Day 20

Speaker: Prof Dhruv Batra, FAIR & Georgia Tech

Title: From Disembodied to Embodied AI.

From Disembodied to Embodied AI

Dhruv Batra

Georgia Tech

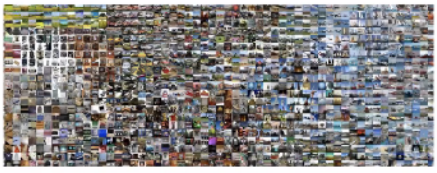
facebook Artificial Intelligence Research

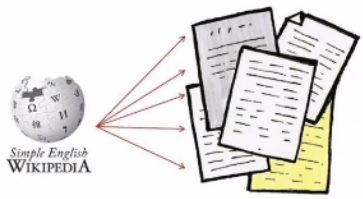
Habitat

What is Embodied AI?

Disembodied AI

Static datasets





Embodied AI

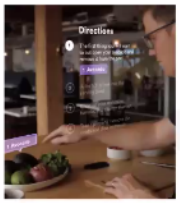
Agents acting in environments

Is there smoke in any room around you?

Yes, in one room.

Go there and look for people.

...



Where did I leave my keys?

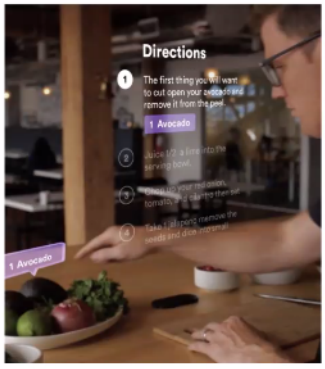
On your coffee table.

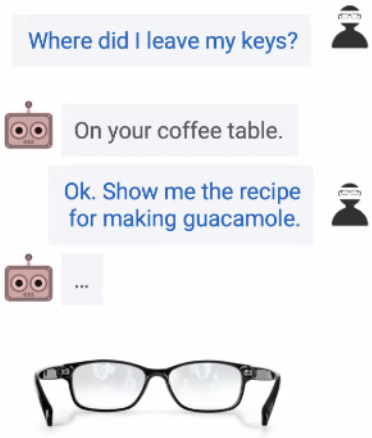
Show me the recipe for making guacamole.

...

Embodied Agents in Real & Virtual Environments

Virtual agent
on smart glasses
taking virtual actions
talking to humans
in natural language





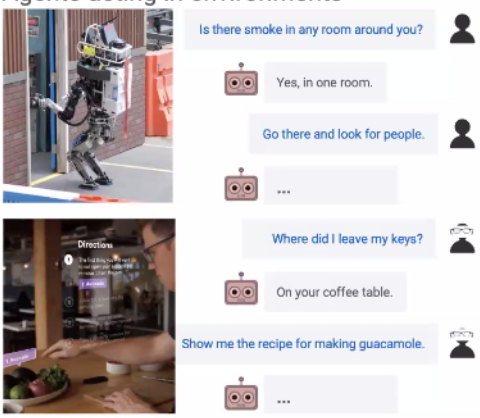
Plan for this talk

- Simulation as a test-bed
  - Habitat: A Platform for Embodied AI [ICCV 19]
  - Habitat 2.0: Training Home Assistants to Rearrange their Habitat [Arxiv 21]
- Emergence of intelligent behavior at scale
  - DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames [ICLR 20]

The Challenge of Embodied AI

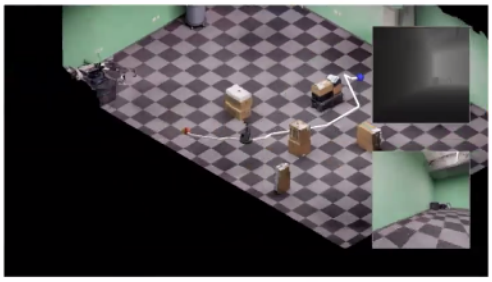
- Hardware doesn't exist, or
- Reality is
  - Slow
  - Dangerous
  - Expensive
  - Not easy reproducible/controllable

Agents acting in environments

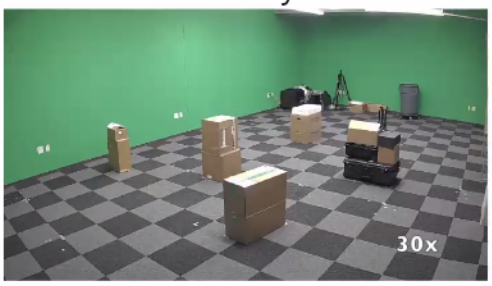


Common Solution: Sim2Real

Simulation

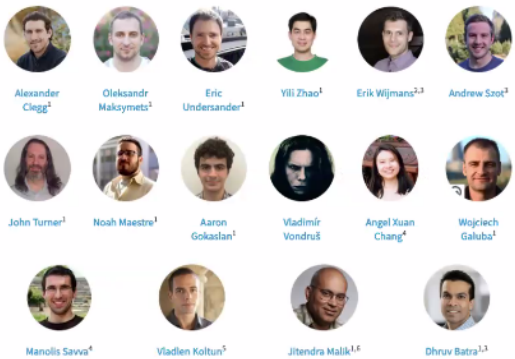


Reality

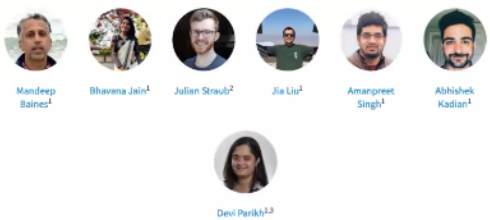


Habitat: A Platform for Embodied AI Research

Team: Current Contributors



Team: Past Contributors



facebook Artificial Intelligence Research 1 | facebook Reality Labs 2 | Georgia Tech 3 | SFU 4 | intel 5 | Berkeley UNIVERSITY OF CALIFORNIA 6

Our Vision

- Create the ImageNet/COCO/VQA of Embodied AI
  - Dataset → Simulator → Task → Benchmark Challenge

