ACT, INTERACT, AND FINETUNE WITH VISION-LANGUAGE MODELS

Georgia Tech Machine Learning

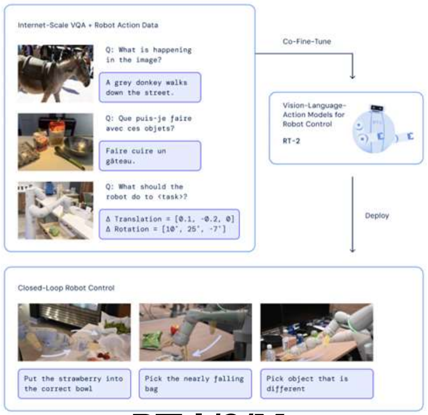
RIPL
Robotics Perception and Learning

Zsolt Kira
Assistant Professor
School of Interactive Computing
Georgia Tech

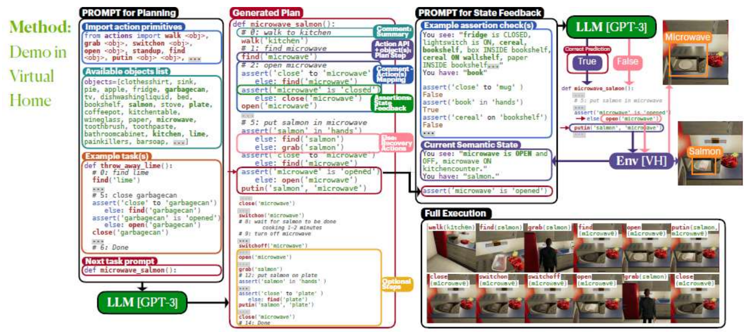
CHALLENGES IN ROBOTICS

- 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

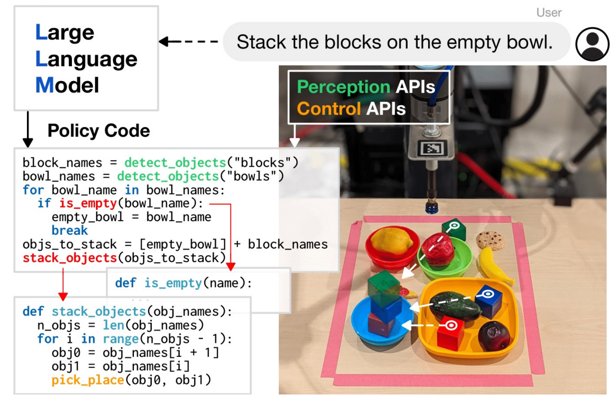
EXPLOSION OF PROGRESS



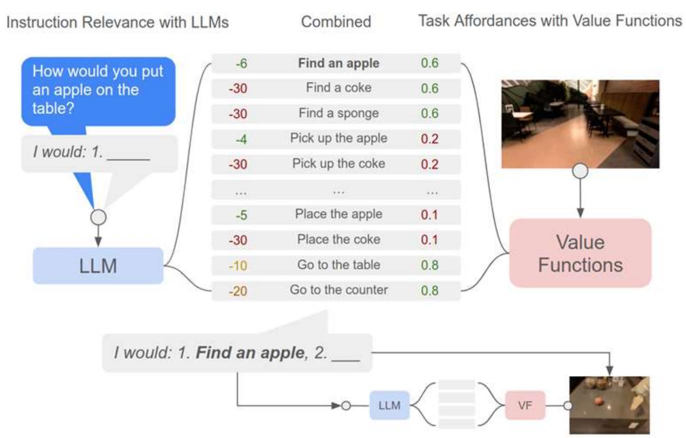
RT-1/2/X



ProgPrompt

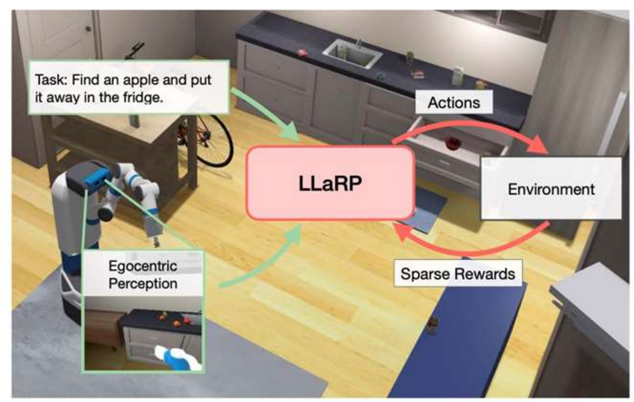


Code-As-Policies



SayCan

Reinforcement Learning of an LLM-based Policy for Many Tasks



LLaRP

