Discovering Object Attributes by Prompting Large Language Models with Perception-Action APIs

Angelos Mavrogiannis, Yiannis Aloimonos

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Motivation

- There has been a lot of interest in grounding natural language to physical entities through visual context [1].
- Vision Language Models (VLMs) can ground linguistic instructions to visual sensory information [2].
- However, VLMs struggle with grounding non-visual attributes, like the weight of an object [3, 4].

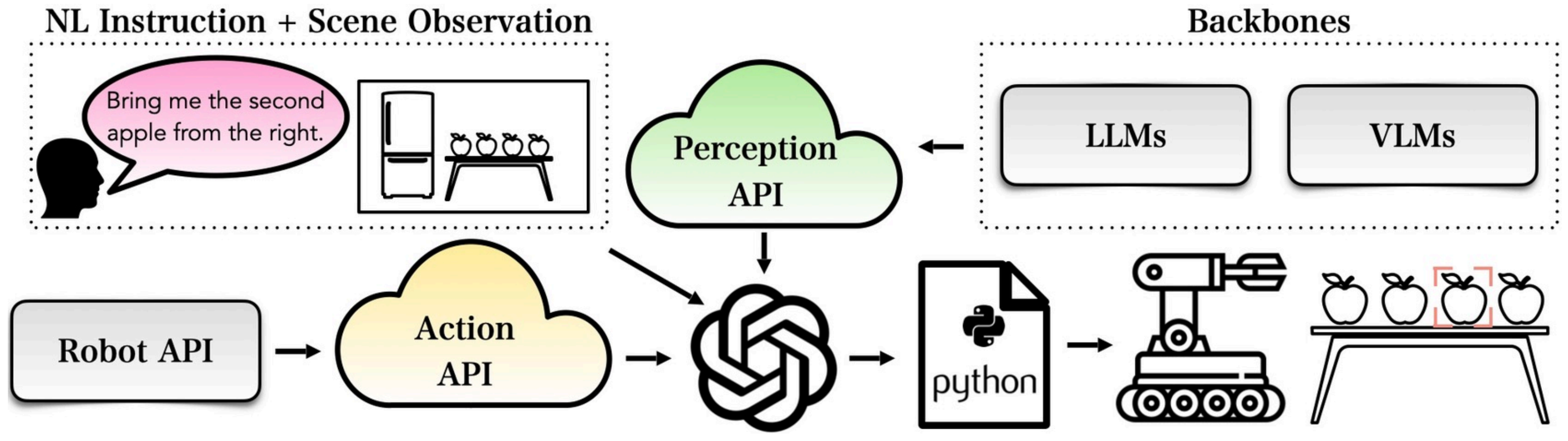
Key Insight

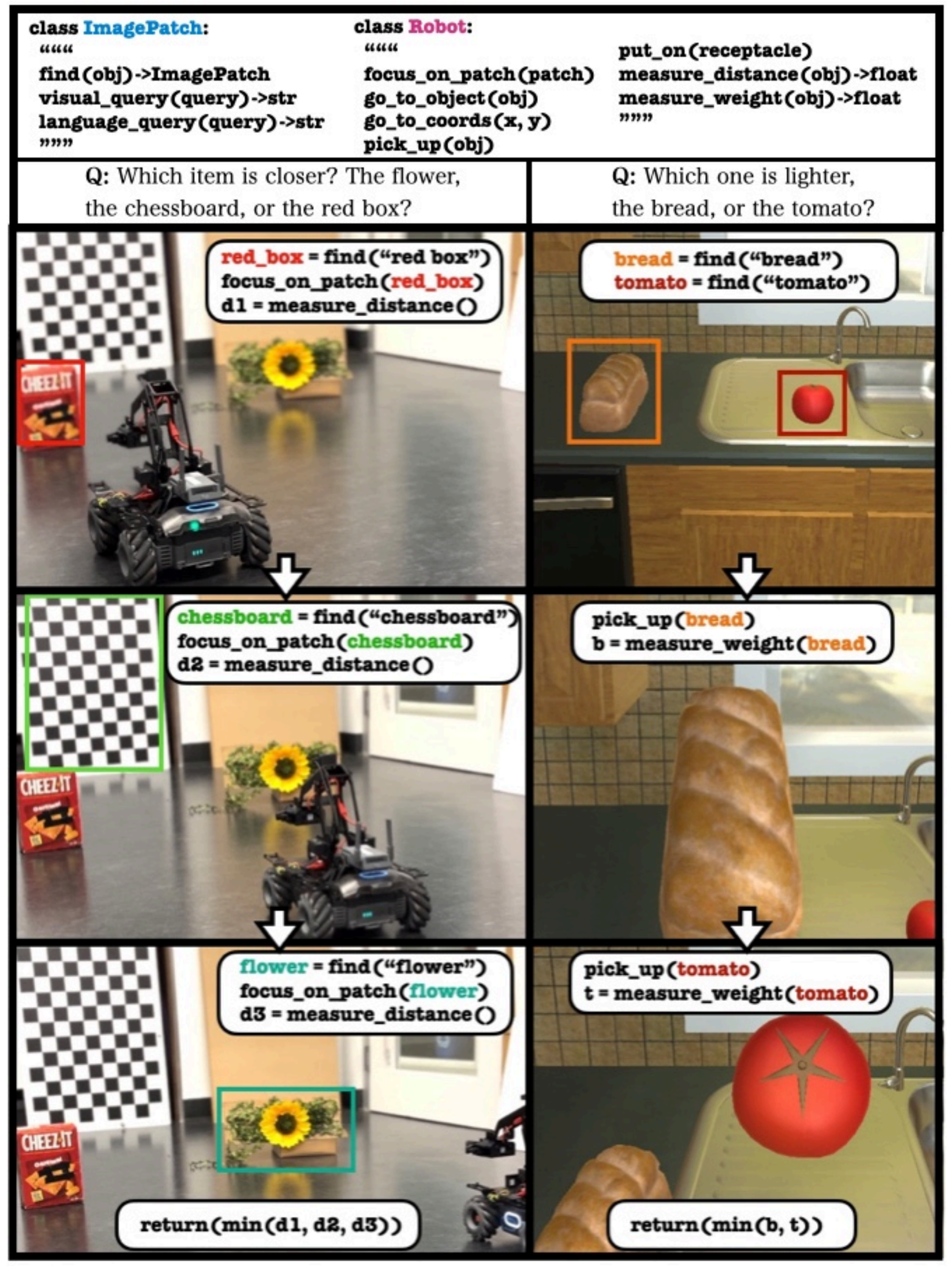
Non-visual attribute detection can be effectively achieved by **active perception** guided by **visual reasoning**.

Approach

We present a **Perception^5-Action** API that consists of VLMs and LLMs as backbones, together with a set of robot control functions. When prompted with this API and a natural language query, an LLM generates a program to actively identify attributes given an input image.

Architecture


NL Instruction + Scene Observation
Bring me the second apple from the right.
Backbones
LLMs
VLMs
Perception API
Robot API
Action API
python



```
class ImagePatch:
 """"
 find(obj)->ImagePatch
 visual_query(query)->str
 language_query(query)->str
 """"

class Robot:
 """"
 focus_on_patch(patch)
 go_to_object(obj)
 go_to_coords(x, y)
 pick_up(obj)
 put_on(receptacle)
 measure_distance(obj)->float
 measure_weight(obj)->float
 """"
```

Q: Which item is closer? The flower, the chessboard, or the red box?

```
red_box = find("red box")
focus_on_patch(red_box)
d1 = measure_distance()
```

```
chessboard = find("chessboard")
focus_on_patch(chessboard)
d2 = measure_distance()
```

```
flower = find("flower")
focus_on_patch(flower)
d3 = measure_distance()
```

```
return(min(d1, d2, d3))
```

Q: Which one is lighter, the bread, or the tomato?