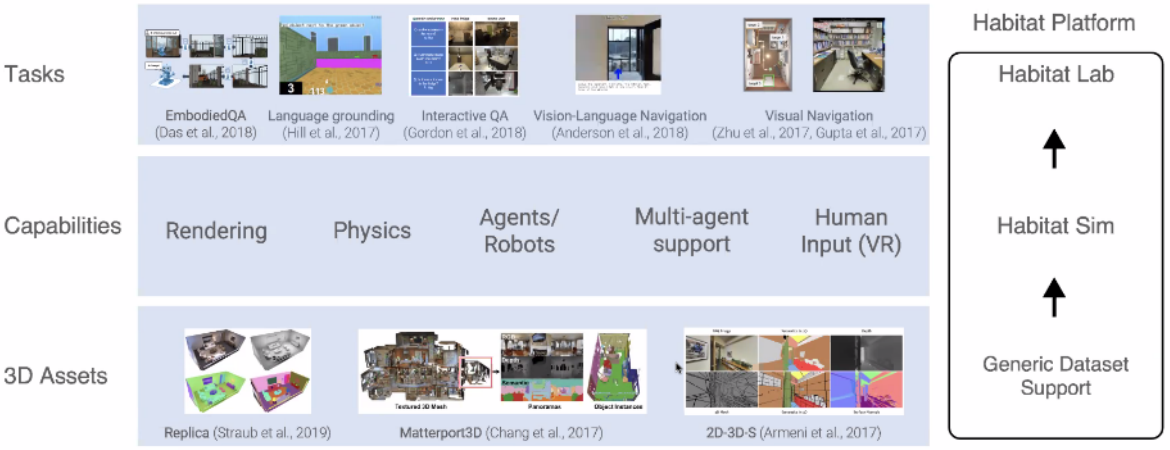
PASCAL2 Pattern Analysis, Statistical Modelling and Computational Learning

COCO Common Objects in Context

IMAGENET

VQA Visual Question Answering

Habitat: A Simulation Platform for Embodied AI



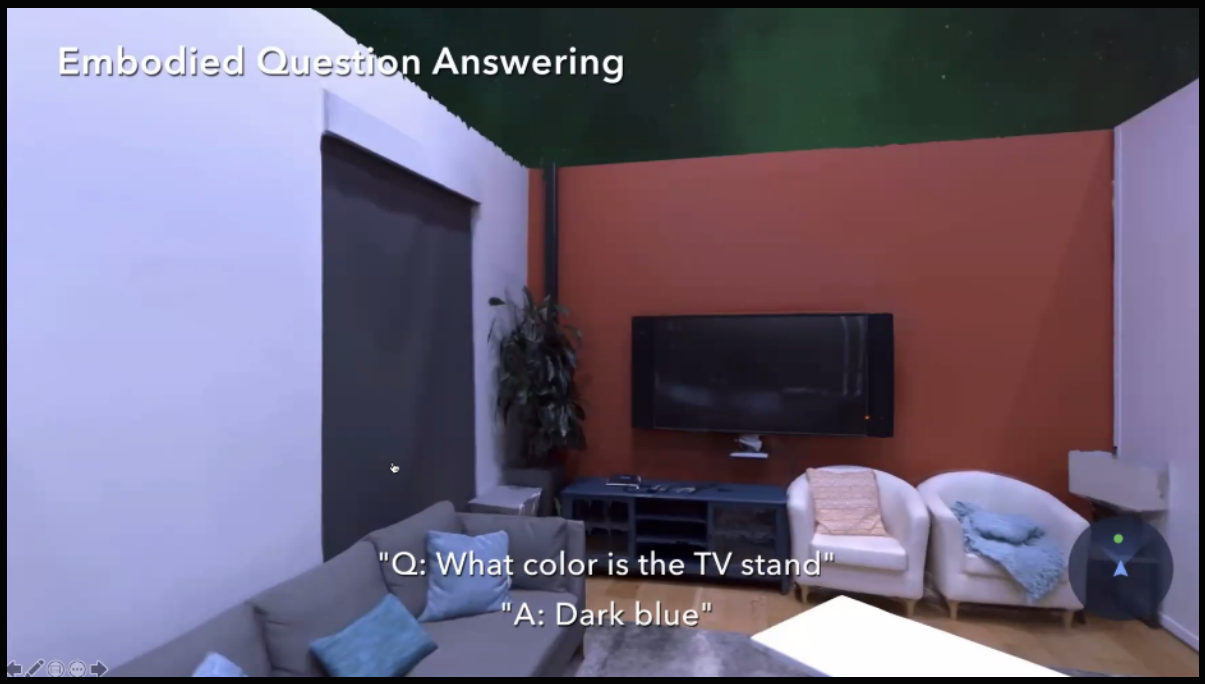
Object datasets: YCB, Ikea

Yale-CMU-Berkeley dataset for robotic manipulation research

Berk Calli[1], Arjun Singh[2], James Bruce[3], Aaron Walsman[3], Kurt Konolige[3], Siddhartha Srinivasa[1], Pieter Abbeel[2] and Aaron M Dollar[1]