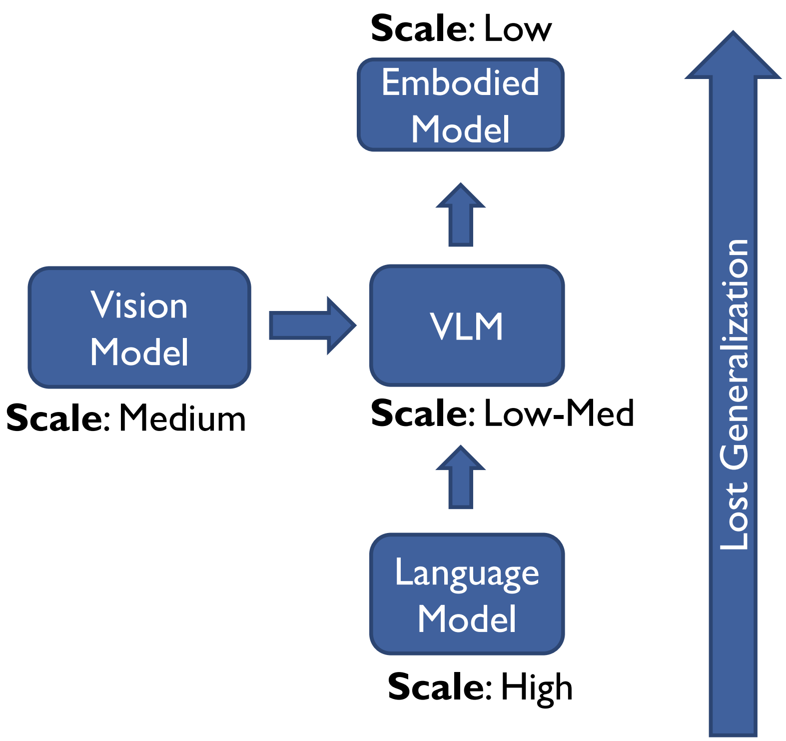
Shameless Plug: https://github.com/GT-RIPL/Awesome-LLM-Robotics

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 ❓

ARE WE DONE?

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

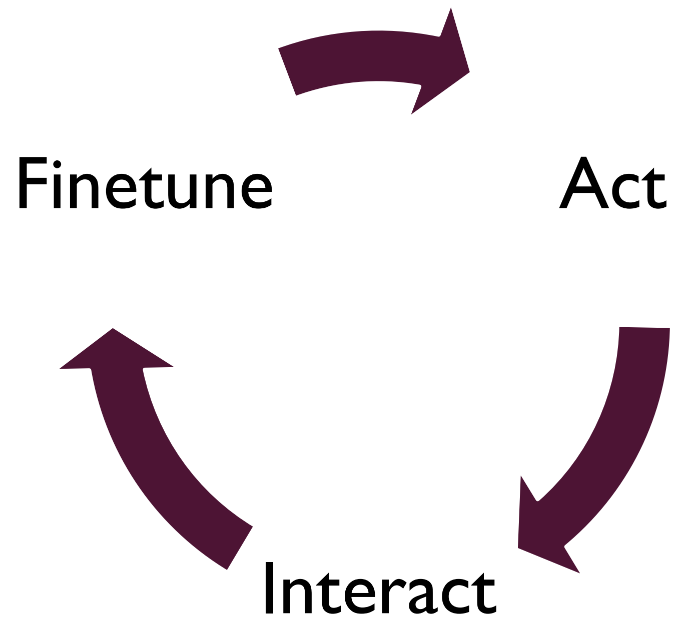


Solutions:

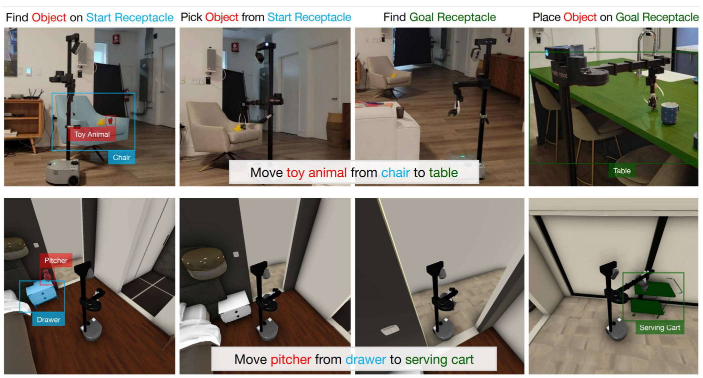
- Scale it all up in real world
- Simulation
- Algorithms/Fine-tuning
- Offline data (videos)
- ?

ARE WE DONE?

- 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

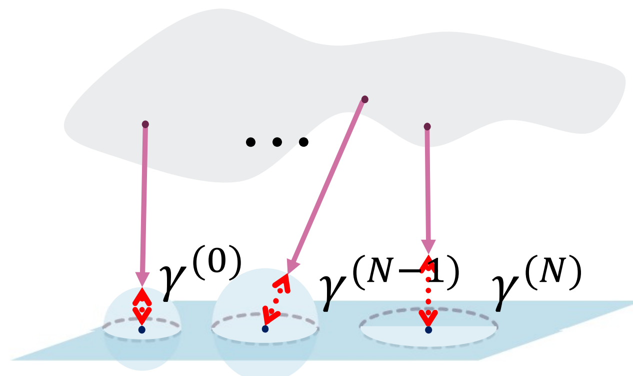


GENERALIZATION AND ROBUSTNESS



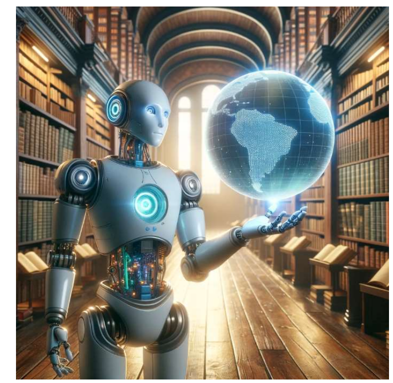
Benchmarks & Datasets

[CoRL 2023, NeuRIPS 2023 Challenge]



