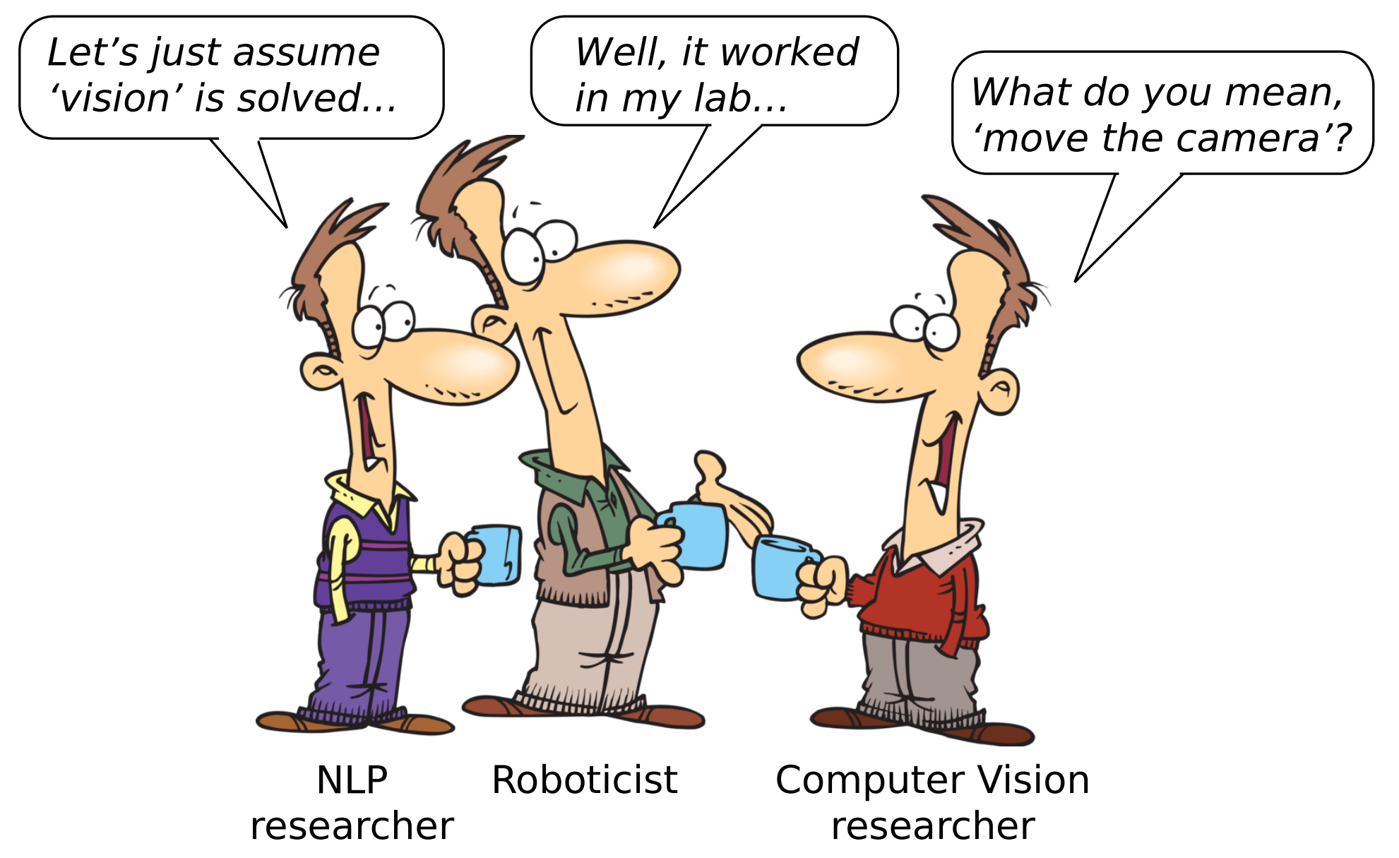
Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments

Peter Anderson[1†], Qi Wu[2], Damien Teney[2], Jake Bruce[3], Mark Johnson[4], Niko Sünderhauf[3], Ian Reid[2], Stephen Gould[1], Anton van den Hengel[2]

[1]Australian National University, [2]University of Adelaide, [3]Queensland University of Technology, [4]Macquarie University, [†]Transitioning to Georgia Tech

1. Motivation

- Connect language and vision to actions.
- Recent availability of 3D reconstructions at large scale is an enabler for research on embodied agents.
- Timely to refocus on the intersection of computer vision, NLP and robotics.