Dataset for IKEA 3D models and aligned images

Joseph J. Lim, Hamed Pirsiavash, and Antonio Torralba

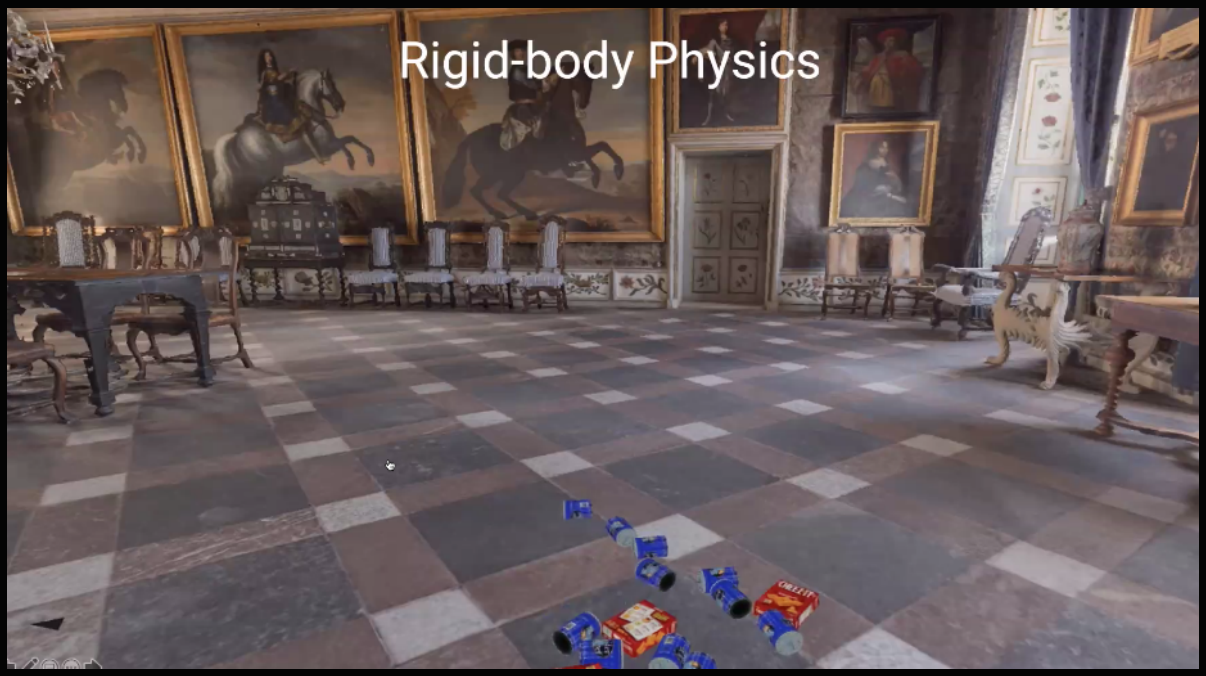


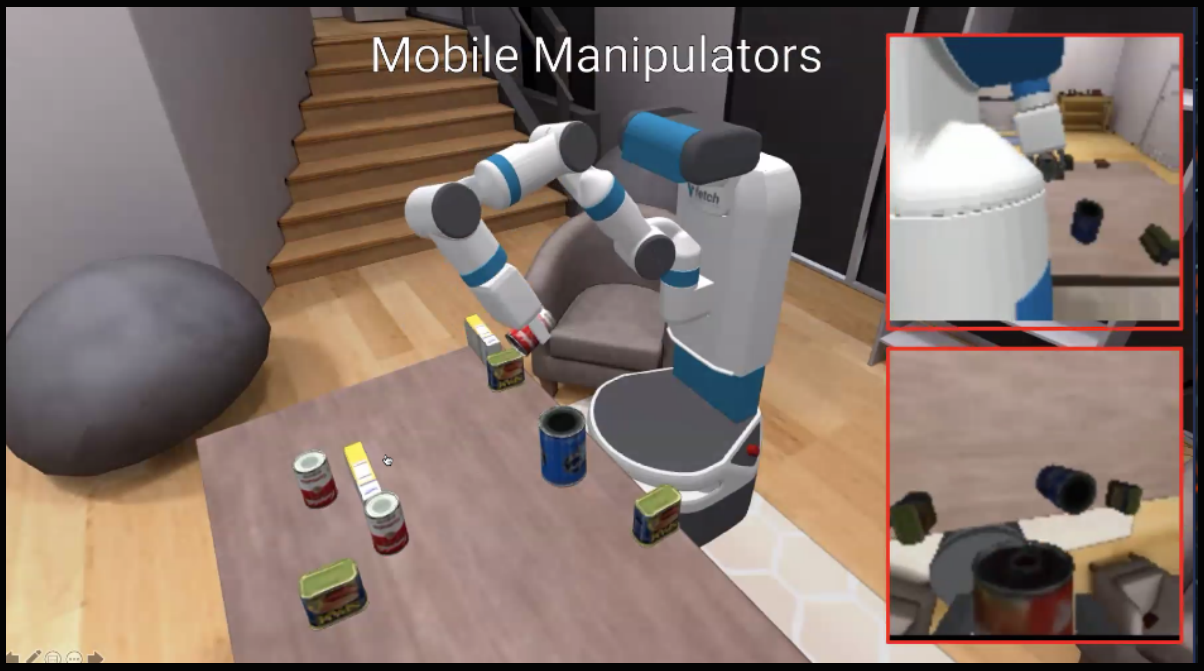
Navigation

Habitat 1.0
[ICCV '19]

Mobile Manipulation

Habitat 2.0


Rigid-body Physics


Mobile Manipulators


Home Assistants

Tidy House

- Move 5 objects from start to goal location

Prepare Groceries

- Object 1 from fridge to counter
- Object 2 from fridge to table
- Object 3 from counter to fridge

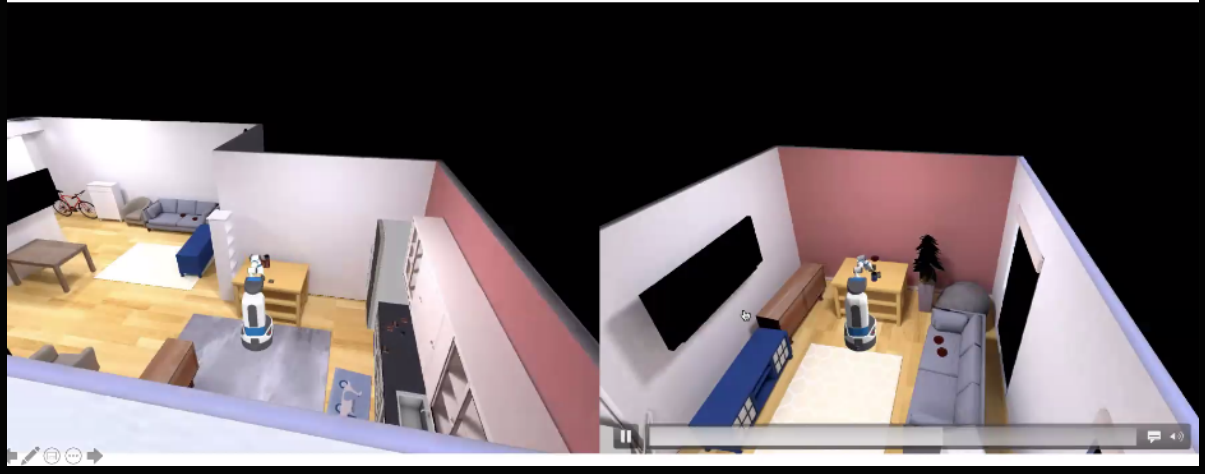
Set Table

- Bowl from drawer to table
- Apple from fridge into bowl
- (Fridge closed)
- (Drawer closed)

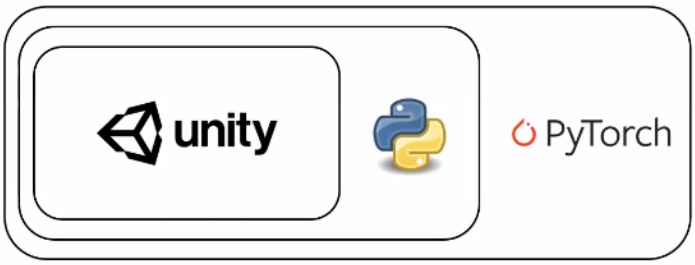
Generalization to New Objects

Generalization to New Layouts

Training Layout **Testing Layout**



Why not use a video game engine?

