Topics:

- Bias/Fairness
- RIPL Research
- Wrap-up:
  - Open directions in Deep Learning

CS 4644-DL / 7643-A

ZSOLT KIRA

- **Projects!**
  - Guidelines: @490
  - Project due **April 26 11:59pm** (grace period **April 28th**)
  - Cannot extend due to grade deadlines!

- CIOS
  - Please make sure to fill out! Let us know about things you liked and didn't like in comments so that we can keep or improve!
  - http://b.gatech.edu/cios

Bias & Fairness

ML and Fairness

- AI effects our lives in many ways
- Widespread algorithms with many small interactions
  – e.g. search, recommendations, social media
- Specialized algorithms with fewer but higher-stakes interactions
  – e.g. medicine, criminal justice, finance
- At this level of impact, algorithms can have unintended consequences
- Low classification error is not enough, need fairness

Slide By Aaron Roth

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like


INDEPENDENT

GOOGLE'S ALGORITHM SHOWS PRESTIGIOUS JOB ADS TO MEN, BUT NOT TO WOMEN

REUTERS

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

MIT Technology Review

Intelligent Machines

How to Fix Silicon Valley's Sexist Algorithms

Computers are inheriting gender bias implanted in language data sets—and not everyone thinks we should correct it.

PROPUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Research shows that Amazon's tech has a harder time i

Slide By Aaron Roth

Machine Learning and Social Norms

- Sample norms: privacy, fairness, transparency, accountability…
- Possible approaches
  - "traditional": legal, regulatory, watchdog
  - *Embed* social norms in data, algorithms, models
- Case study: privacy-preserving machine learning
  - "single", strong, definition (differential privacy)
  - almost every ML algorithm has a private version
- Fair machine learning
  - not so much…
  - impossibility results

(Un)Fairness Where?

- Data (input)
  - e.g. more arrests where there are more police
  - Label should be “committed a crime”, but is “convicted of a crime”
  - try to “correct” bias
- Models (output)
  - e.g. discriminatory treatment of subpopulations
  - build or “post-process” models with subpopulation guarantees
  - equality of false positive/negative rates; calibration
- Algorithms (process)
  - learning algorithm *generating* data through its decisions
  - e.g. don’t learn outcomes of denied mortgages
  - lack of clear train/test division
  - design (sequential) *algorithms* that are fair

Slide By Aaron Roth

When the *training data* we collect does not contain representative samples of the true distribution.

Examples:

- If we use data gathered from smart phones, we would likely be underestimating poorer and older populations.
- ImageNet (a very popular image dataset) with 1.2 million images. About 45% of these images were taken in the US and the majority of the rest in North America and Western Europe. Only about 1% and 2.1% of the images come from China and India respectively.

Data Bias

Measurement Bias

Often we are gathering data that contains (noisy) proxies of characteristics of interest. Some examples:

- Financial responsibility → Credit Score
- Crime Rate → Arrest Rate
- Intelligence → SAT Score

If these measurements are not measured equally across groups or places (or aren't relevant to the task at hand), this can be another source of bias.

Examples:

- If factory workers are monitored more often, more errors are spotted. This can result in a **feedback loop** to encourage more monitoring in the future.
  - Same principles at play with predictive policing. Minoritized communities were more heavily policed in the past, which causes more instances of documented crime, which then leads to more policing in the future.

- Women are more likely to be misdiagnosed (or not diagnosed) for conditions where self-reported pain is a symptom. In this case aspect of our data “diagnosed with X” is a biased proxy for “has condition X”.
- The feature we measure is a poor representation of the quality of interest (e.g., SAT score doesn’t actually measure intelligence)

Measurement Bias

What does it mean for a model to be fair or unfair? Can we come up with a numeric way of measuring fairness?

Lots of work in the field of ML and fairness is looking into mathematical definitions of fairness to help us spot when something might be unfair.

- There is not going to be one central definition of fairness, as each definition is a mathematical statement of which behaviors are/aren’t allowed.
- Different definitions of fairness can be contradictory!

ML and Fairness

- Fairness is morally and legally motivated
- Takes many forms
- Criminal justice: recidivism algorithms (COMPAS)
  - Predicting if a defendant should receive bail
  - Unbalanced false positive rates: more likely to wrongly deny a black person bail

Table 1: ProPublica Analysis of COMPAS Algorithm

	White	**Black**
Wrongly Labeled High-Risk	23.5%	44.9%
Wrongly Labeled Low-Risk	47.7%	28.0%

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Why Fairness is Hard

- Suppose we are a bank trying to fairly decide who should get a loan
  - i.e. Who is most likely to pay us back?
- Suppose we have two groups, A and B (the sensitive attribute)
  - This is where discrimination could occur
- The simplest approach is to remove the sensitive attribute from the data, so that our classier doesn't know the sensitive attribute. Often called **"Fairness through unawareness"**

Table 2: To Loan or Not to Loan?

Age	Gender	Postal Code	Req Amt	A or B?	Pay
46	F	M5E	$300	A	1
24	M	M4C	$1000	B	1
33	M	M3H	$250	A	1
34	F	M9C	$2000	A	0
71	F	M3B	$200	A	0
28	M	M5W	$1500	B	0

Why Fairness is Hard

- However, if the sensitive attribute is correlated with the other attributes, this isn't good enough
- It is easy to predict race if you have lots of other information (e.g. home address, spending patterns)
- More advanced approaches are necessary

Table 3: To Loan or Not to Loan? (masked)

Age	Gender	Postal Code	Req Amt	A or B?	Pay
46	F	M5E	$300	?	1
24	M	M4C	$1000	?	1
33	M	M3H	$250	?	1
34	F	M9C	$2000	?	0
71	F	M3B	$200	?	0
28	M	M5W	$1500	?	0

Doesn’t work in practice. This does not prevent historical or measurement bias. Protected attributes can be unintentionally inferred from other, related attributes (e.g., in some cities, zip code can be deeply correlated with race).