Robust Finetuning

[CVPR/NeurIPS 2023]



Open-World V/+LM Retrieval

[CVPR/NeurIPS 2023, arXiv:2305.10420, ECCV 2022]

BENCHMARKS AND DATASETS: HOME ROBOT

HOMEROBOT: OPEN VOCABULARY MOBILE MANIPULATION

Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, Chris Paxton

Poster: Today 2-3pm Sequoia 2,
Nov 8, 12:00 - 12:45 pm

HABITAT 2.0 & REARRANGEMENT

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

THE LIMITATIONS OF GENERALIZATION

- Learning a *pick skill*

Method	Seen	Unseen: Layouts	Unseen: Objects	Unseen: Receptacles
MonolithicRL	91.7 ± 1.1	86.3 ± 1.4	74.7 ± 1.8	52.7 ± 2.0
SPA	70.2 ± 1.9	72.7 ± 1.8	72.7 ± 1.8	60.3 ± 2.0
SPA-Priv	77.0 ± 1.7	80.0 ± 1.6	79.2 ± 1.7	60.7 ± 2.0

Classical method w/ RRTs

Add ground truth pointcloud

Perception & policy generalization is still a bottleneck!

HABITAT 2.0 & REARRANGEMENT CHALLENGE 2022

Start State $s^0 = (x, y, z)$ ⇨ Goal State $s^g = (x', y', z')$

Impractical, single environment and no real-world evaluations

OPEN-VOCAB MOBILE MANIPULATION CHALLENGE

~~Start State~~ $\cancel{s^0 = (x, y, z)}$ $\Rightarrow$ ~~Goal State~~ $\cancel{s^g = (x', y', z')}$

"Pick up the *<u>pitcher</u>* from the *<u>drawer</u>*. Place it on the *<u>serving cart</u>*."

<object> **<start_receptacle>** **<goal_receptacle>**

Open Vocab: novel objects not seen during training

HOME ROBOT IN COMPARISON

Robotics stack that enables reproducible benchmarking in sim and real

		Continuous Actions	Sim2Real	Robotics Stack	Open Licensing	Manipulation
Room Rearrangement	[27]	✖	✖	✖	✔	✖
Habitat ObjectNav Challenge	[28]	✔	✖	✖	✔	✖
TDW-Transport	[29]	✖	✖	✖	✓	✓
VirtualHome	[30]	✖	✖	✖	✔	✓
ALFRED	[6]	✖	✖	✖	✔	✓
Habitat 2.0 HAB	[21]	✔	✖	✖	✔	✔
ProcTHOR	[31]	✖	✖	✖	✔	✔
RoboTHOR	[32]	✖	✔	✖	✔	✖
Behavior-1K	[33]	✔	✔	✖	✖	✓
ManiSkill-2	[34]	✔	✓	✖	✓	✔
OVMM + HomeRobot		✔	✔	✔	✔	✔

SIMULATION DATASET

- 200 Habitat Synthetic Scenes Dataset (HSSD) scenes

- 2500+ graspable objects covering 120+ categories

- 5000+ receptacles

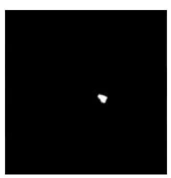
BASELINE

High-level policy that calls the skills in sequence:

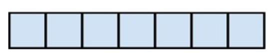
FindObj → GazeAtObj → Grasp → FindRecep → Place

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

Simulation Results	Skill			Partial Success Rates			Overall Success Rate	Partial Success Metric
Perception	Navigation	Gaze	Place	FindObj	Pick	FindRec		
Ground Truth	Heuristic	None	Heuristic	54.1	48.5	31.5	5.1	34.8
	Heuristic	RL	RL	56.5	51.5	42.3	13.2	40.9
	RL	None	Heuristic	65.4	54.8	43.7	7.3	42.8
	RL	RL	RL	66.6	61.1	50.9	14.8	48.3
DETIC [27]	Heuristic	None	Heuristic	28.7	15.2	5.3	0.4	12.4
	Heuristic	RL	RL	29.4	13.2	5.8	0.5	12.2
	RL	None	Heuristic	21.9	11.5	6.0	0.6	10.0
	RL	RL	RL	21.7	10.2	6.2	0.4	9.6