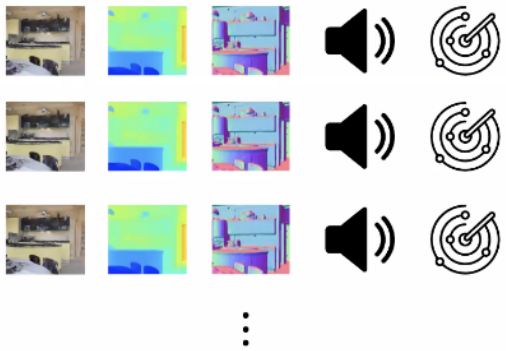


AI2-THOR [Kolve et al. 2017]
architecture example sketch

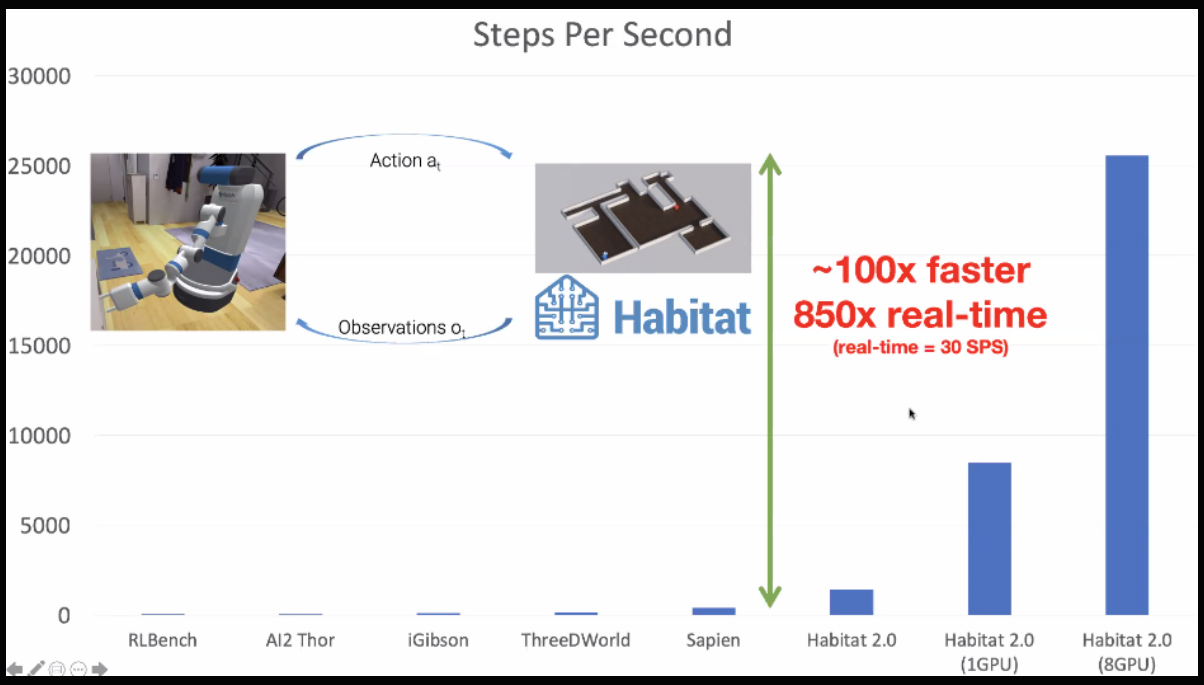
Human needs vs AI needs

Human: 1080p @ 60Hz

AI: 84x84 @ 1000+ Hz



Steps Per Second

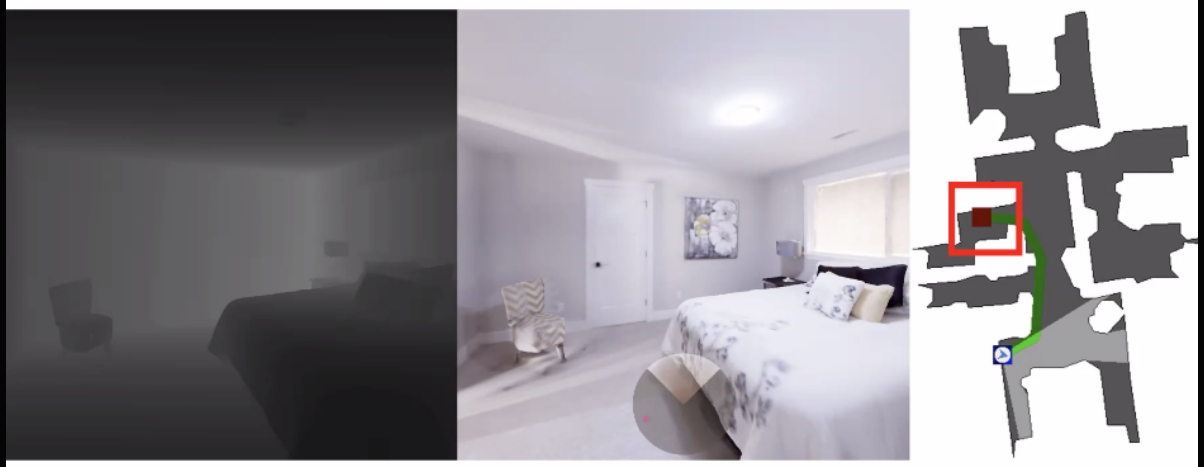


Why does speed matter?

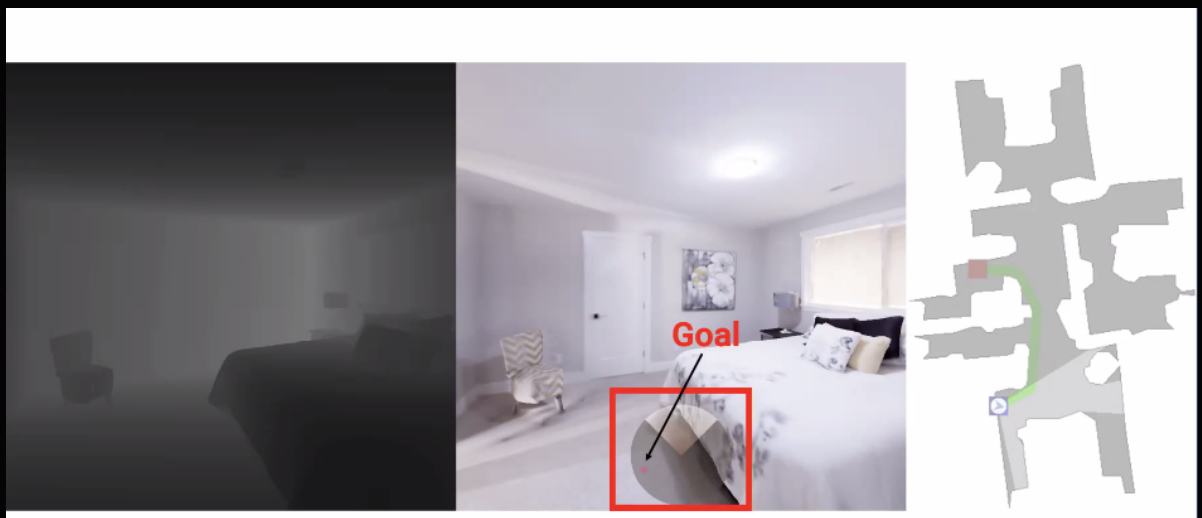
Because you can now run experiments you couldn't before.