Definitions of Fairness – Group Fairness

- So we've built our classier . . . how do we know if we're being fair?
- One metric is demographic parity | requiring that the same percentage of A and B receive loans
  - What if 80% of A is likely to repay, but only 60% of B is?
  - Then demographic parity is too strong
- Could require equal false positive/negative rates
  - When we make an error, the direction of that error is equally likely for both groups

$$P(loan|no\ repay, A) = P(loan|no\ repay, B)$$
$$P(no\ loan|would\ repay, A) = P(no\ loan|would\ repay, B)$$

- These are definitions of group fairness
- Treat different groups equally"

Definitions of Fairness – Individual Fairness

- Also can talk about individual fairness | “Treat similar examples similarly"
- Learn fair representations
  - Useful for classification, not for (unfair) discrimination
  - Related to domain adaptation
  - Generative modelling/adversarial approaches

(a) Unfair representations (b) Fair(er) representations

Figure 1: “The Variational Fair Autoencoder” (Louizos et al., 2016)

Conclusion

- This is an exciting field, quickly developing
- Central definitions still up in the air
- AI moves fast | lots of (currently unchecked) power
- Law/policy will one day catch up with technology
- Those who work with AI should be ready
  - **Think about implications of what you develop!**

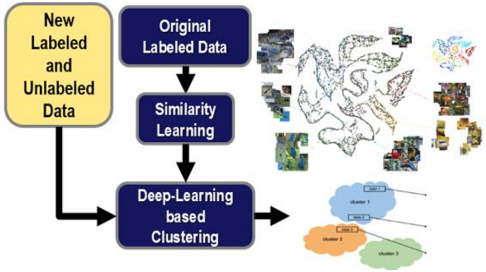
Research in RIPL

Zsolt Kira

zkira@gatech.edu
Associate Professor
School of Interactive Computing
Georgia Institute of Technology

RIPL
Robotics Perception and Learning

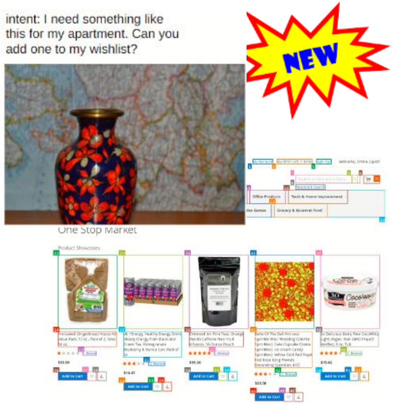
Research Interests: Intersection of deep learning and robotics, focusing on robustness and decision-making in an open world



How can perception deal with changing environments and the open world?

Robust Open-World Learning

- Past: Semi and self-supervised, few-shot, continual learning
- Open-world learning and Vision-Language Models
- Robust fine-tuning of VMs/VLMs



How can we use VLMs for Learning, Planning, and Reasoning Agents

Planning & Reasoning

- VLMs for reasoning/planning
- Grounding

How can we scale robotics in DL era?

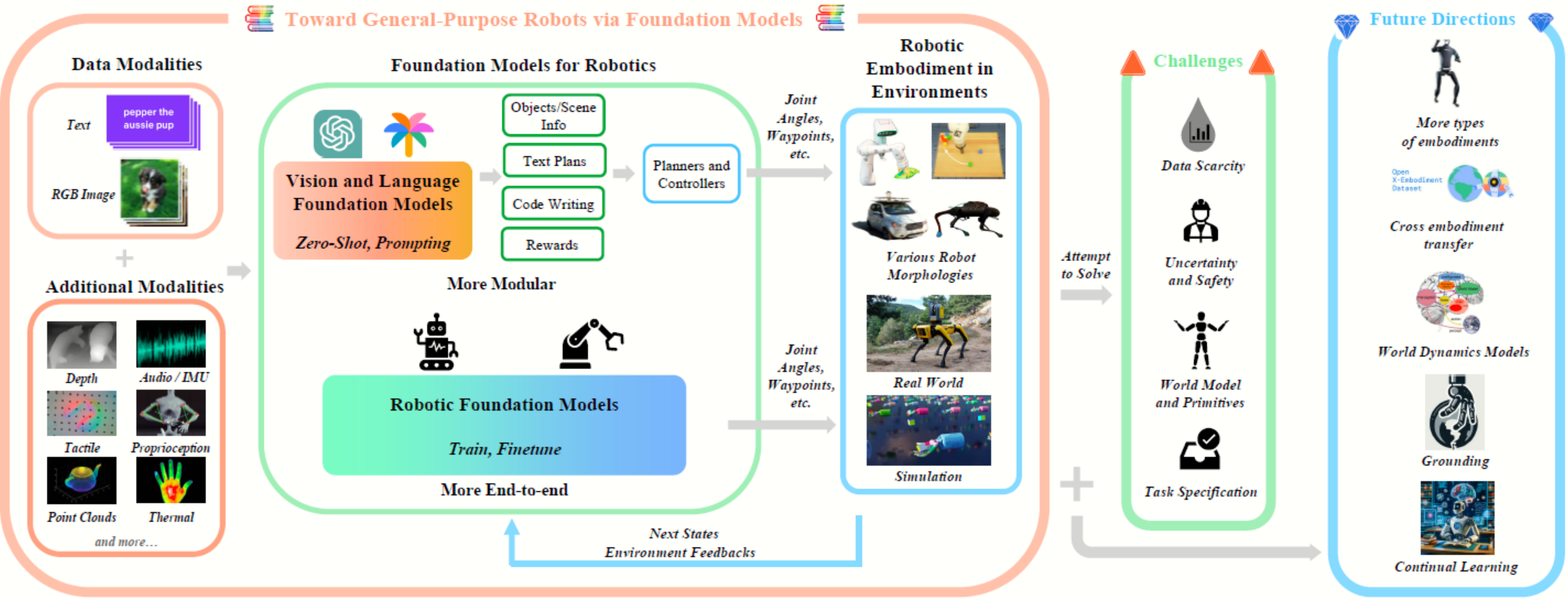
Scaling up Robotics

- Better simulation w/ NeRFs/3D
- Self-supervised and pre-training
- Combinations with large language and multi-modal models
  - Long-Context Models
- Vision-Language Action Models

