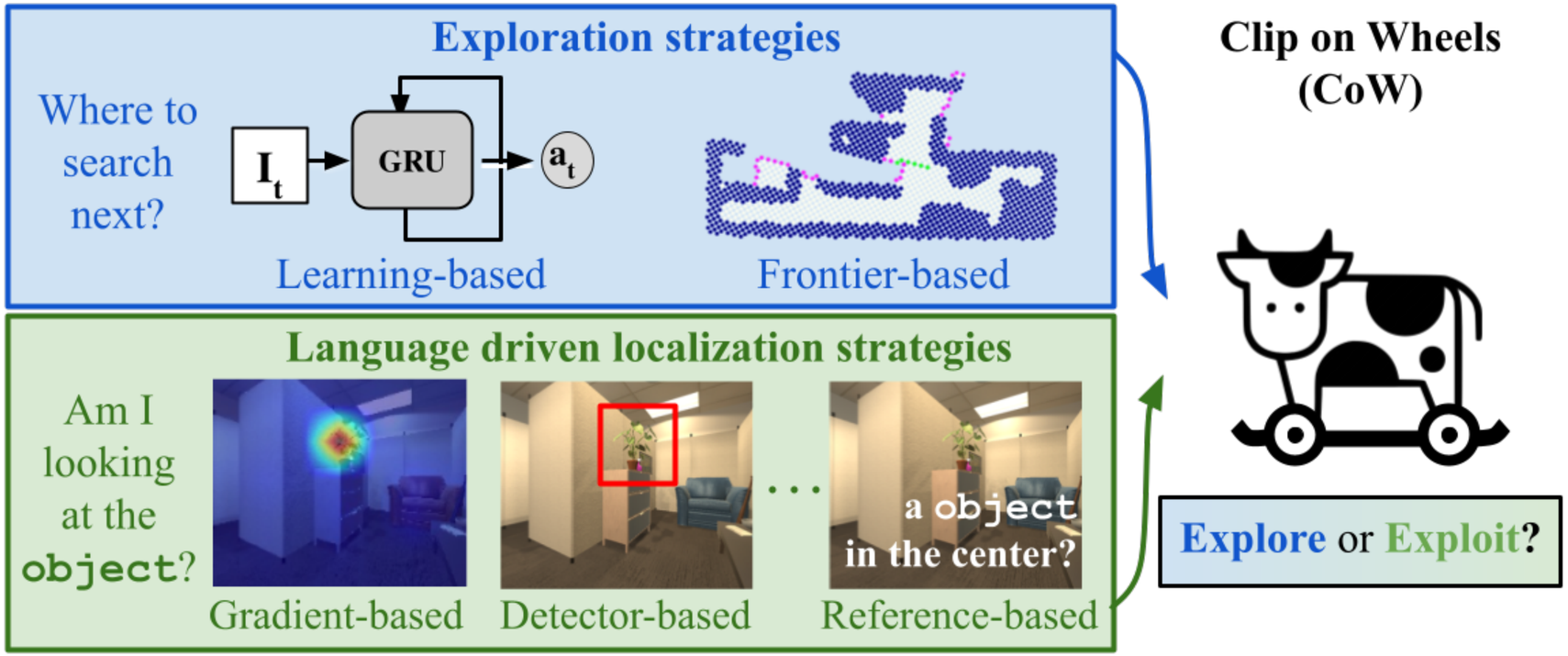






# Application: CLIP on Wheels

Language Driven Zero-Shot Object Navigation




The **modularized design** (i.e., localization & exploration) allow us to directly apply VLMs to object navigation **without additional training on navigation**.

# Evaluating CoWs on Pasture Benchmark

~20 different variant of CoWs

Study capabilities that closed-vocabulary object navigation agents do not possess

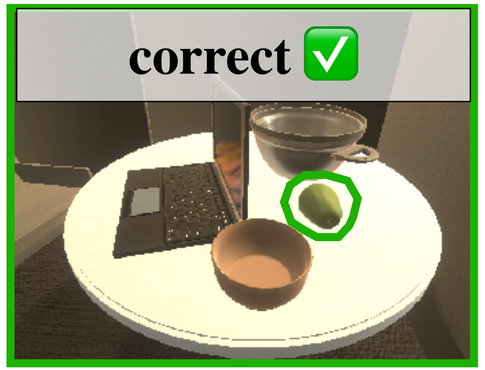
# Evaluating CoWs on Pasture Benchmark

## 1) Find uncommon objects

llama wicker basket


## 3) Finding object based on attributes

"…small, green apple…" (appearance)
"…apple on a coffee table near a laptop…" (spatial)


## 4) Hidden objects

"mug under the bed"


