ACT, INTERACT, AND FINETUNE WITH VISION-LANGUAGE MODELS

Zsolt Kira
Assistant Professor
School of Interactive Computing
Georgia Tech
What are some challenges in robotics?

Sense, Plan, Act

- **Sense:** Sensor data to observations
- **Plan:** Making plans to achieve complex goals
  - **Related:** World knowledge
- **Act:** From plans to low-level actions
EXPLOSION OF PROGRESS

Shameless Plug: https://github.com/GT-RIPL/Awesome-LLM-Robotics

RT-1/2/X  ProgPrompt  Code-As-Policies

Instruction Relevance with LLMs

Combined  Task Affordances with Value Functions

How would you put an apple on the table?

Find an apple 0.6
Find a coke 0.6
Pick up the apple 0.2
Pick up the coke 0.2
Place the apple 0.1
Place the coke 0.1
Go to the table 0.8
Go to the counter 0.8

I would: 1. Find an apple, 2. ...

Shameless Plug: https://github.com/GT-RIPL/Awesome-LLM-Robotics
WHAT DO LLMS & VLMS PROGRESS?

- **Sense**: Sensor data to observations  ✓ (+ Embodied Foundation Models)

- **Plan**: Making plans to achieve complex goals  ✓
  - **Related**: World knowledge  ✓

- **Act**: From plans to low-level actions  ?
But what about:

**Robustness and Generalization**: Handling in-the-wild objects, environments, and task/plan space

**Solutions**:
- Scale it all up in real world
- Simulation
- Algorithms/Fine-tuning
- Offline data (videos)
- ?
But what about:

- **Robustness and Generalization**: Handling in-the-wild objects, environments, and task/plan space
- **Interaction: Agent-Robot & Human-Robot**
  - Interacting with VLM Models and Agents
  - Two-way communication
    - Natural specification of tasks by humans
    - Feedback
GENERALIZATION AND ROBUSTNESS

**Benchmarks & Datasets**
[CoRL 2023, NeuRIPS 2023 Challenge]

**Robust Finetuning**
[CVPR/NeurIPS 2023]

**Open-World V/+LM Retrieval**
BENCHMARKS AND DATASETS: HOME ROBOT

HOMEROBOT: OPEN VOCABULARY
MOBILE MANIPULATION


Poster: Today 2-3pm Sequoia 2,
Nov 8, 12:00 - 12:45 pm
HABITAT 2.0 & REARRANGEMENT

Move the box from here

Start State $s^0 = (x, y, z)$

Goal State $s^g = (x', y', z')$
### The Limitations of Generalization

- **Learning a *pick skill***

<table>
<thead>
<tr>
<th>Method</th>
<th>Seen</th>
<th>Unseen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Layouts</td>
</tr>
<tr>
<td>MonolithicRL</td>
<td>91.7 ±1.1</td>
<td>86.3 ±1.4</td>
</tr>
<tr>
<td>SPA</td>
<td>70.2 ±1.9</td>
<td>72.7 ±1.8</td>
</tr>
<tr>
<td>SPA-Priv</td>
<td>77.0 ±1.7</td>
<td>80.0 ±1.6</td>
</tr>
</tbody>
</table>

Perception & policy generalization is still a bottleneck!
HABITAT 2.0 & REARRANGEMENT CHALLENGE 2022

Impractical, single environment and no real-world evaluations
To a sofa

"Pick up the **pitcher** from the **drawer**. Place it on the **serving cart**."

Move object from here

Move pitcher from drawer

To here

To serving cart

Start State  $s^0 = (x, y, z)$  $\xrightarrow[]{}$  Goal State  $s^g = (x', y', z')$

"Open Vocab: novel objects not seen during training"
## HOME ROBOT IN COMPARISON

Robotics stack that enables reproducible benchmarking in sim and real

<table>
<thead>
<tr>
<th></th>
<th>Continuous Actions</th>
<th>Sim2Real</th>
<th>Robotics Stack</th>
<th>Open Licensing</th>
<th>Manipulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Room Rearrangement</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Habitat ObjectNav Challenge</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>TDW-Transport</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>VirtualHome</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ALFRED</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Habitat 2.0 HAB</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ProcTHOR</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RoboTHOR</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Behavior-1K</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>ManiSkill-2</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- **OVMM + HomeRobot**
  - ✓ Continuous Actions
  - ✓ Sim2Real
  - ✓ Robotics Stack
  - ✓ Open Licensing
  - ✓ Manipulation
SIMULATION DATASET