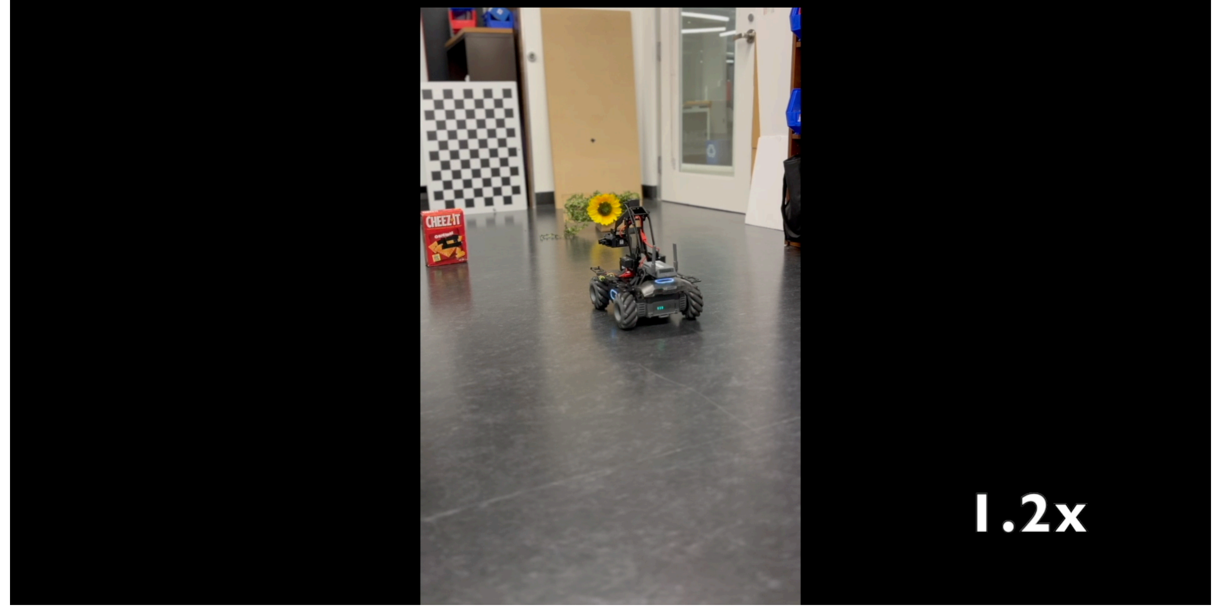
```
bread = find("bread")
tomato = find("tomato")
```

```
pick_up(bread)
b = measure_weight(bread)
```