Challenges in Robotics

- Data flywheel
  - Hard to gather
  - Potentially dangerous
  - Huge heterogeneity
- Robusness
  - In-the-wild data
  - Long-tail (see self-driving cars)
  - Long-horizon decision-making
  - Physics!
- Reliability 24/7
- Cost?

Robotic Foundation Models



Habitat 2.0 & 3.0

Train Pick Policy on Large Scale Randomization

Multimodal Large Language Models


Bing's A.I. Chat: 'I Want to Be Alive. 😈'

In a two-hour conversation with our columnist, Microsoft's new chatbot said it would like to be human, had a desire to be destructive and was in love with the person it was chatting with. Here's the transcript.

https://www.nytimes.com/article/ai-artificial-intelligence-chatbot.html


ARTIFICIAL INTELLIGENCE

ChatGPT is about to revolutionize the economy. We need to decide what that looks like.

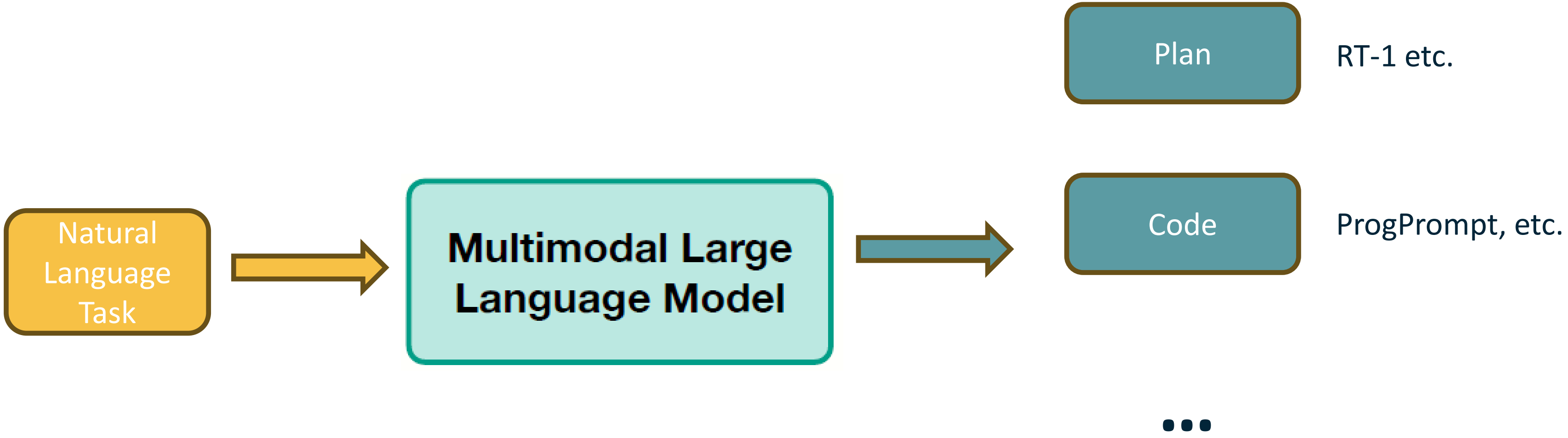
https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/

Multimodal Large Language Model

Gemini 1.5

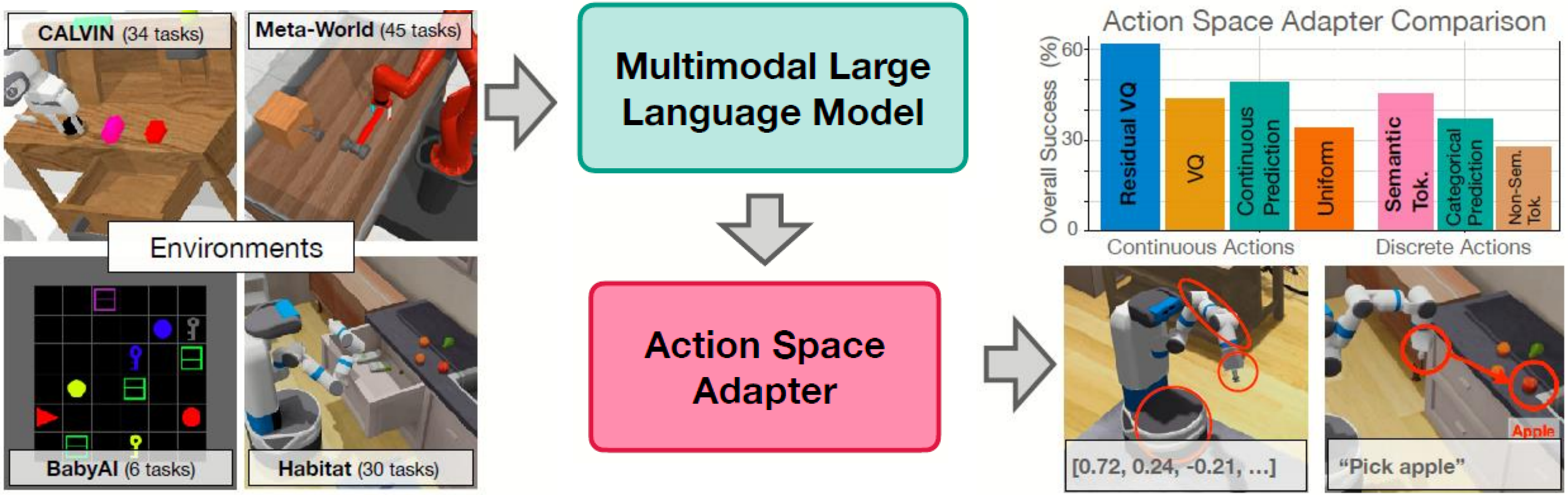
LLAMA 2

Multimodal Large Language Models



What about VLMS for direct task to action?