- 200 Habitat Synthetic Scenes Dataset (HSSD) scenes
- 2500+ graspable objects covering 120+ categories
- 5000+ receptacles
High-level policy that calls the skills in sequence:

- FindObj
- GazeAtObj
- Grasp
- FindRecep
- Place

Heuristic and RL variants for skills taking as inputs:

- Depth
- DETIC’s detections
- CLIP (object_name)
- joint states
Move soap_dispenser from couch to chair
Move soap Dispenser from couch to chair
Move soap_dispenser from couch to chair
Move soap_dispenser from couch to chair
## RESULTS - SIMULATION

1200 episodes using 12 validation scenes

<table>
<thead>
<tr>
<th>Perception</th>
<th>Skill</th>
<th>Partial Success Rates</th>
<th>Overall Success Rate</th>
<th>Partial Success Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FindObj</td>
<td>Pick</td>
<td>FindRec</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>Heuristic</td>
<td>None</td>
<td>Heuristic</td>
<td>54.1</td>
</tr>
<tr>
<td></td>
<td>Heuristic</td>
<td>RL</td>
<td>RL</td>
<td>56.5</td>
</tr>
<tr>
<td></td>
<td>RL</td>
<td>None</td>
<td>Heuristic</td>
<td>65.4</td>
</tr>
<tr>
<td></td>
<td>RL</td>
<td>RL</td>
<td>RL</td>
<td>66.6</td>
</tr>
<tr>
<td>DETIC [27]</td>
<td>Heuristic</td>
<td>None</td>
<td>Heuristic</td>
<td>28.7</td>
</tr>
<tr>
<td></td>
<td>Heuristic</td>
<td>RL</td>
<td>RL</td>
<td>29.4</td>
</tr>
<tr>
<td></td>
<td>RL</td>
<td>None</td>
<td>Heuristic</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td>RL</td>
<td>RL</td>
<td>RL</td>
<td>21.7</td>
</tr>
</tbody>
</table>
RESULTS – REAL WORLD

20 experiments in a three-room apartment

<table>
<thead>
<tr>
<th>Real World</th>
<th>FindObj</th>
<th>Pick</th>
<th>FindRec</th>
<th>Overall Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic Only</td>
<td>70</td>
<td>35</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>RL Only</td>
<td>70</td>
<td>45</td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>
Detic fails to detect objects entirely (e.g. cellphone here) or results in fluctuating detections (e.g. counter)
CHALLENGES – PLACE POLICIES ARE NOT ROBUST

Poster: Today 2-3pm Sequoia 2,
Nov 8, 12:00 - 12:45 pm
10. HomeRobot: Open-Vocabulary Mobile Manipulation
The Importance of Interaction

- We now have robots, VLM Agents, and humans!

- How can they interact to further increase generalization?
LARGE VISION-LANGUAGE MODELS AS EMBODIED AGENTS

Input Prompt:
You are a robot exploring an environment for the first time. You will be given a target object to look for and should provide guidance of where to explore based on a single input image. First, you need to list out all the objects that are visible in the scene in a detailed manner. Then, your job is to provide guidance about which object should be explored next in the given image. In this attached image, you need to find the television.

For example if we are in a house and looking for a tv we should explore areas that typically have tv's such as bedrooms and living rooms. You should always provide reasoning for selecting a particular area or object to explore, if there are multiple right answers you should separate them with commas. Always include Reasoning: < your reasoning > and Answer: < your answer ( s ) >. If there are no suitable answers leave the space after Answer: blank.

Objects Visible in the Scene:
1. Striped couch or seating area.
2. Green painted wall.
3. White door.
4. Hanging jackets or clothes.
5. Coal rack.
7. Black bags or equipment on the seating.
8. Ceiling with lights.

Reasoning:
Televisions are typically found in areas where there is comfortable seating or where people gather to relax. The room appears to be a lounge or common area, given the seating and the general layout. While we cannot directly see a television in this image, based on typical room arrangements, televisions are often placed opposite seating areas for optimal viewing. In this case, the television might be placed on the wall adjacent to the green wall or behind the point of view of the image.