Let's just assume 'vision' is solved...
Well, it worked in my lab...
What do you mean, 'move the camera'?
NLP researcher
Roboticist
Computer Vision researcher

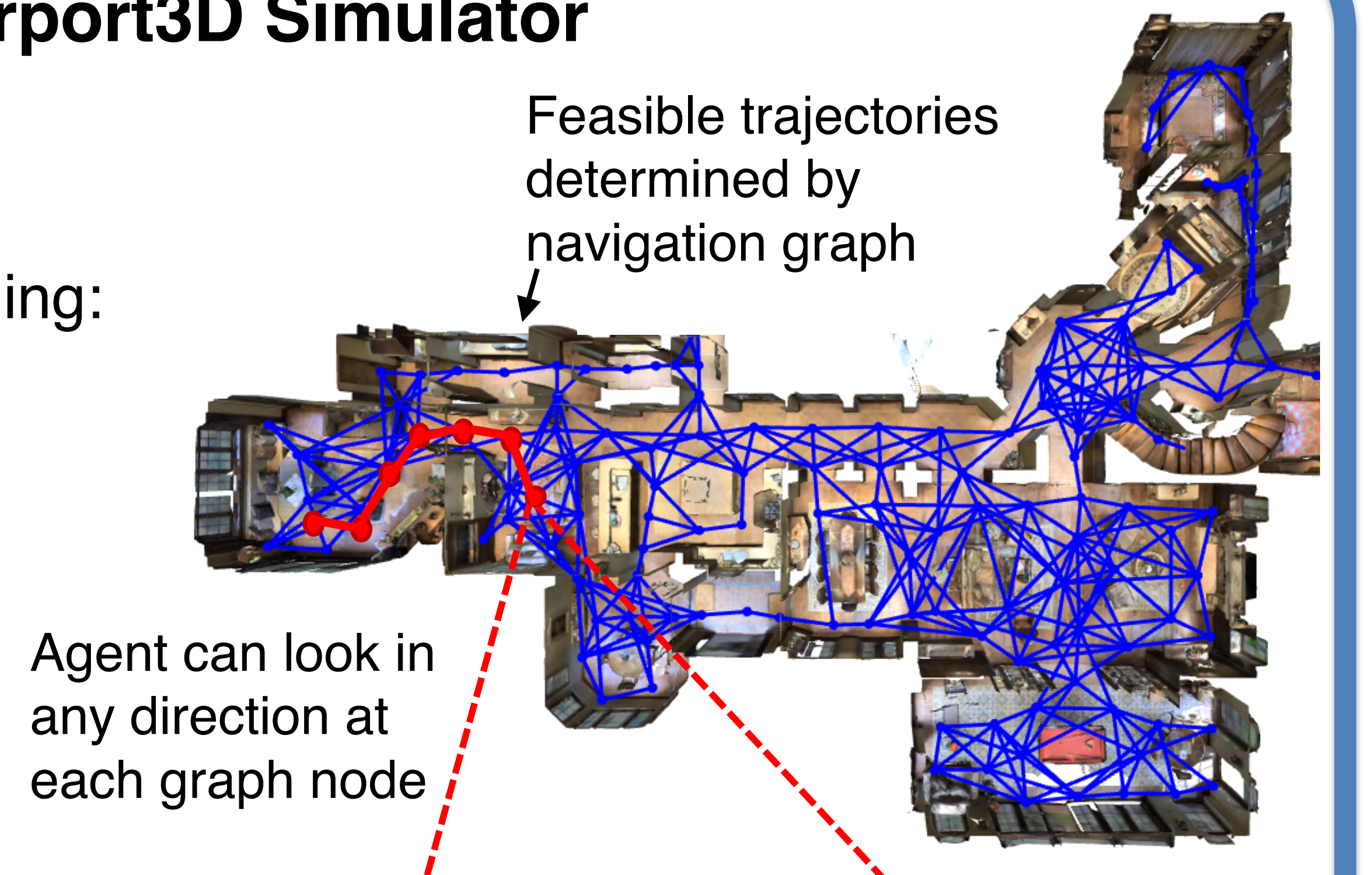
2. Vision-and-Language Navigation

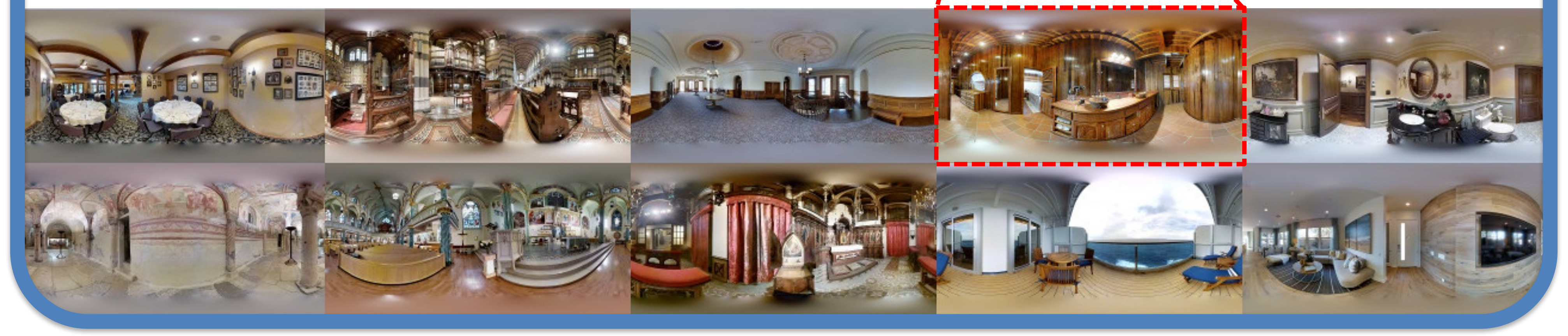
- Given a natural language navigation instruction, navigate through a real environment to find the goal location.

`Instruction:` Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

3. Matterport3D Simulator

- Simulator for embodied visual agents, based on the Matterport3D dataset[5] containing:
  - 10,800 panoramas
  - 90 diverse buildings
- Discrete motion but with continuous camera control and real images.


Feasible trajectories determined by navigation graph
Agent can look in any direction at each graph node