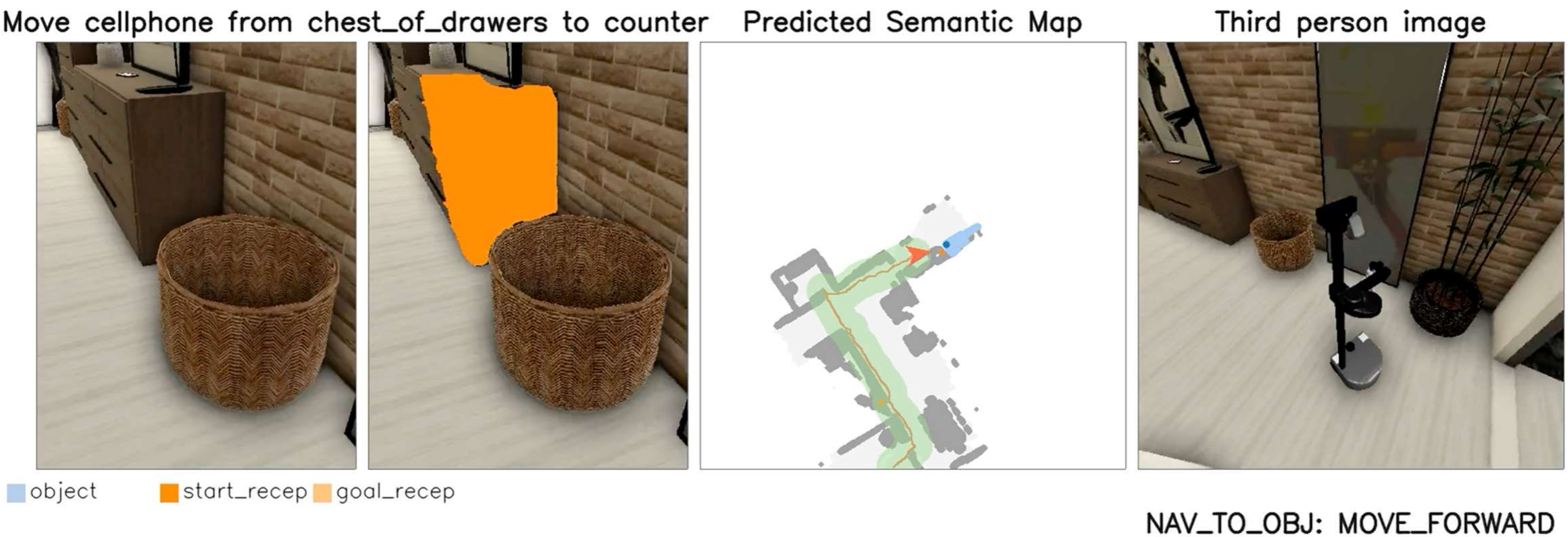
RESULTS – REAL WORLD

20 experiments in a three-room apartment

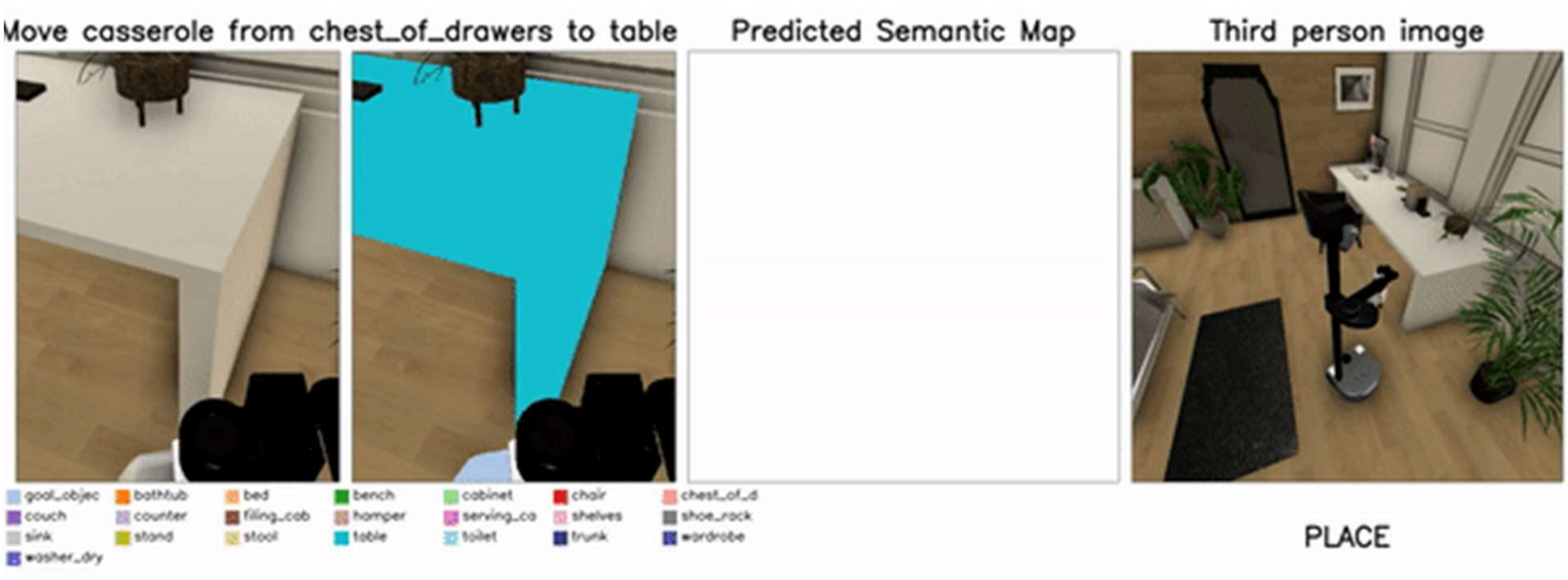
Real World	FindObj	Pick	FindRec	Overall Success
Heuristic Only	70	35	30	15
RL Only	70	45	30	20

CHALLENGES – IMPERFECT PERCEPTION

Detic fails to detect objects entirely (eg. cellphone here) or results in fluctuating detections (eg. counter)