Answer:
Explore the wall opposite the striped couch/seating area or turn the viewpoint around to check behind the current field of vision.

Dialogue exchange with GPT-4-Vision is really useful for commonsense spatial reasoning to locate unseen objects.
But using GPT-4 in real-time on a robot is infeasible:
- **No official OpenAI API**
  - Limited to 50 queries per 3 hrs
- **Slow** for real-time robot deployment


Led by Yusuf Ali
Work with Larry Heck & Ritwik Kotra
A SIMPLE SOLUTION

Take your favorite robot on a walk on campus

Generate multiple images based on the robot walks

Use OWL-ViT\(^{(1)}\) for localising object selected by LLaVa

Look around for the target. If found, move towards target and terminate. If not restart the navigation routine.

Select closest object to the unseen object

Use fine tuned LLaVa-v1.5 to find unseen target objects in new environments based on a simple navigation routine

Use the GPT4-V generated dialogue exchanges to

For each image, generate multiple dialogue exchanges where the robot tries to find unseen objects using GPT4-V

Find a study lamp.

The visible objects include bookshelves, books, a table, and stools. Study lamps are often found on tables where people read or study.

Object To Explore: table

Use the GPT4-V generated dialogue exchanges to

LoRA\(^{(1)}\)-finetune SoTA LMM (LLaVa-v1.5\(^{(2)}\))


Demo: Tues/Wed 10.30 - 11 am
Thurs 3.30 - 4.15 pm
Can we simulate robot-agent-human interaction?

Demo: Tues/Wed 10.30 - 11 am
Thurs 3.30 - 4.15 pm
HABITAT 3.0 – COLLABORATION WITH HUMANS IN SIMULATION

Habitat 3.0
A Co-habitat for Humans, Avatars, and Robots


https://aihabitat.org/habitat3/
Potential solutions to the problem of generalization:

- Algorithms/Fine-tuning

Example 1: Robust fine-tuning of foundation models

- How do we not lose generalization during fine-tuning?
How can we finetune based on feedback?

- **Tool**: Projected Gradient Descent
- Observation: some projection operations have closed form solutions
  - Maximum Row Sum Norm (MARS norm)

\[
\Pi(\theta_0, \theta, \gamma) : \theta^* = \theta_0 + \frac{1}{\max\left(1, \frac{\|\theta - \theta_0\|}{\gamma}\right)}(\theta - \theta_0)
\]

- **IDEA**: We can incorporate this operation into the computational graph.
  - Use the bi-level minimization formulation to optimize them
  - See paper for the theoretical analysis

Junjiao Tian, Xiaoliang Dai, Chih-Yao Ma, Zecheng He, Yen-Cheng Liu, Zsolt Kira, *Trainable Projected Gradient Method for Robust Fine-tuning*
HOW CAN WE FINETUNE BASED ON FEEDBACK?

- **Potential solutions to the problem of generalization:**
  - Algorithms/Fine-tuning

- **Example 1:** Robust fine-tuning of foundation models
  - How do we not lose generalization during fine-tuning?

---

Junjiao Tian, Xiaoliang Dai, Chih-Yao Ma, Zecheng He, Yen-Cheng Liu, Zsolt Kira, *Trainable Projected Gradient Method for Robust Fine-tuning*
**TRAINABLE PROJECTED GRADIENT**

- **Idea:** Constrain optimization per layer via projected gradient descent
- **Bi-level optimization**

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\min_{\theta, \gamma} L(x, y; \theta, \gamma)) s.t. (|\theta - \theta_0|_* \leq \gamma)</td>
<td>(\min_{\lambda, \gamma} L(x, y; \theta, \lambda)) s.t. (|\theta - \theta_0|_* \leq \gamma)</td>
<td>(\theta_{t+1} = \Pi(\theta_0, \theta_{t+1}, \gamma_{t+1}))</td>
</tr>
</tbody>
</table>

**Algorithm 1: TPGM**

**Data:** \(\mathcal{D}_{tr}, \mathcal{D}_{val}\)  
**Result:** \(\theta\)

Initialize \(\theta_0^* = \theta_0, \gamma_0 = \epsilon\)

for \(t = \{0, \ldots, T - 1\}\) do

- **Step 1**  
  \(\theta_{t+1} = \arg \min_{\theta} L(x, y; \theta^*)\)  
  \(x, y \in \mathcal{D}_{tr}\)