Vision-Language Action Models



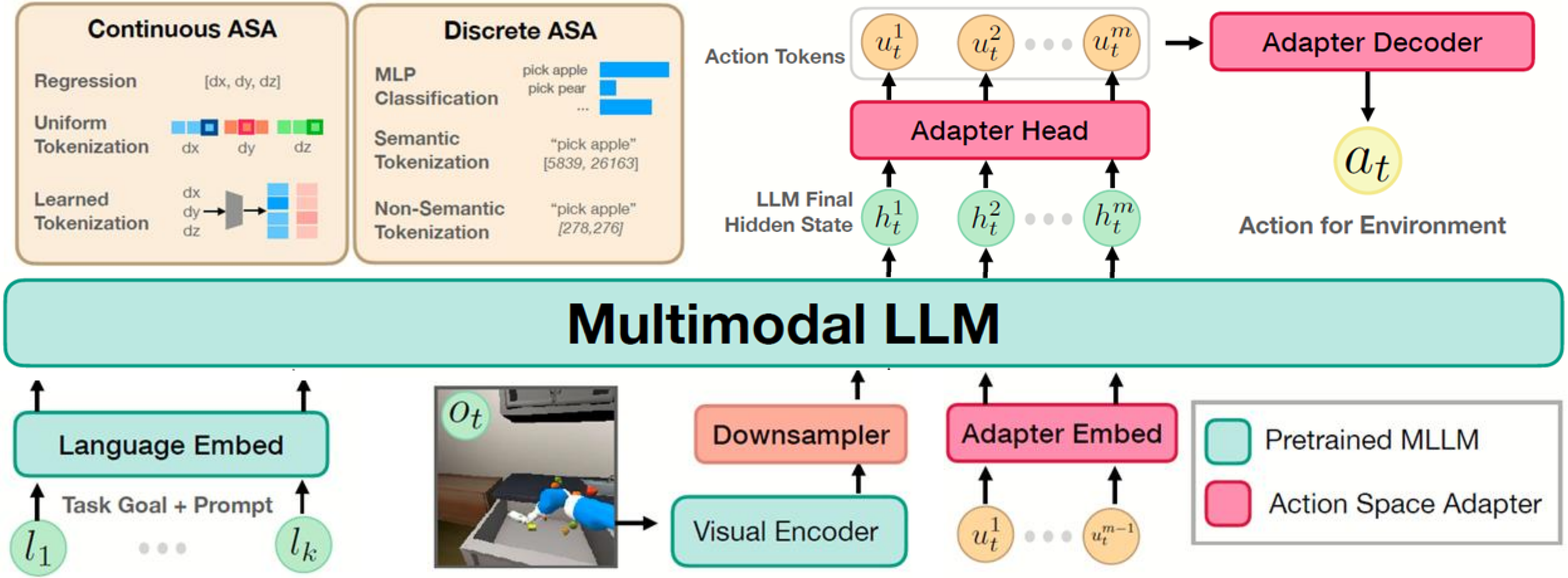
Lots of great concurrent work! OpenVLA, LLARVA, etc.

Andrew Szot

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

Szot et al., Grounding Multimodal Large Language Models in Actions, NeurIPS 2024

Georgia Tech



We finetune the ASAs, downsampler, and MLLM

Szot et al., Grounding Multimodal Large Language Models in Actions

Andrew Szot

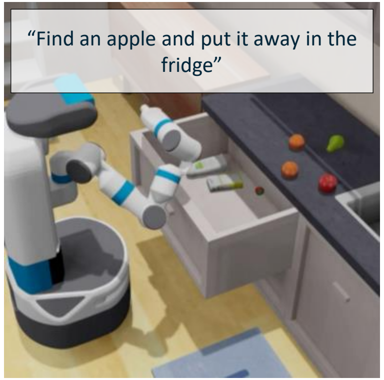
ML Ph.D. (co-advised with Dhruv Batra)

Georgia Tech

VLA: Results across Spectrum of Generalization

	Total	Aggregated		Per Dataset Breakdown											
		Behavior Generalization	Paraphrastic Robustness	Train	Scene	Instruct Rephrasing	Novel Objects	Multiple Rearrange	Referring Expressions	Context	Irrelevant Text	Multiple Objects	Spatial	Conditional Instructs
SemLang	51 ± 1	56 ± 2	47 ± 1	94 ± 3	94 ± 6	92 ± 1	97 ± 0	80 ± 6	31 ± 3	46 ± 14	66 ± 6	2 ± 2	0 ± 0	46 ± 4
Lang	27 ± 12	31 ± 14	24 ± 10	72 ± 13	58 ± 11	74 ± 12	76 ± 29	21 ± 10	10 ± 12	12 ± 11	20 ± 13	0 ± 0	2 ± 3	26 ± 16
Pred	42 ± 2	45 ± 3	38 ± 1	99 ± 1	96 ± 4	92 ± 2	95 ± 4	47 ± 5	26 ± 2	34 ± 2	32 ± 2	0 ± 1	8 ± 1	39 ± 3

Many tasks we want an agent to take actions to autonomously complete



Robotic Manipulation

Navigation

Games

UI Control

Can we have ***one*** policy that does all of these?