[5] Chang *et al.* 3DV, 2017

4. Room-to-Room (R2R) Navigation Dataset

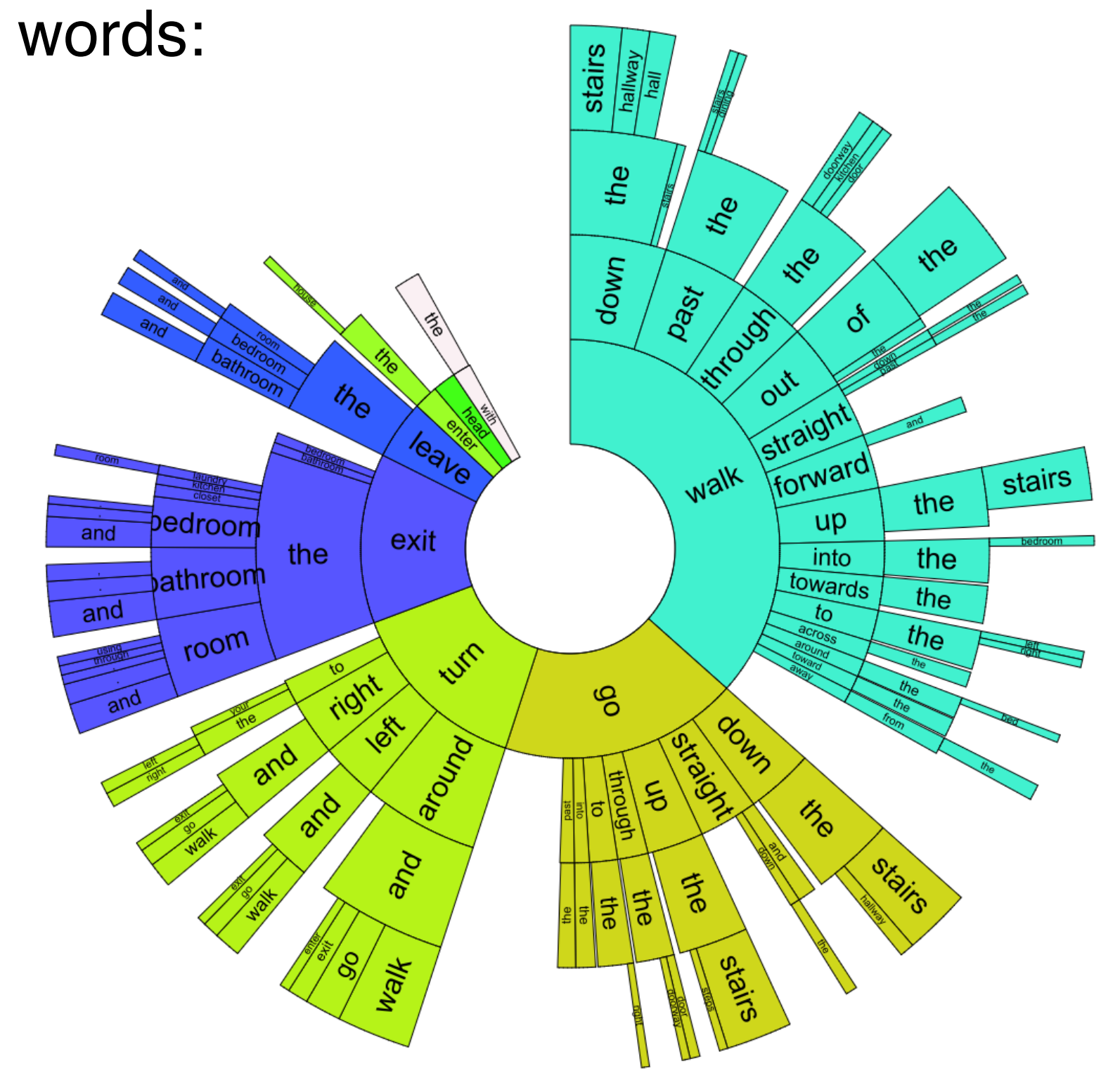
Data Collection[6]:

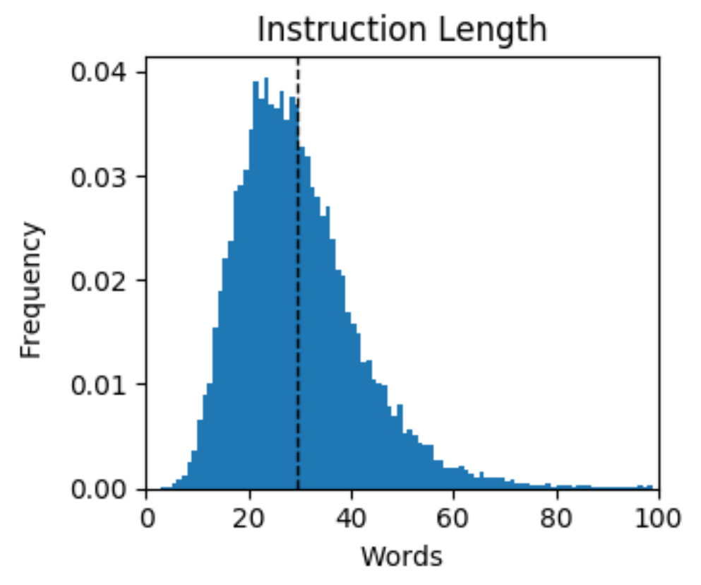
- Sampled 7,189 shortest paths between locations (mostly) in different rooms.
- Collected 21,567 navigation instructions (3 per path) using crowd workers and a WebGL interface (1,600 hours).