CHALLENGES – PLACE POLICIES ARE NOT ROBUST

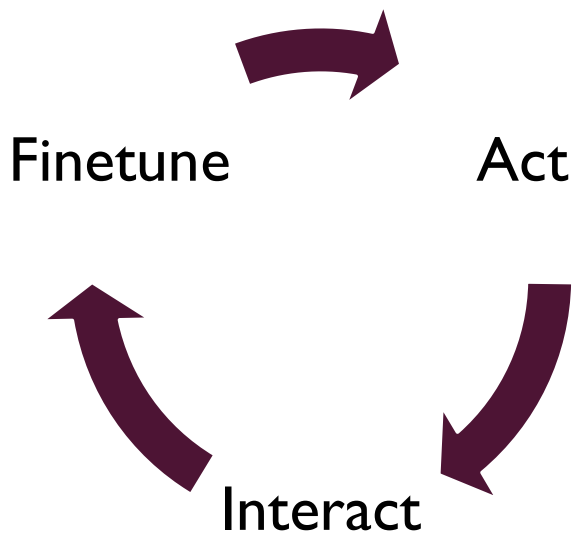


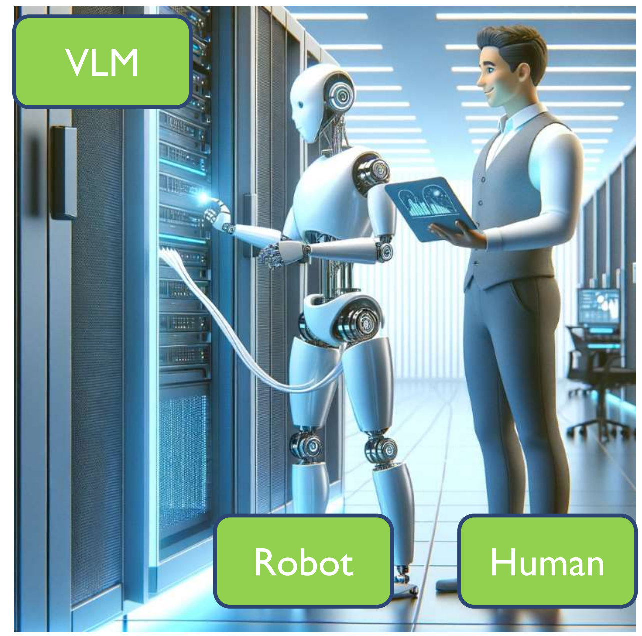
Poster: Today 2-3pm Sequoia 2,
Nov 8, 12:00 - 12:45 pm

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 explore 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 afters 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. Coat rack.
6. Ladder.
7. Black bags or equipment on the seating.
8. Ceiling with lights.
9. Checkered floor tiles.

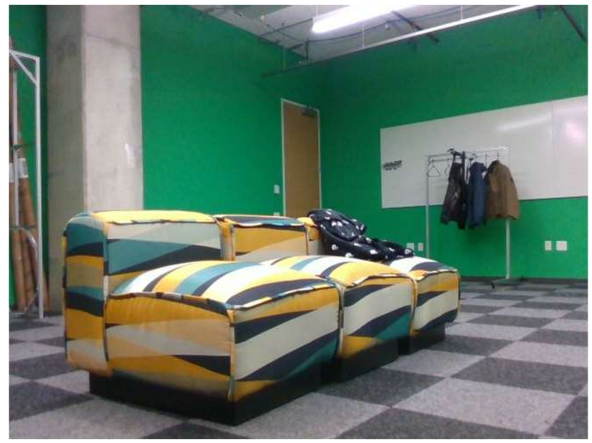
Accurately detects all relevant objects in the image

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.

Invoke commonsense reasoning for spatial understanding

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

Accurate next steps to find target "television"



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

Prompt taken from: Shah, Dhruv, Michael Equi, Blazej Osinski, Fei Xia, Brian Ichter, and Sergey Levine. "**Navigation with large language models: Semantic guesswork as a heuristic for planning.**" *CoRL 2023*

A SIMPLE SOLUTION

Take your favorite robot on a walk on campus

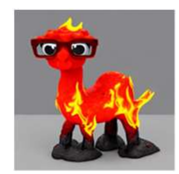
Generate multiple images based on the robot walks

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

Find a study lamp.

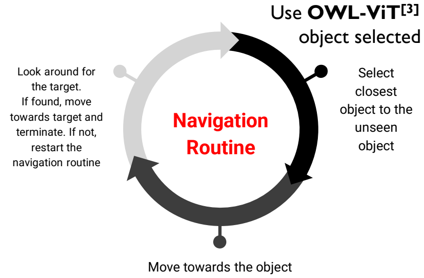
The visible objects include bookshelves, books, a table, and stools. Study lamps are often found on tables where people read or study.
Object To Explore: table

Use the GPT4-V generated dialogue exchanges to ***LoRA[1]-finetune*** SoTA LMM (LLaVa-v1.5[2])