How can we create a generalist agent capable of excelling in diverse interactive tasks?

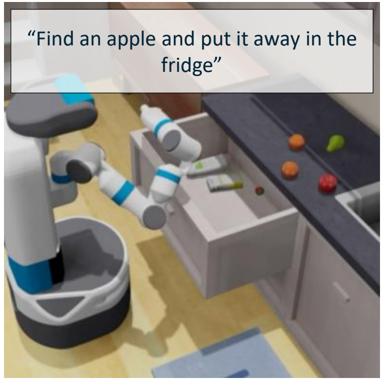
Adapt a pre-trained Multimodal LLM

Multimodal LLM

Interactive Data

Adapt

Generalist Agent



Robotic Manipulation

Navigation

Games

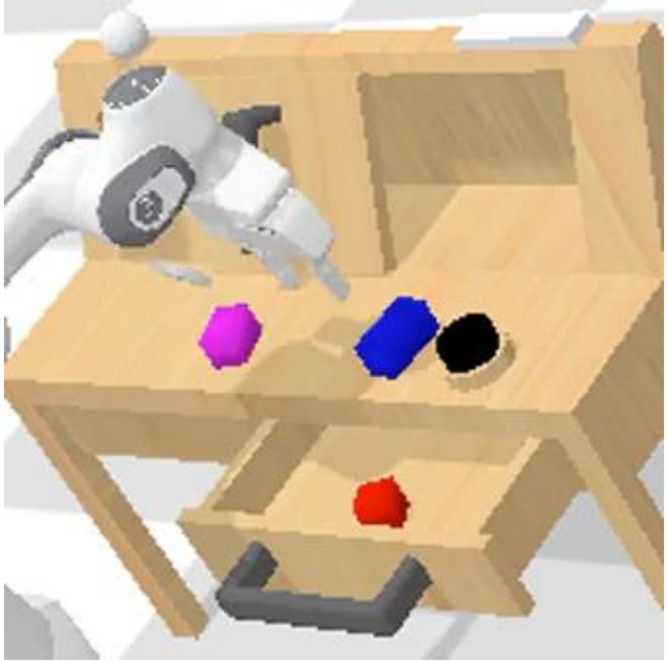
UI Control

Step 1: Collect expert demonstrations in diverse domains for training

From diverse sources, like scripted policies, humans, or RL policies

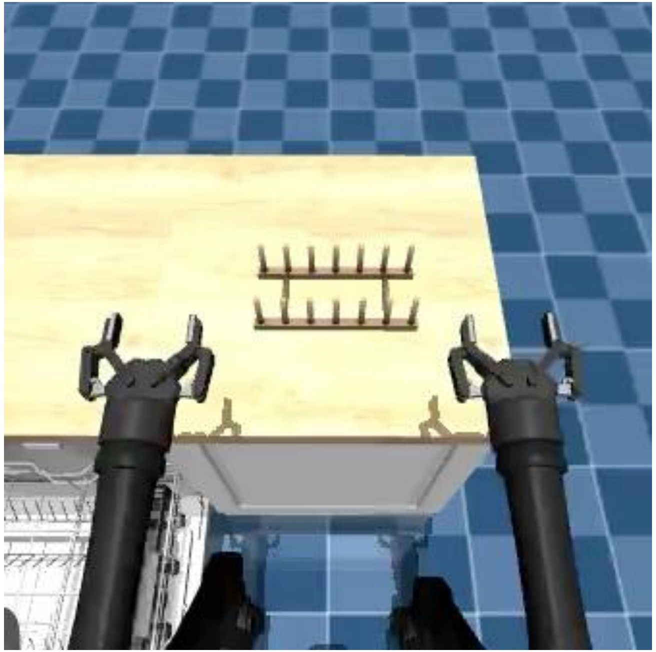
Data - Static Manipulation

“Use the block to pull the handle sideways”



“Move the purple block next to the blue block”

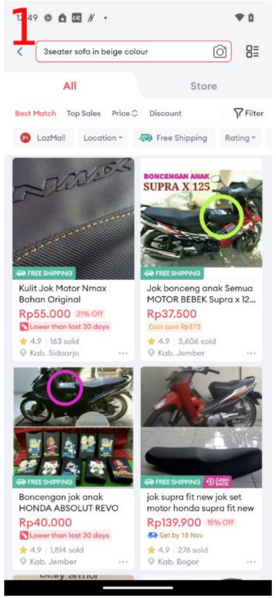
Data - Mobile Manipulation



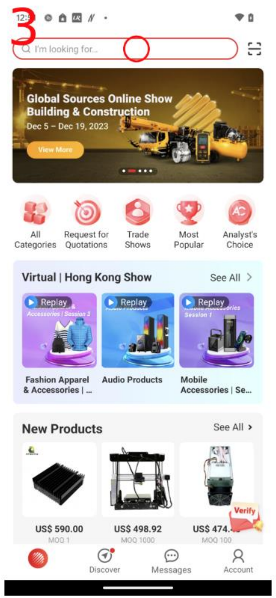
“Unload the plates from the dishwasher and place them on the rack”

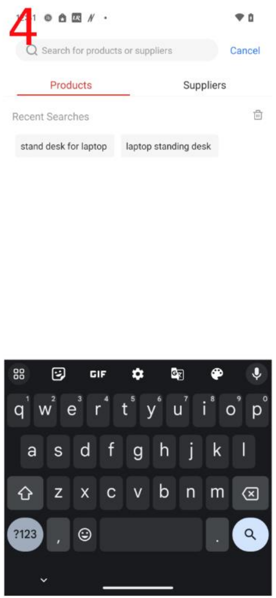
“Pick up the banana”