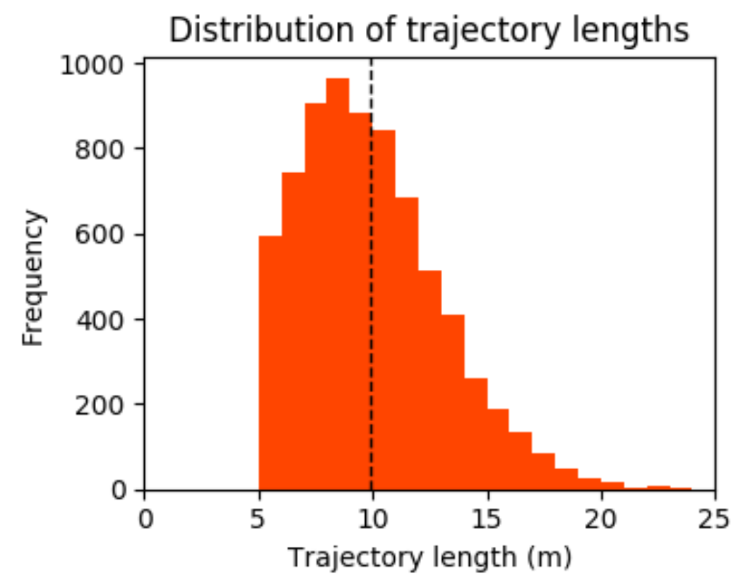
Environment splits:

- 61 training / val-seen, 11 val-unseen, 18 test (unseen).

Distribution of navigation instructions based on their first words:







Examples of new vocabulary encountered in unseen environments:

hieroglyphs

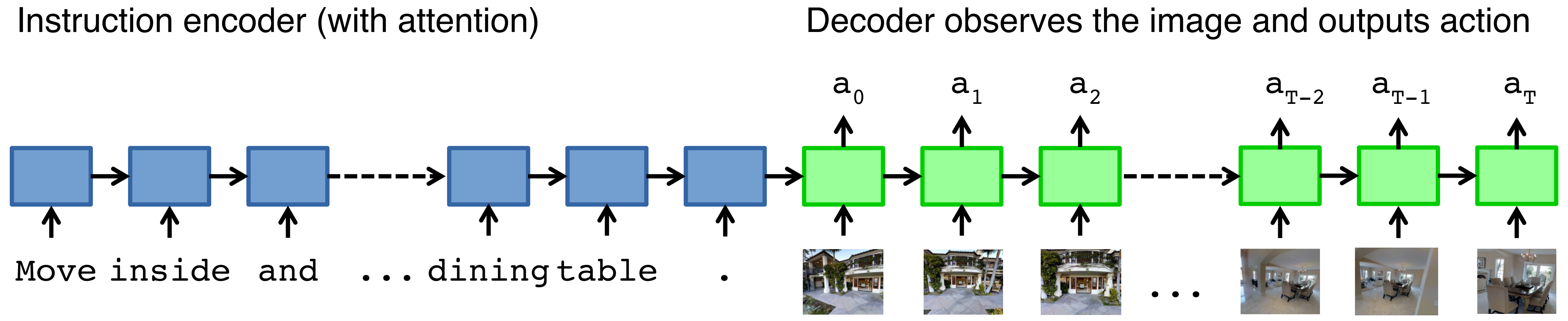
Squiggle painting

mannequins

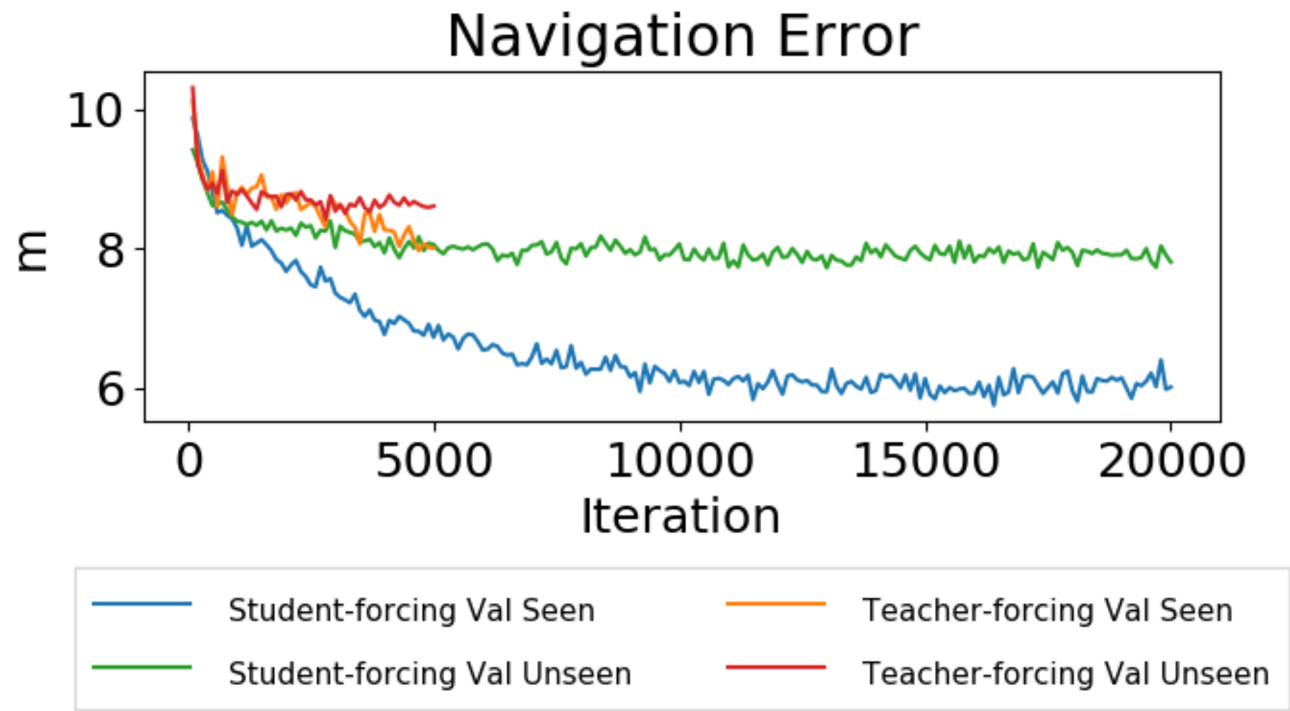
teapot

[6] Data collection was generously supported by a Facebook ParlAI Research Award.

5. Baseline Seq2Seq Agent


Instruction encoder (with attention)
Decoder observes the image and outputs action
a_0
a_1
a_2
a_{T-2}
a_{T-1}
a_T
Move inside and ... dining table .
...