Plan for this talk

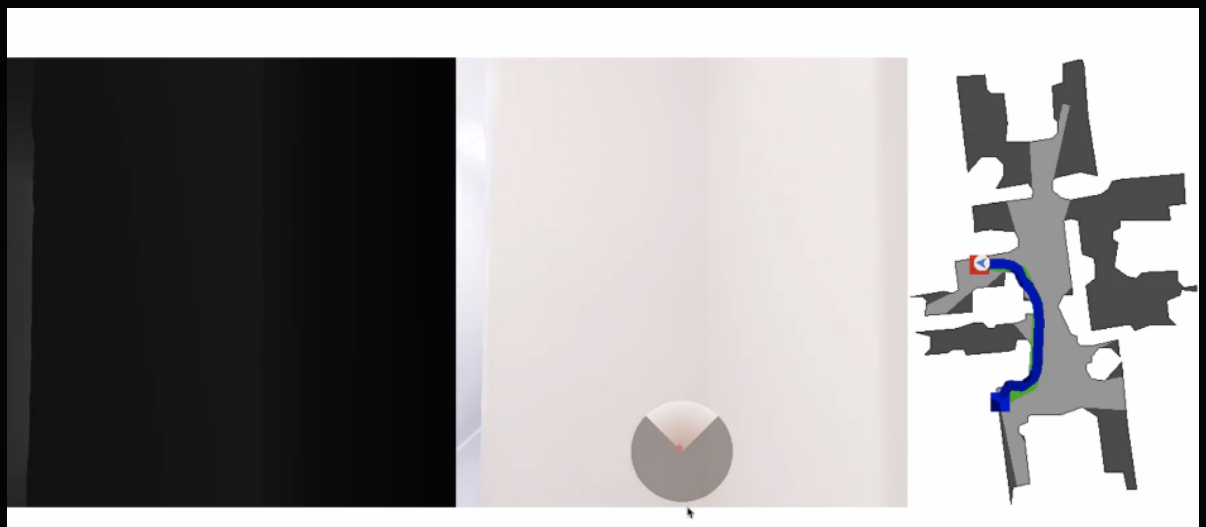
- Simulation as a test-bed
  - Habitat: A Platform for Embodied AI [ICCV 19]
  - Habitat 2.0: Training Home Assistants to Rearrange their Habitat [Arxiv 21]
- Emergence of intelligent behavior at scale
  - DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames [ICLR 20]



Depth RGB and GPS+Compass Top Down Map

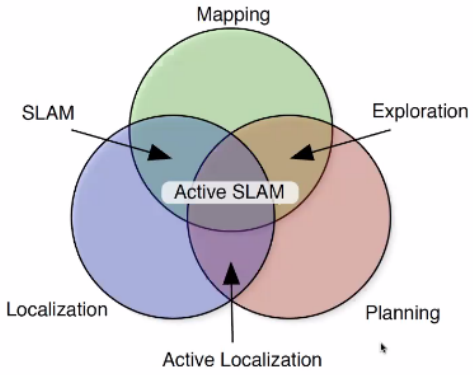


Depth RGB and GPS+Compass Top Down Map



Depth RGB and GPS+Compass Top Down Map

Classical Robots Pipeline



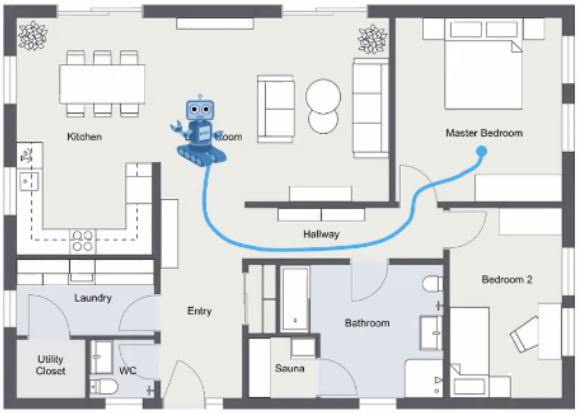


Image Credit: Nathaniel Fairfield, PhD Thesis 2009

Decentralized Distributed PPO:

Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames


Erik Wijmans


Abhishek Kadian


Ari Morcos


Stefan Lee


Irfan Essa


Devi Parikh


Manolis Savva


Dhruv Batra

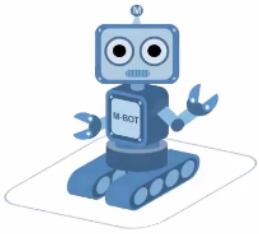
facebook Artificial Intelligence Research

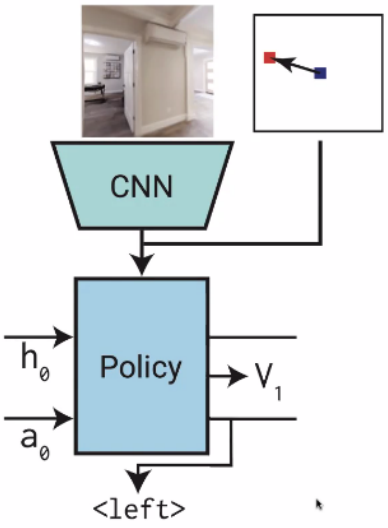
Georgia Tech

SFU

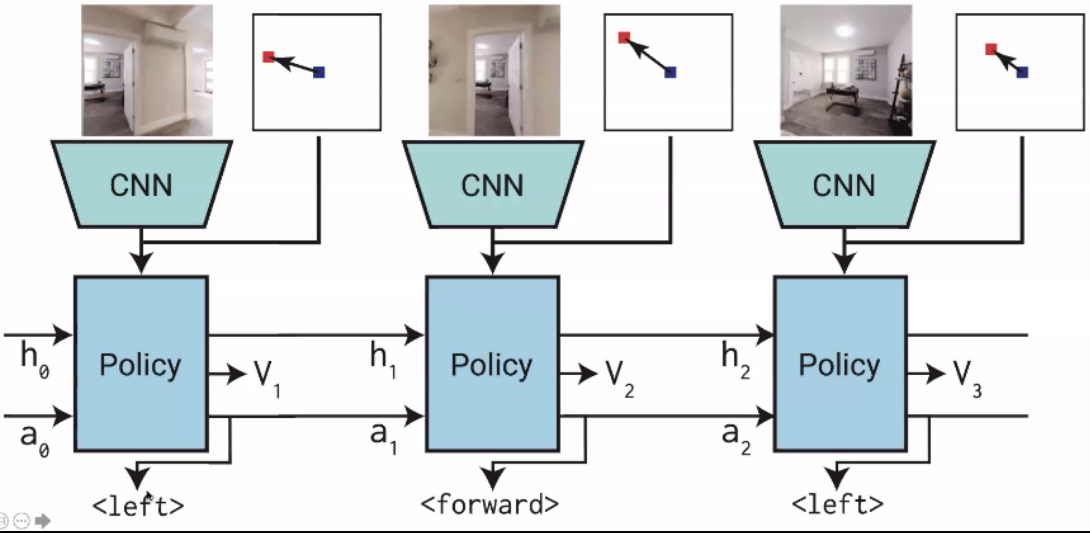
Oregon State University

Agent and Model Design



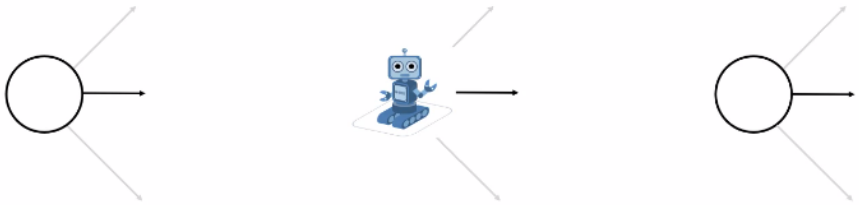


Agent and Model Design



Agent and Model Design

$$r_t = \begin{cases} s & \text{if goal is reached} \\ \underbrace{(d_{t-1} - d_t)}_{\text{reward shaping}} - \underbrace{\lambda}_{\text{efficiency}} & \text{otherwise} \end{cases}$$