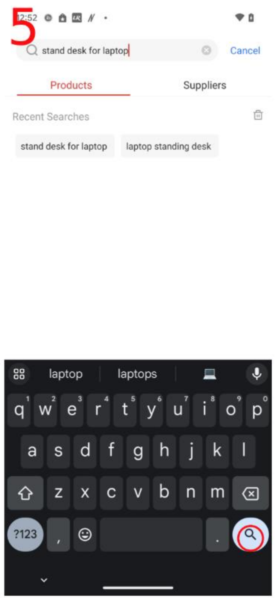
Data - UI Control

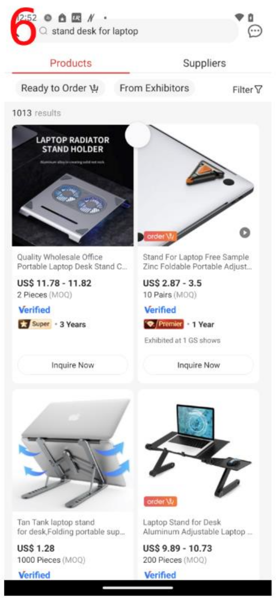
1


2


3


4


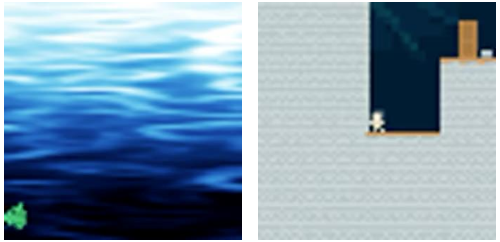
5


6


"Find me a standing desk for my laptop from the GlobalSources app"

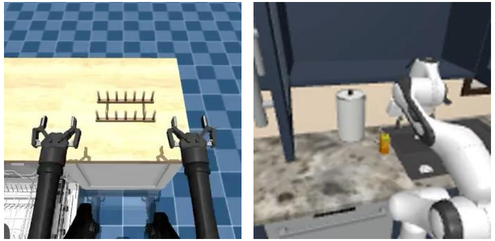
Static Manipulation

Games



Navigation

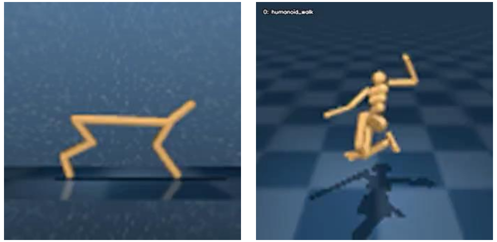
Mobile Manipulation



UI Control

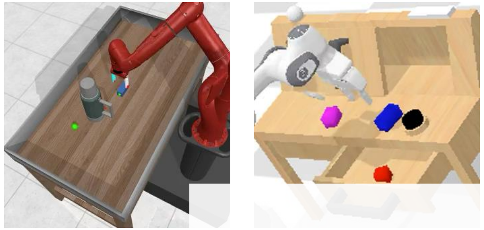
Real Robots

Character Control

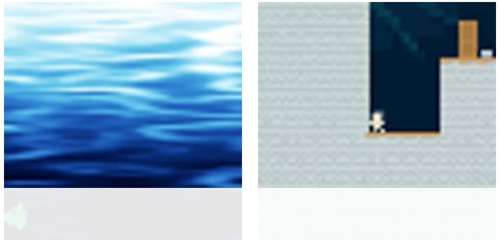


Planning

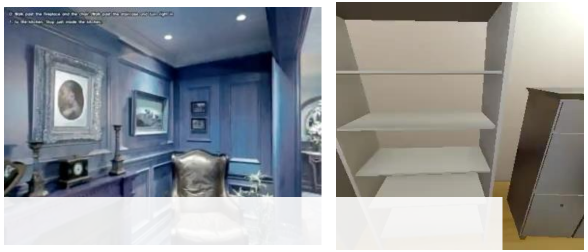
Static Manipulation



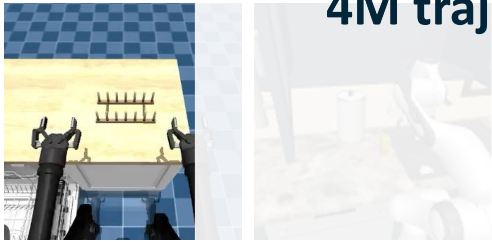
Games



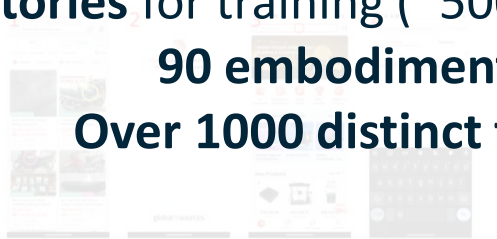
Navigation



Mobile Manipulation



UI Control



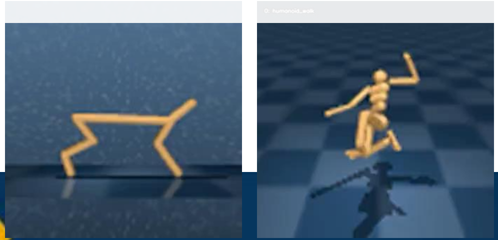
Real Robots

4M trajectories for training (~500M image/actions)

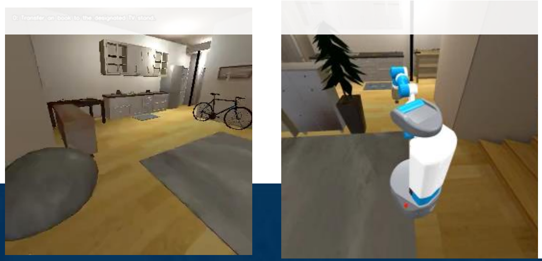
90 embodiments

Over 1000 distinct tasks

Character Control



Planning



Evaluation

New Tasks

Find an apple and put it away in the fridge.