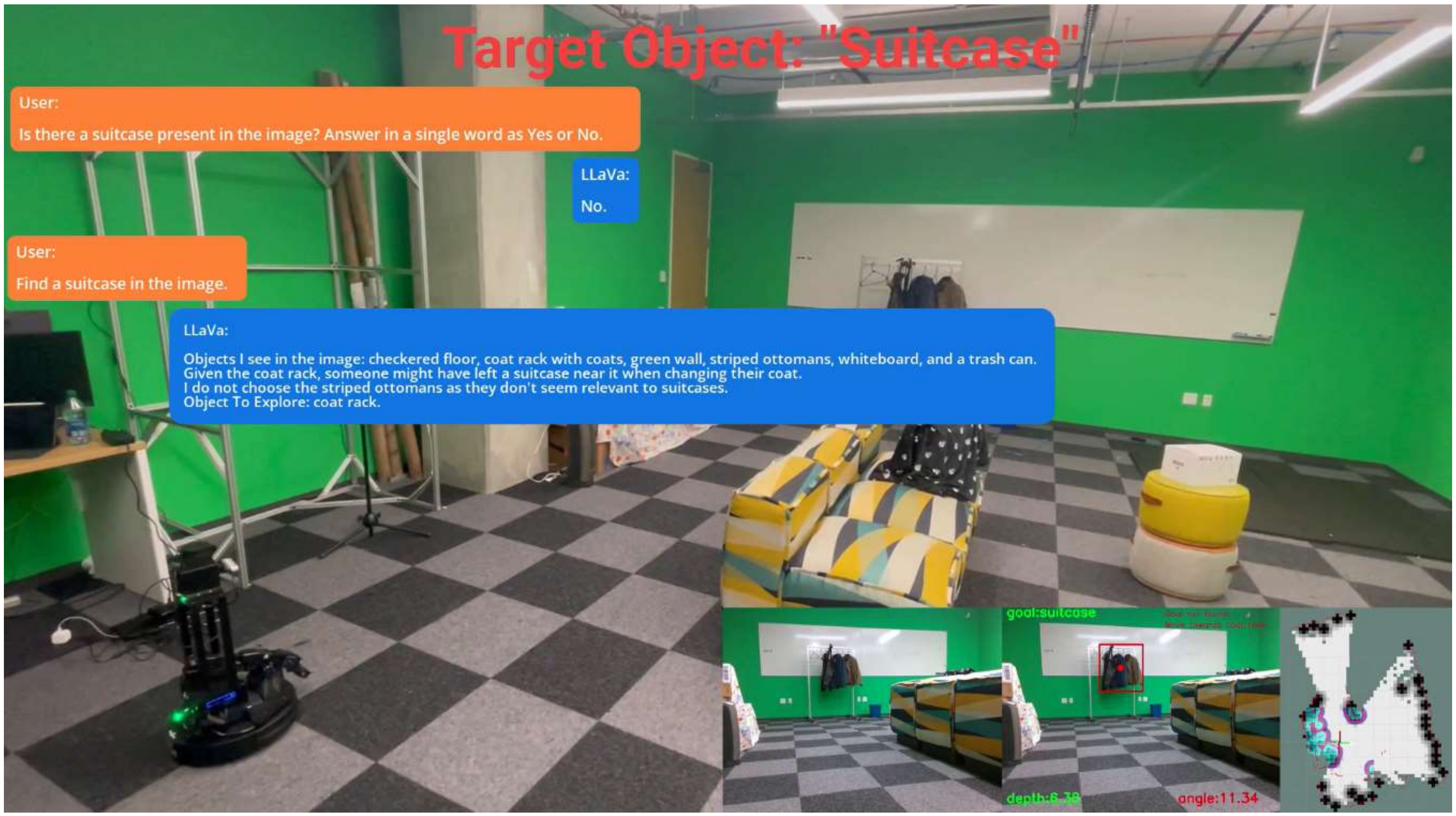
Use fine tuned LLaVa-v1.5 to **find unseen target objects** in new environments based on a simple navigation routine

Use **OWL-ViT[3]** for localising object selected by LLaVa



[1] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. **"LoRA: Low-Rank Adaptation of Large Language Models."** *ICLR 2022*
[2] Liu, Haotian, Chunyuan Li, Yuheng Li, and Yong Jae Lee. **"Improved Baselines with Visual Instruction Tuning."** *arXiv preprint arXiv:2310.03744* (2023)
[3] M. Minderer, A.Gritsenko, A.Stone, M.Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, N. Houlsby. **"Simple Open-Vocabulary Object Detection with Vision Transformers"** *ECCV 2022*

VIDEO



Demo: Tues/Wed 10.30 - 11 am
Thurs 3.30 - 4.15 pm

VIDEO

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

Meta AI

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, Roozbeh Mottaghi

https://aihabitat.org/habitat3/

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

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Junjiao Tian

Robotics Ph.D.

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

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Junjiao Tian

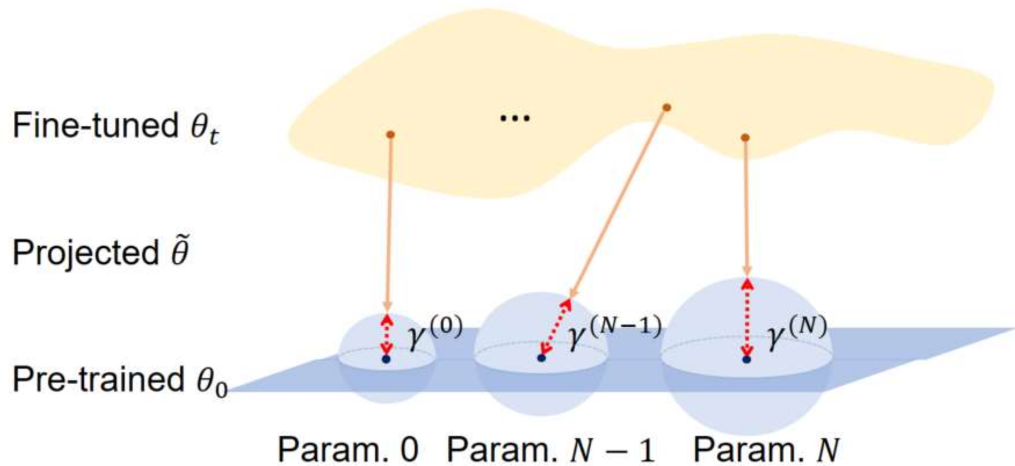
Robotics Ph.D.