Radical Empiricism

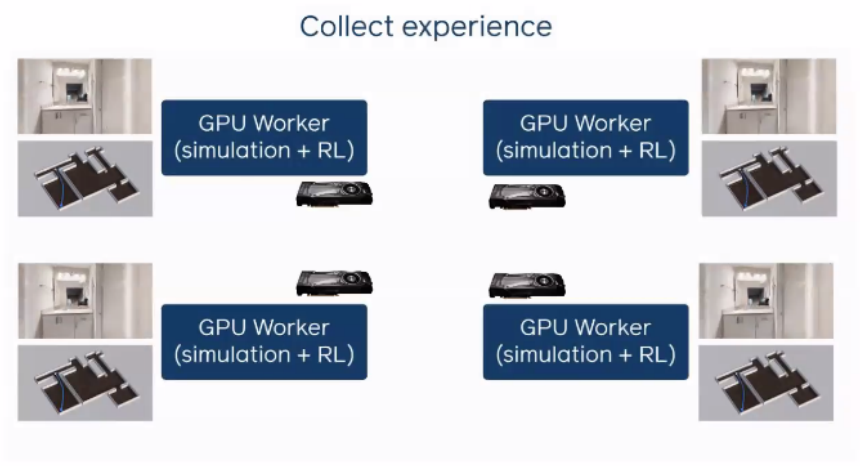
- No task-specific modules
  - No SLAM, no mapping, no path planning
- No domain-specific inductive biases
  - No spatial memory, knowledge of projective geometry or 3D
- No additional learning signals
  - No mapping supervision, auxiliary tasks
  - No expert demonstrations or imitation learning
  - No pre-training of representations
  - No look-ahead search trees
- No learning tricks
  - No curriculum, replay buffer, reset to partially-successful-states

Radical Empiricism

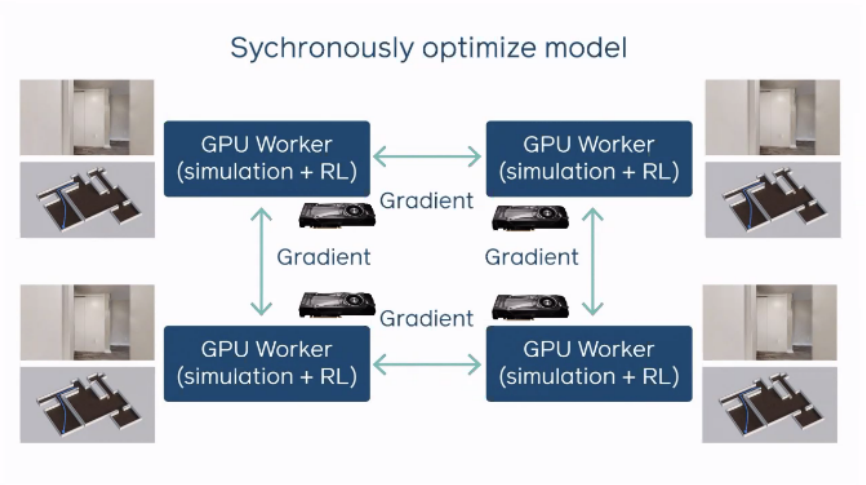
- On-policy Episodic Reinforcement Learning (on steroids)
  - Dense rewards

How far can we **scale** model-free RL for embodied visual navigation?