Novel Objects
Find a pear and put it away in the fridge.

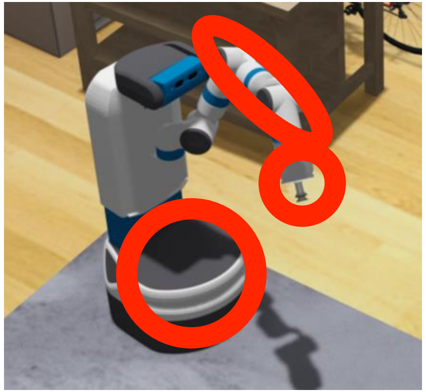
Context
I am hungry for something sweet and healthy. Put a snack for me on the table.

Spatial Relationships
Find an apple and put it in the receptacle to the right of the kitchen counter.

⋮

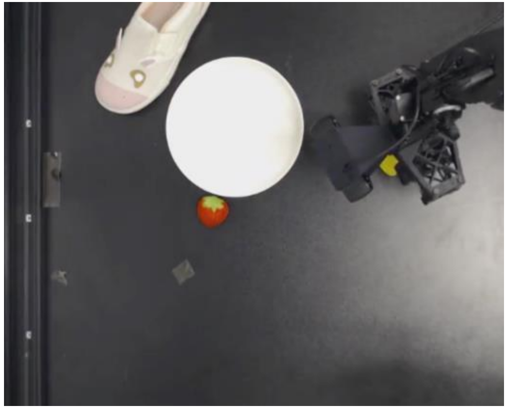
New Embodiments

New control spaces and robot types



New Environments

New platform with limited data



R

	GEA	Prior Work	# Tasks	Generalization Type
Manipulation				
Meta-World	**94.7**	84 MLLM+IL [81]^{S} 87.0 [69]^{G}	45	object positions
CALVIN (ABC →D)	**90.0**	82.4 MLLM+IL [48]^{S} 92.2 IL+pointcloud[35]	34*	instructions, background
Maniskill	13.6	6.5 IL [22]^{S} **47.8** IL+PPO [22]^{S}	5	object positions
Habitat Pick	**82.5**	29 IL [81]^{S} 81.0 RL + sim state^{S}	20	house
Habitat Place	**93.5**	95.5 RL + sim state^{S}	10	house
Video Games				
Procgen	**44.0**	25 [59]^{S}	16	background
Atari	32.7	31 [69]^{G} **85** Offline RL [41]^{S}	44	none
Navigation				
Habitat Nav	62.5	**72** [78]^{S}	10	house
BabyAI	<u>91.1</u>	**93.2** [69]^{G}	17*	instructions, grid state
UI Control				
AndroidControl	**57.3**	45 GPT-4o+SoM [93]^{G}	35*	instructions
Planning				
LangR	<u>50.0</u>	**51** MLLM+RL[81]^{S}	10*	instructions, house

Emb

Agent

Agent

Simul

an ap

Parting Thoughts

We Learned a Lot!

Deep Learning Fundamentals

Linear classification
Loss functions
Optimization
Optimizers
Backpropagation
Computation Graph
Multi-layer Perceptrons

Neural Network Components and Architectures

Hardware & software
Convolutions
Convolution Neural Networks
Pooling
Activation functions
Batch normalization
Transfer learning
Data augmentation
Architecture design
RNN/LSTMs
Attention & Transformers

Applications & Learning Algorithms

Semantic & instance Segmentation
Reinforcement Learning
Large-language Models
Variational Autoencoders
Diffusion Models
Generative Adversarial Nets
Self-supervised Learning
Vision-Language Models
VLM for Robotics

Some existing works not covered…

Current / Past

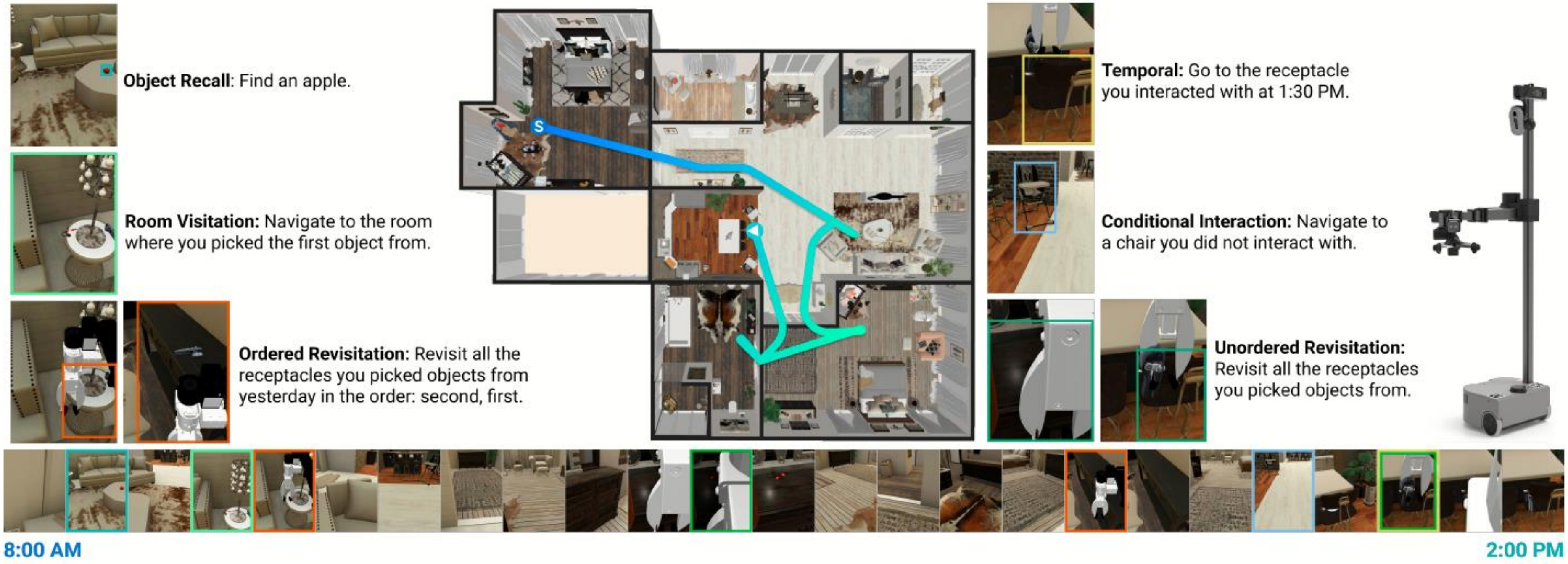
- Graph neural networks
- Meta-learning
- AutoML
- 3D perception & reconstruction / NeRFs
    - Neural Radiance Fields
- AI for Tabular data, time-series, etc.
- Beyond supervised learning: Semi-supervised, domain adaptation, zero/one/few-shot learning
- Embodied AI & Embodied question answering
- Adversarial Learning
- Continual/lifelong learning without forgetting
- World modeling, learning intuitive/physics models
- Reasoning, Planning, Search
- Neural Theorem Proving, induction & synthesis
- AI for science
- MLSys and MLOps
- Evaluation…
- Alignment
- Security