- **Step 2**  
  \(\gamma_{t+1} = \text{ProjectTune}(\mathcal{D}_{val}, \theta_0, \theta_{t+1}, \gamma_t)\)

- **Step 3**  
  \(\theta^*_{t+1} = \Pi(\theta_0, \theta_{t+1}, \gamma_{t+1})\)
Idea: Constrain optimization per layer via projected gradient descent

- Bi-level optimization

Table 3. **DomainNet Results using CLIP pre-trained ResNet50 with 10% Real Data.** TPGM adjusts to the size of the fine-tuning dataset by imposing stronger per-layer constraints.

<table>
<thead>
<tr>
<th>Method</th>
<th>ID Real</th>
<th>Sketch</th>
<th>Painting</th>
<th>Infograph</th>
<th>Clipart</th>
<th>OOD Avg.</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla FT</td>
<td>57.35 (1.43)</td>
<td>17.48 (0.68)</td>
<td>25.60 (0.70)</td>
<td>10.30 (1.57)</td>
<td>23.01 (0.65)</td>
<td>19.10</td>
<td>0.00</td>
</tr>
<tr>
<td>LP</td>
<td>47.19 (0.93)</td>
<td>17.81 (0.25)</td>
<td>22.71 (2.08)</td>
<td>17.13 (0.75)</td>
<td>17.59 (0.69)</td>
<td>18.81</td>
<td>-17.71</td>
</tr>
<tr>
<td>PF [19]</td>
<td>71.04 (0.91)</td>
<td>27.87 (1.04)</td>
<td><strong>38.31</strong> (1.05)</td>
<td><strong>19.85</strong> (0.70)</td>
<td>33.92 (1.53)</td>
<td>29.99</td>
<td>23.86</td>
</tr>
<tr>
<td>L2-SP [44]</td>
<td>61.41 (0.92)</td>
<td>22.61 (0.52)</td>
<td>30.48 (0.42)</td>
<td>12.28 (0.50)</td>
<td>26.59 (0.57)</td>
<td>22.99</td>
<td>7.08</td>
</tr>
<tr>
<td>MARS-SP [9]</td>
<td>52.53 (0.84)</td>
<td>15.34 (0.54)</td>
<td>21.57 (0.45)</td>
<td>8.49 (0.60)</td>
<td>19.96 (0.01)</td>
<td>16.34</td>
<td>-8.41</td>
</tr>
<tr>
<td>LP-FT [21]</td>
<td>64.11 (0.78)</td>
<td>20.54 (0.27)</td>
<td>30.89 (0.41)</td>
<td>13.58 (0.63)</td>
<td>29.55 (0.82)</td>
<td>23.64</td>
<td>11.78</td>
</tr>
<tr>
<td>TPGM</td>
<td><strong>73.16</strong> (1.27)</td>
<td><strong>29.88</strong> (0.81)</td>
<td>36.80 (1.42)</td>
<td>19.72 (0.12)</td>
<td><strong>35.28</strong> (0.74)</td>
<td><strong>30.42</strong></td>
<td><strong>27.56</strong></td>
</tr>
</tbody>
</table>
UPCOMING ENHANCEMENT - FTP

- Follow-on with several enhancements
  - Remove need for validation set: use split training mini-batch!

Junjiao Tian, Yen-Cheng Liu, James Seale Smith, Zsolt Kira,
Fast Trainable Projection for Robust Fine-Tuning
https://github.com/GT-RIPL/FTP
Potential solutions to the problem of generalization:
- Algorithms/Fine-tuning

Example 2: Open-World Learning through Retrieval

Rabah Ouldoughi, Chia-Wen Kuo, Zsolt Kira
CLIP-GCD: Simple Language Guided Generalized Category Discovery
CONCLUSIONS

- Huge advancements through LLMs & VLMs in Sense, Plan, and maybe even Act

- Open field on:
  - Generalization & Robustness
  - Interaction & Finetuning

- We have pushed towards better benchmarks, datasets, and algorithms in these areas, but much work remains

- [Diagram showing cycles of Finetune, Act, and Interact]