## 4) Zero-shot to different domains


# Evaluating CoWs on Pasture Benchmark

1) Find uncommon objects

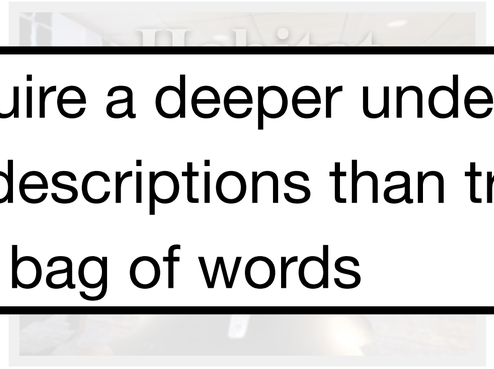
llama wicker basket

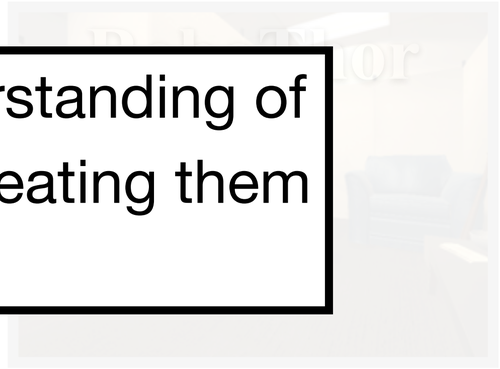

3) Finding object based on attributes

"…small, green apple…" (appearance)
"…apple on a coffee table near a laptop…" (spatial)



4) Zero-shot to different domains





**4) Hidden objects**

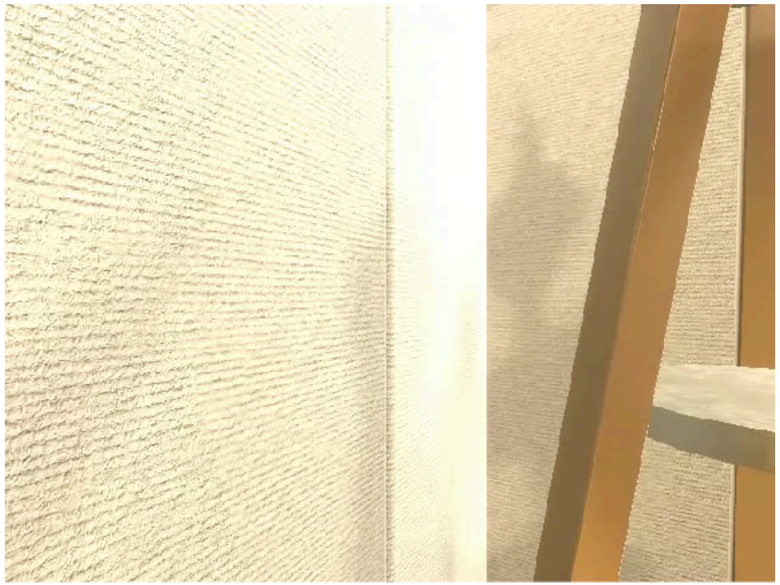
"mug under the bed"


Require a deeper understanding of the descriptions than treating them as a bag of words

# Evaluating CoWs on Pasture Benchmark

| ID | CoW breeds Loc. | Arch. | Post | 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-ZS [37] | | | – | – | – | 8.1 | – | 14.0* | 60M |

**Zero-shot CoW (no training on navigation task) is better than fine-tuned SOTA model!**

**CoWs on PASTURE: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation.** Samir Gadre et al

# Abstraction as Inductive Bias

# Abstraction as Inductive Bias

**Advantage:**

- Learn from less data
- Better generalization

**Semantic Abstraction**

# Abstraction as Inductive Bias

**Advantage:**

- Learn from less data
- Better generalization

**Semantic Abstraction**

**Address limitation:**

- Adapt to different tasks with different reusable modules
  - e.g., shape completion and  localization
- Learnable downstream model to reduce error propagation

# Abstraction as Inductive Bias

In the age large pre-trained models …

**PaLM**



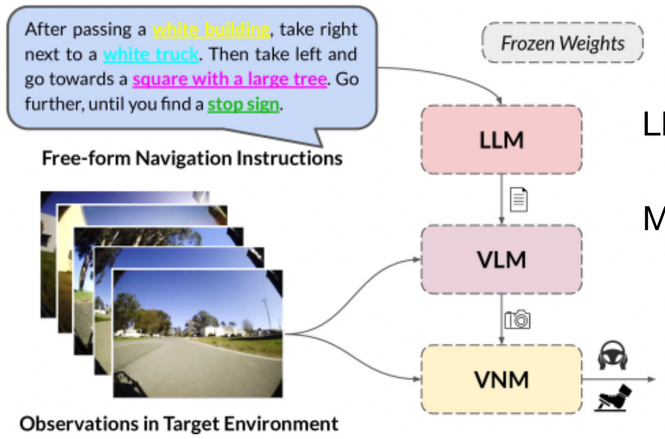
**CLIP**


…

- Large models learned from Internet knowledge
- Fine-tuning them for specific application can be impractical (e.g., compute) or challenging (e.g., hurting robustness)
- The right abstraction allow us to extract the relevant knowledge from them in a zero-shot manner (w.o. fine-tuning)