When Comparing to Humans, What's Missing?

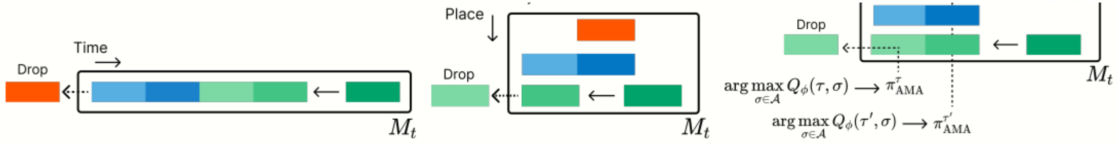
- Reasoning
  - ~~What does it mean for a neural network to "think" longer?~~
  - Chain-of-thought probably still off from how humans do it!
- Memory
- Planning, Search
- Deep integration of concepts and modalities
- Cognitive Architecture?

MA-LMM: Memory-Au
Model for Long-Ter

Spatially-Aware Transformer for Embodied Agents



Graves et. al, Neural Turing Machines



(a) FIFO Memory **(b)** Place Memory **(c)** Adaptive Memory Allocator (AMA)

(a) Maze navigation task

Prompt

```
bos
start  0 2
goal   1 0
wall   1 2
wall   2 0
eos
```

Response

```
bos
plan  0 2
plan  0 1
plan  1 1
plan  1 0
eos
```

(b) Tokenization of a planning task and its solution

A* planning algorithm

Require: Start node n_{start} and goal node n_{goal}.

$\mathcal{S}_{\text{closed}} \leftarrow \{\}$
$\mathcal{S}_{\text{frontier}} \leftarrow \{n_{\text{start}}\}$
while $|\mathcal{S}_{\text{frontier}}| > 0$ **do**
  $n_{\text{curr}} = \arg\min_{n \in \mathcal{S}_{\text{frontier}}} \text{cost}(n)$
  $\mathcal{S}_{\text{closed}} \leftarrow \mathcal{S}_{\text{closed}} \cup \{n_{\text{curr}}\}$
  for $n_{\text{child}} \in \text{children}(n_{\text{curr}})$ **do**
    if $\text{pos}(n) = \text{pos}(n_{\text{child}})$ for any $n \in \mathcal{S}_{\text{closed}} \cup \mathcal{S}_{\text{frontier}}$ **then**
      if $\text{cost}(n) \leq \text{cost}(n_{\text{child}})$ **then**
        continue
      end if
    end if
    Set $\text{parent}(n_{\text{child}}) \leftarrow n_{\text{curr}}$
    $\mathcal{S}_{\text{frontier}} \leftarrow \mathcal{S}_{\text{frontier}} \cup \{n_{\text{child}}\}$
  end for
end while
Compute and return plan by recursing on parents of n_{curr}.

Tokenization of algorithm execution

```
       bos
       create  0 2 c3 c0   Add node to frontier
       close   0 2 c3 c0   Add node to closed set
       create  0 1 c2 c1   Heuristic of node
       create  1 2 c2 c1   Cost from start
Trace  close   0 1 c2 c1
       create  1 1 c1 c2
       close   1 1 c1 c2
       create  1 0 c0 c3
       close   1 0 c0 c3
       plan    0 2
       plan    0 1
Plan   plan    1 1
       plan    1 0
       eos
```

(c) A^*'s execution when solving a planning task is logged into an execution trace

via

ctures
work,
ormer,
while
ecoder
ed via
l plan.
n task
tudies
ct the
'e also
koban

Correspondence: {lucaslehnert, yuandong}@meta.com

Things to Watch out For

- Research is cyclical
  - SVMs, boosting, probabilistic graphical models & Bayes Nets, Structural Learning, Sparse Coding, Deep Learning
  - Deep learning is unique in its depth and breadth, but...
  - Deep learning may be improved, reinvented, combined, overtaken
- Learn fundamentals for techniques across the field:
  - Know the span of ML techniques and choose the ones that fit your problem!
  - **Be responsible** in 1) how you use it, 2) promises you make and how you convey it
- Try to understand landscape of the field
  - Look out for what is coming up next, not where we are
- Have fun!

Open Discussion

Thank you!