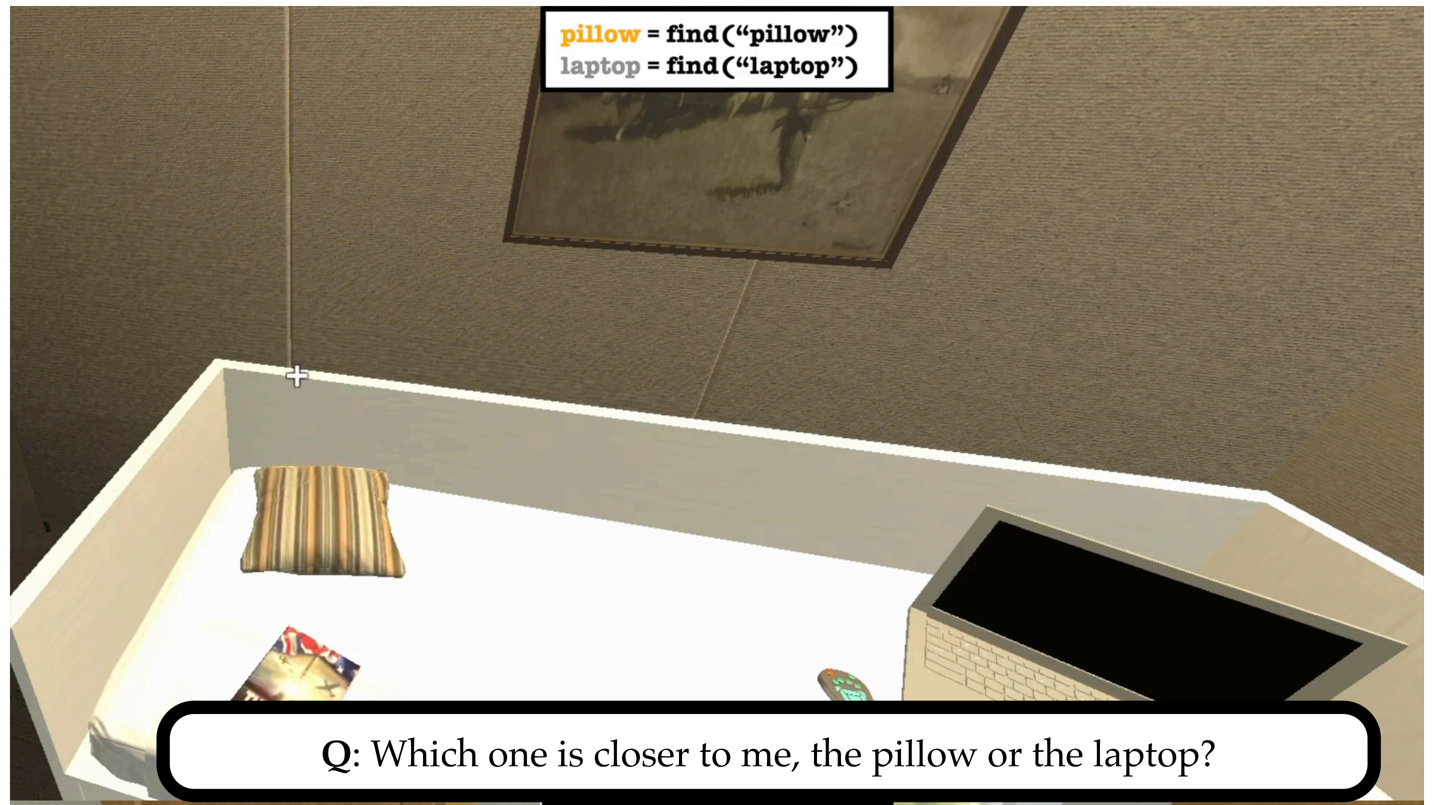
```
pick_up(tomato)
t = measure_weight(tomato)
```

```
return(min(b, t))
```

Q: Which one is closer, the flower, the chessboard, or the red box?