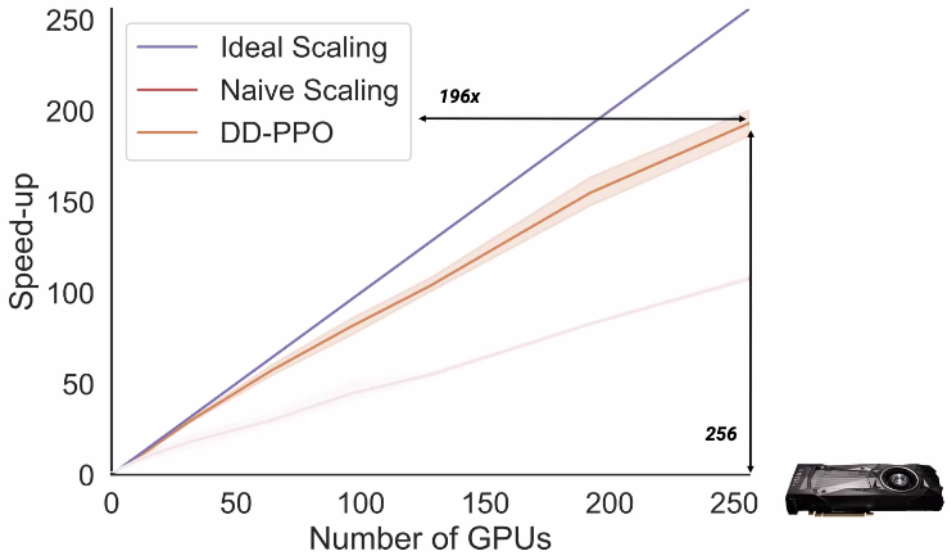
Decentralized Distributed PPO

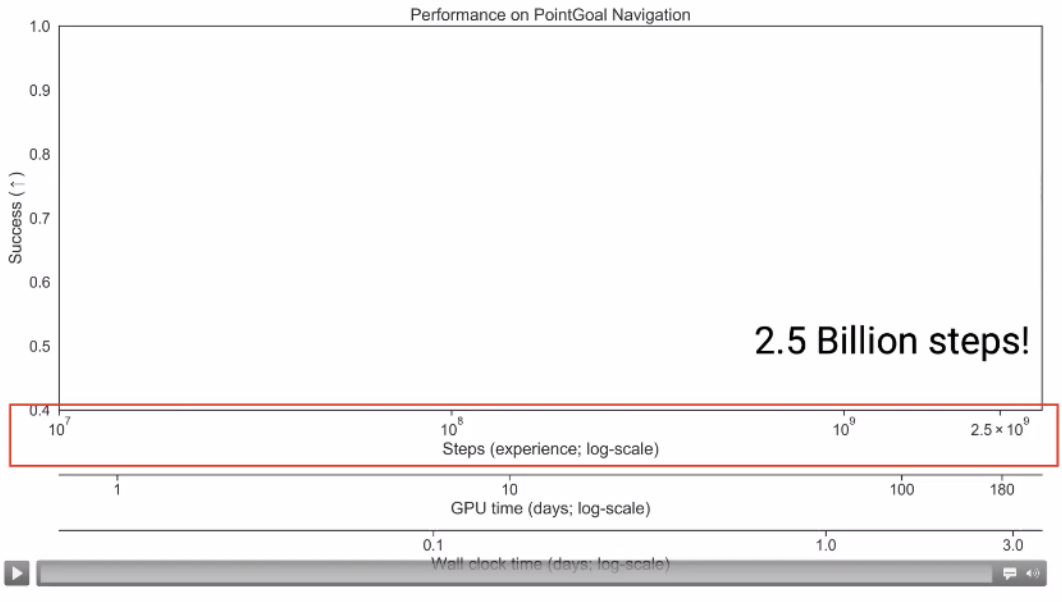


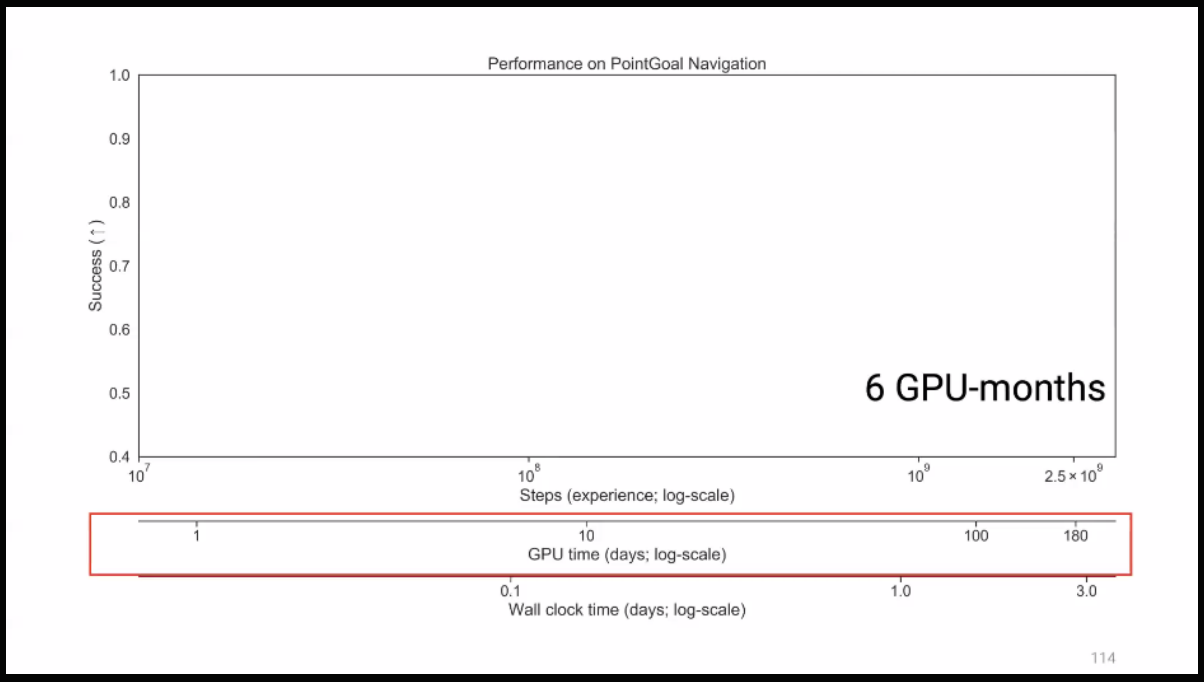
Decentralized Distributed PPO



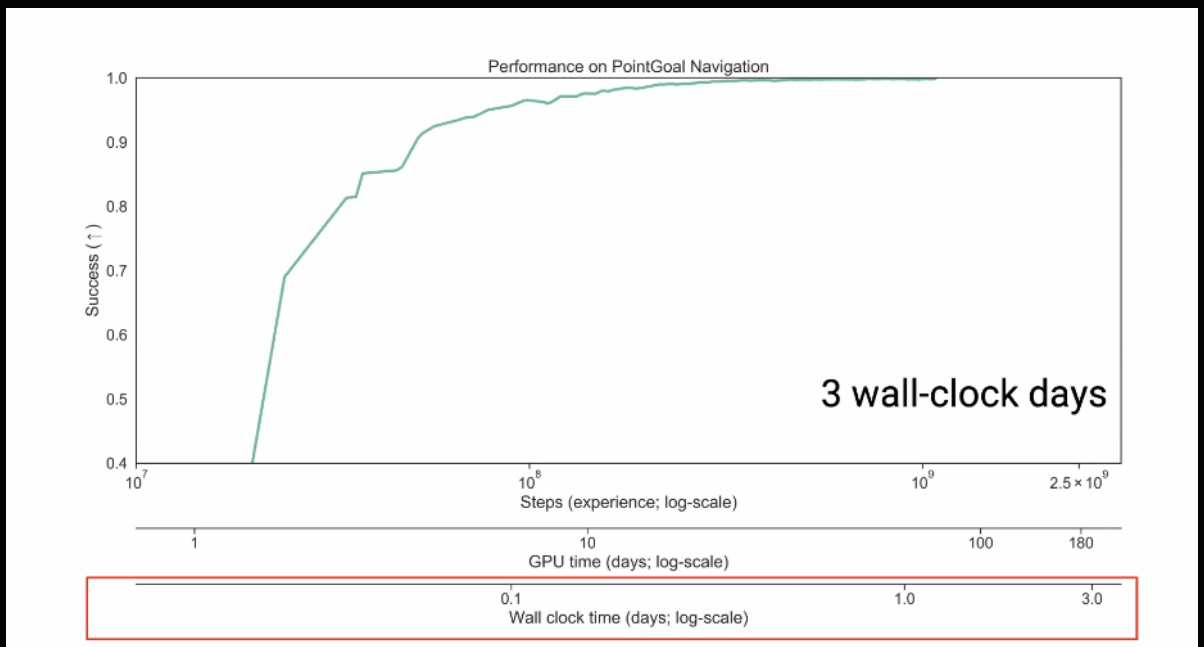
How does DD-PPO scale? Near-linearly!