# Abstraction as Inductive Bias

## *Extend New skills …*



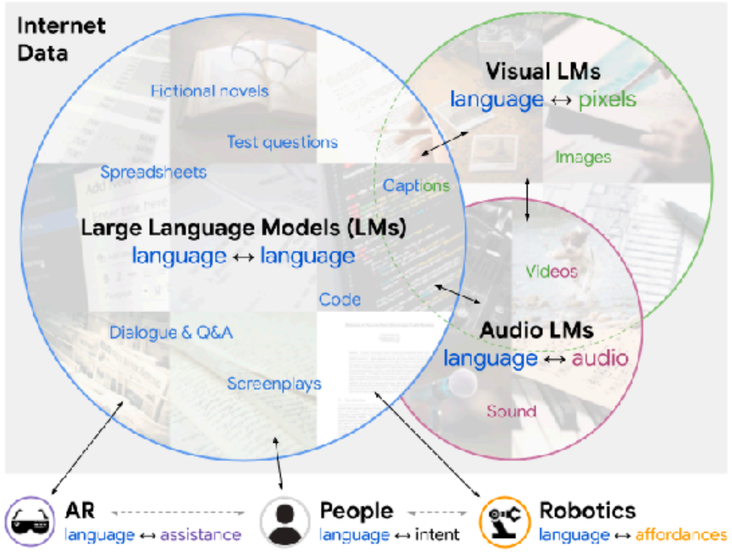

### Navigation

LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action


### Manipulation

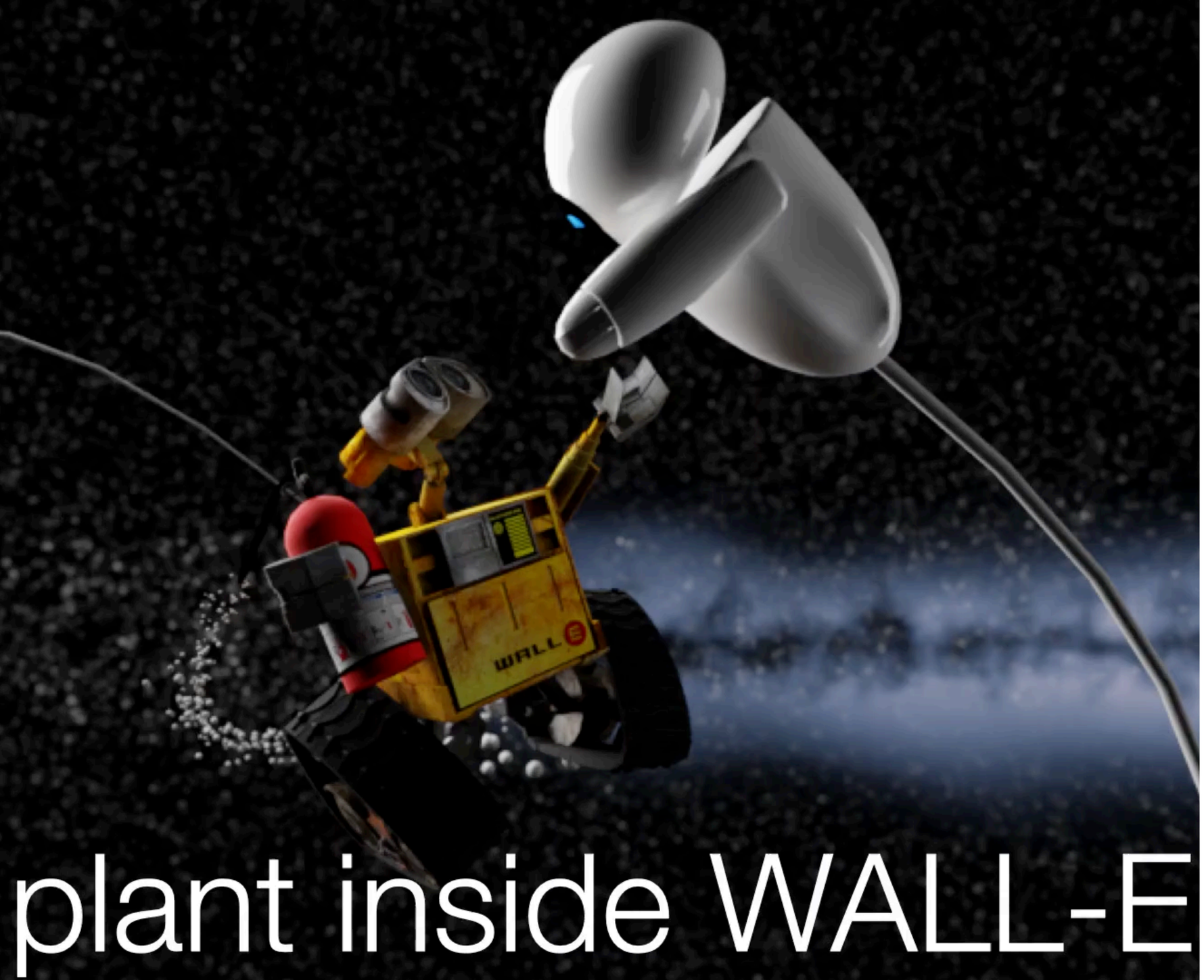
Do As I Can, Not As I Say: Grounding Language in Robotic Affordances

## *Combine Different Expertises …*




Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language

***What's Next?*** *Your Work!*




**Thank You!**
**Question?**