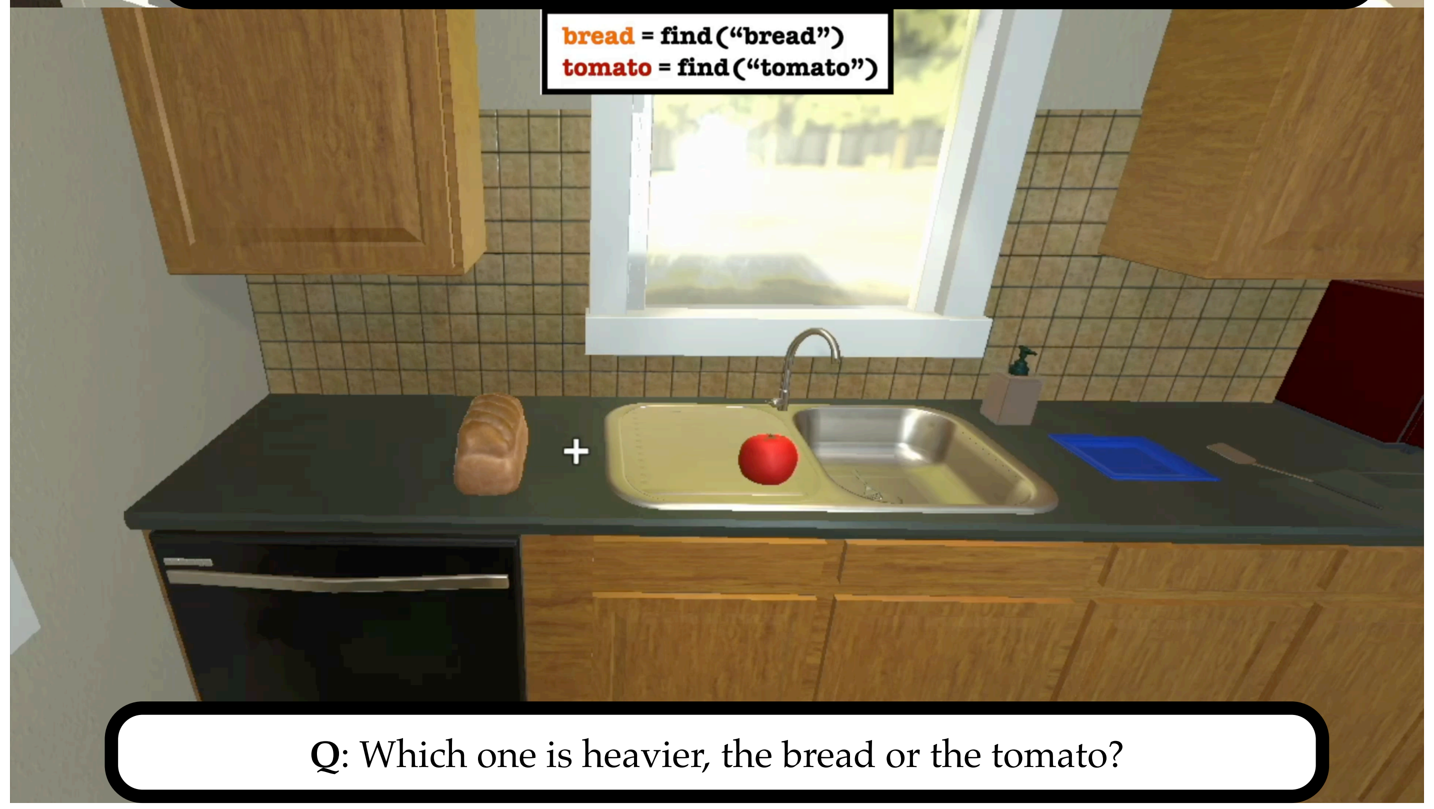


Evaluation on AI2-THOR



```
pillow = find("pillow")
laptop = find("laptop")
```

Q: Which one is closer to me, the pillow or the laptop?



```
bread = find("bread")
tomato = find("tomato")
```

Q: Which one is heavier, the bread or the tomato?

Method	Task	
	Weight	Distance
OVD (GLIP)	0.14	0.64
VQA (BLIP-2)	0.64	0.56
Attribute Detection API	0.90	0.22
GPT-4o	0.88	0.70
Perception-Action API	**0.96**	**0.94**

References

[1] Ichter et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. CoRL 2023

[2] Huang et al. Instruct2Act: Mapping Multi-modality Instructions to Robotic actions with Large Language Model. arXiv 2023

[3] Yi et al. NEWTON: Are Large Language Models Capable of Physical Reasoning? EMNLP 2023

[4] Gao et al. Physically Grounded Vision-Language Models for Robotic Manipulation. ICRA 2024

[5] Surís et al. ViperGPT: Visual Inference via Python Execution for Reasoning. ICCV 2023