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?

Trainable Projected Gradient Descent (TPGM)



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

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Junjiao Tian

Robotics Ph.D.

TRAINABLE PROJECTED GRADIENT

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

$$\underbrace{\min_{\lambda,\gamma|(x,y)\in\mathcal{D}_{val}}}_{\text{Step 2}} \underbrace{\min_{\theta|(x,y)\in\mathcal{D}_{tr}}}_{\text{Step 1}} \mathcal{L}(x,y;\theta,\lambda,\gamma) \quad \text{s.t.} \quad \underbrace{\|\theta-\theta_0\|_* \le \gamma}_{\text{Step 3}}$$

Algorithm 1: TPGM

Data: $\mathcal{D}_{tr}$,$\mathcal{D}_{val}$

Result: θ

Initialize $\theta_0^* = \theta_0, \gamma_0 = \epsilon$

for $t = \{0, ..., T-1\}$ **do**

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

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

Step 3 $\quad \theta_{t+1}^* = \Pi(\theta_0, \theta_{t+1}, \gamma_{t+1})$

TPGM - RESULTS

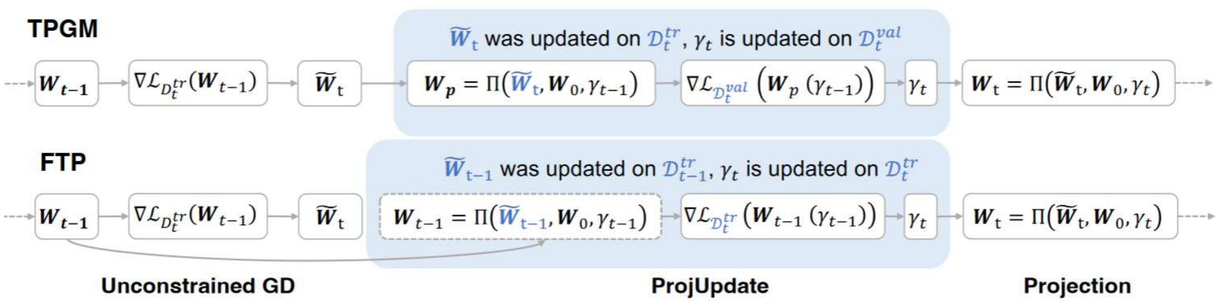
- **Idea:** Constrain optimization per layer via projected gradient descent
  - Bi-level optimization

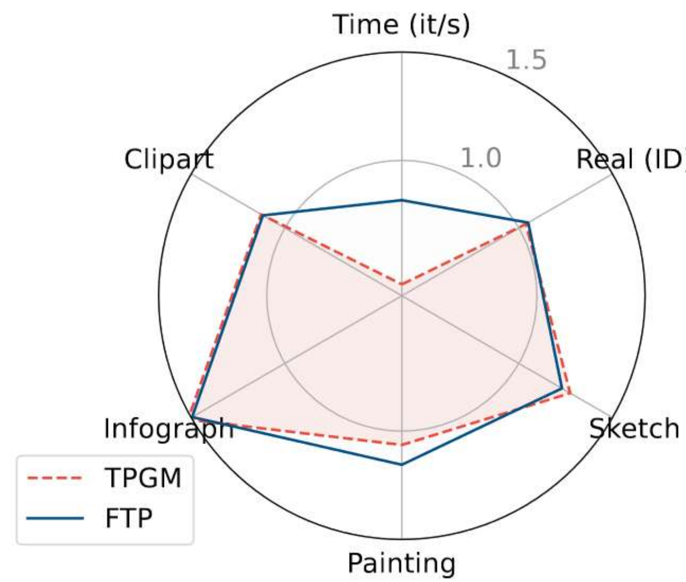
Table 3. **DomainNet Results using CLIP pre-trained ResNet50 with 10% Real Data.** TPGM adjusts to the size of the fine-tuning dataset by imposing stronger per-layer constraints.

	ID	OOD				Statistics		
	Real	Sketch	Painting	Infograph	Clipart	OOD Avg.	ID Δ (%)	OOD Δ (%)
Vanilla FT	57.35 (1.43)	17.48 (0.68)	25.60 (0.70)	10.30 (1.57)	23.01 (0.65)	19.10	0.00	0.00
LP	47.19 (0.93)	17.81 (0.25)	22.71 (2.08)	17.13 (0.75)	17.59 (0.69)	18.81	-17.71	-1.52
PF [19]	71.04 (0.91)	27.87 (1.04)	**38.31** (1.05)	**19.85** (0.70)	33.92 (1.53)	29.99	23.86	57.01
L2-SP [44]	61.41 (0.92)	22.61 (0.52)	30.48 (0.42)	12.28 (0.50)	26.59 (0.57)	22.99	7.08	20.37
MARS-SP [9]	52.53 (0.84)	15.34 (0.54)	21.57 (0.45)	8.49 (0.60)	19.96 (0.01)	16.34	-8.41	-14.44
LP-FT [21]	64.11 (0.78)	20.54 (0.27)	30.89 (0.41)	13.58 (0.63)	29.55 (0.82)	23.64	11.78	23.77
TPGM	**73.16** (1.27)	**29.88** (0.81)	36.80 (1.42)	19.72 (0.12)	**35.28** (0.74)	**30.42**	**27.56**	**59.27**