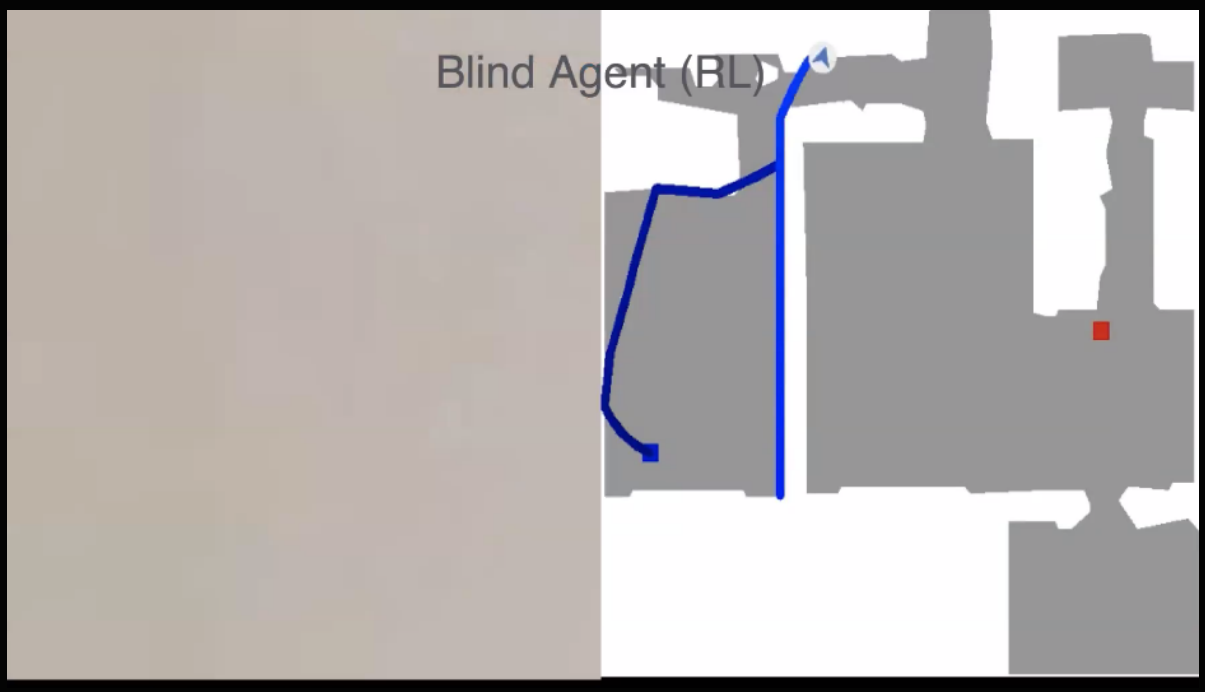


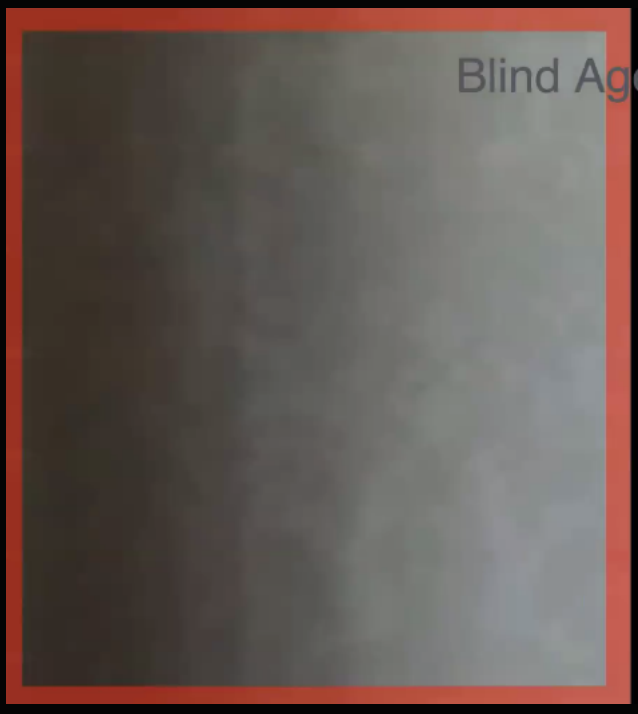

Performance on PointGoal Navigation
Success (↑)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
6 GPU-months
10^7
10^8
10^9
2.5×10^9
Steps (experience; log-scale)
1
10
100
180
GPU time (days; log-scale)
0.1
1.0
3.0
Wall clock time (days; log-scale)

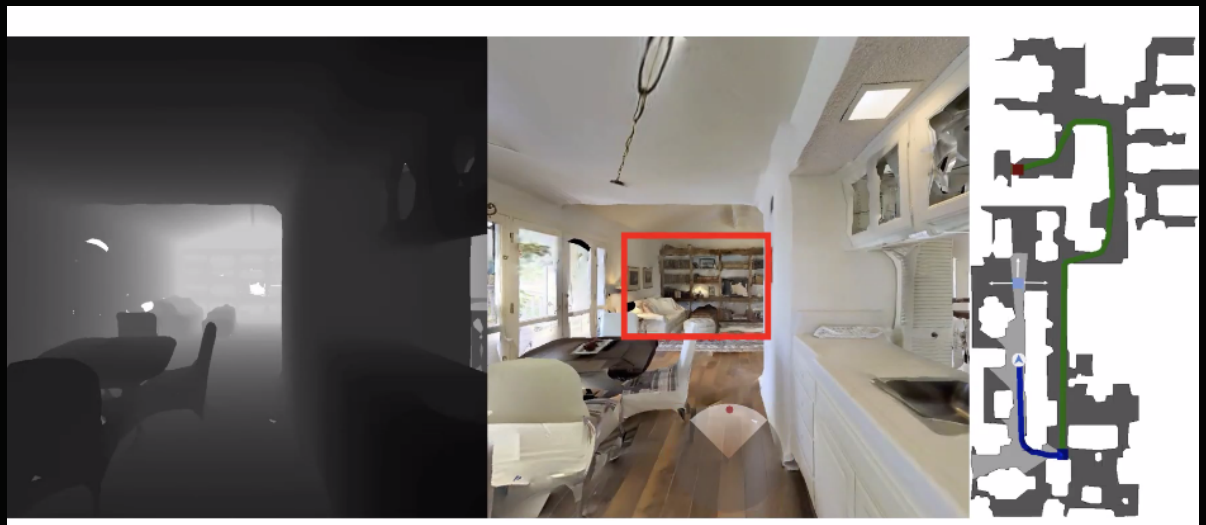
114


Performance on PointGoal Navigation
Success (↑)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
3 wall-clock days
10^7
10^8
10^9
2.5×10^9
Steps (experience; log-scale)
1
10
100
180
GPU time (days; log-scale)
0.1
1.0
3.0
Wall clock time (days; log-scale)

115


Blind Agent (RL)





Depth RGB and GPS+Compass Top Down Map

However, there is wall straight

Depth RGB and GPS+Compass Top Down Map

Depth RGB and GPS+Compass Top Down Map

How far can we scale model-free RL for embodied visual navigation?

- Surprisingly far!
  - PointNav w/ RGBD + GPS+Compass: Near-perfect
- Not far enough yet (but rapidly improving) 2019 2020 2021
  - PointNav w/ RGBD (no localization): 15% → 28% → 96%

How far can we scale model-free RL for embodied visual navigation?

- Surprisingly far!
  - PointNav w/ RGBD + GPS+Compass: Near-perfect

- Not far enough yet (but rapidly improving)

	2019	2020	2021
PointNav w/ RGBD (no localization):	15%	→ 28%	→ 96%
Object-Goal Navigation:	3-5%	→ 21%	→ 30%