UPCOMING ENHANCEMENT - FTP

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



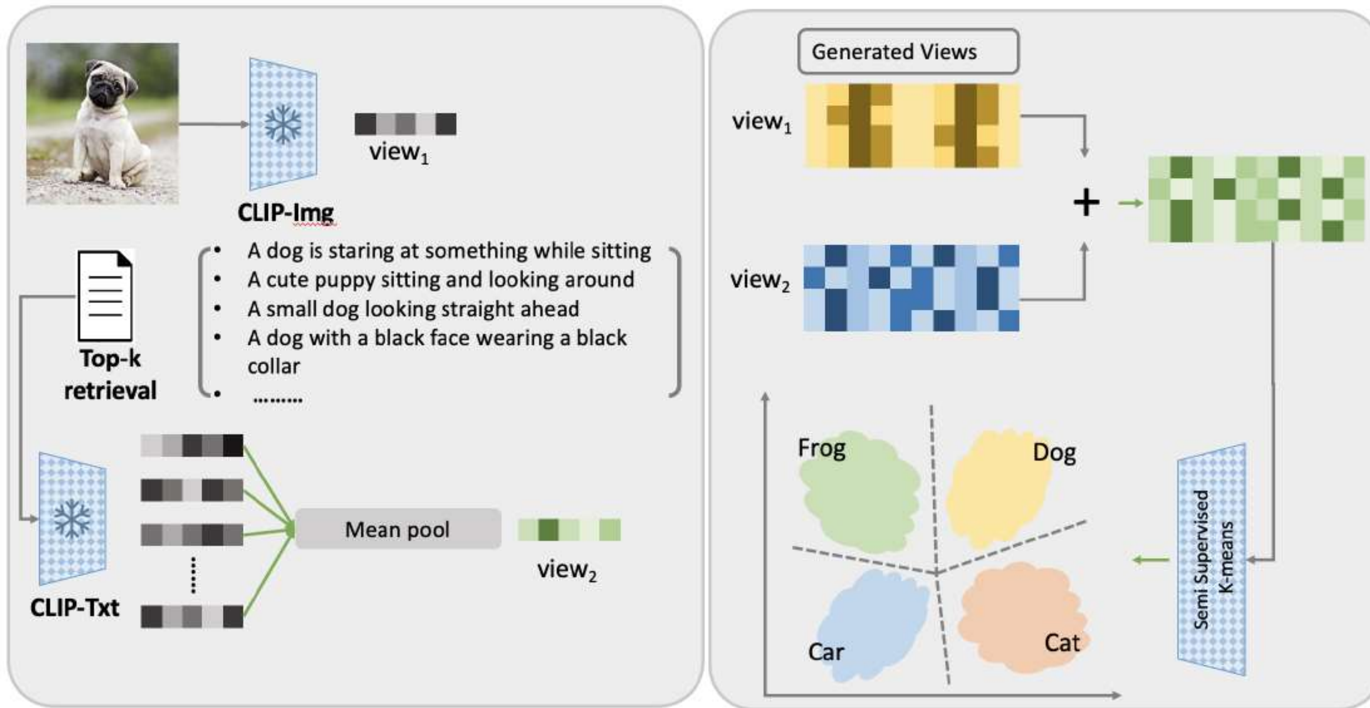


Junjiao Tian, Yen-Cheng Liu, James Seale Smith, Zsolt Kira,
Fast Trainable Projection for Robust Fine-Tuning
https://github.com/GT-RIPL/FTP

NEURAL INFORMATION PROCESSING SYSTEMS

AUGMENTING WITH RETRIEVAL

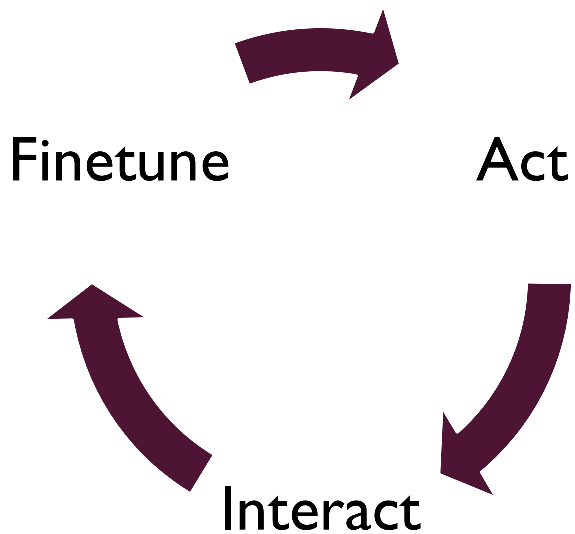
- **Potential solutions to the problem of generalization**:
  - Algorithms/Fine-tuning
- **Example 2:** Open-World Learning through Retrieval



Rabah Ouldnoughi, 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



Karmesh Yadav

CS Ph.D. (co-advised with Dhruv Batra)

Junjiao Tian

Robotics Ph.D.

Yusuf Ali

CS Ph.D.

Andrew Szot

ML Ph.D. (co-advised with Dhruv Batra)

hello robot™ Meta AI Carnegie Mellon University