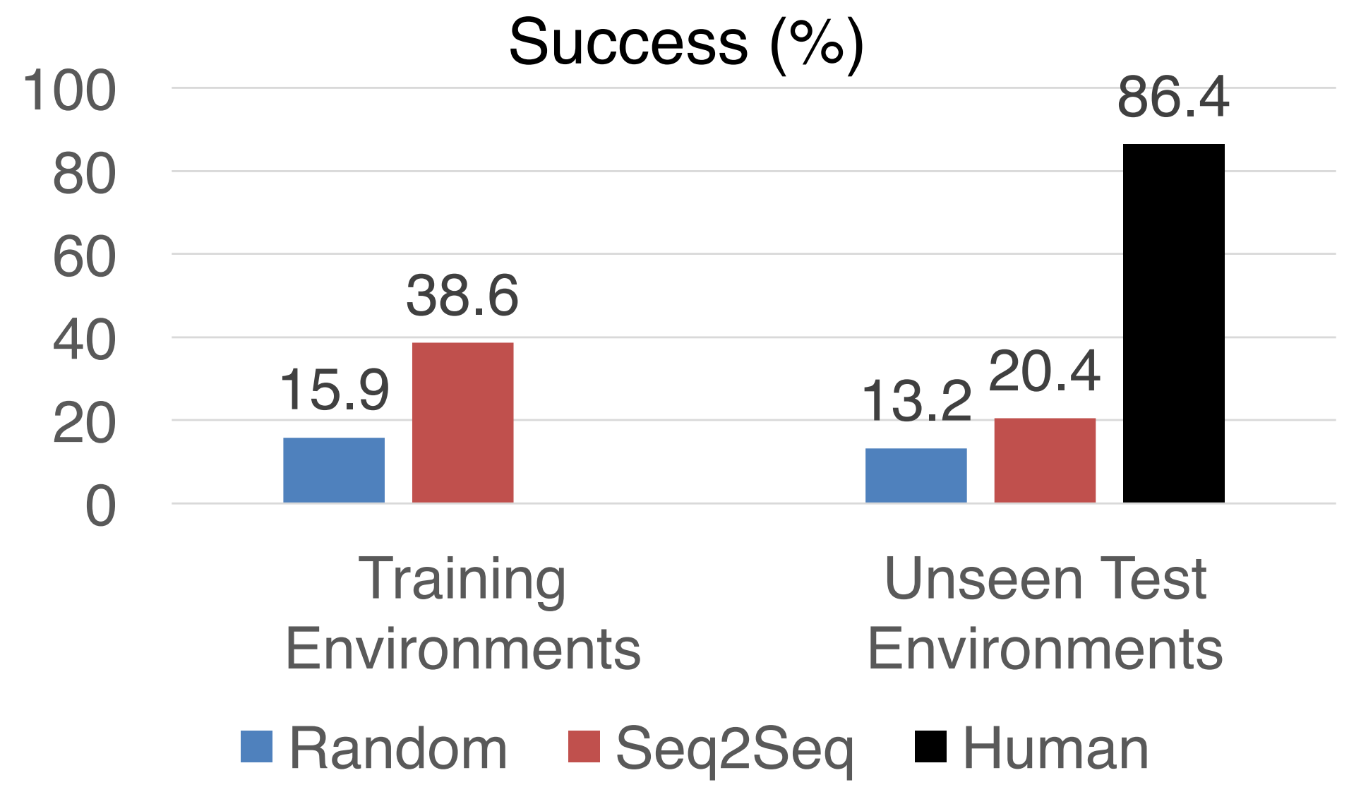
- LSTM-based Seq2Seq baseline model outputting a distribution over 6 actions: `left`, `right`, `up`, `down`, `forward` & `stop`.
- Image features from ResNet-152.
- Training with 'student-forcing' (sampling the next action) outperforms 'teacher-forcing' (selecting the ground-truth action).



6. Evaluation

Clear Evaluation Protocol:

- Report navigation error (distance from goal) for each instruction in the unseen test environments.
- 'Success' when navigation error < 3m.
- Agent must choose to stop (also report success rate with oracle stopping).



Test (unseen) performance:

	Trajectory Length (m)	Navigation Error (m)	Success (%)	Oracle Success (%)
Random	9.93	9.77	13.2	18.3
Seq2Seq	8.13	7.85	20.4	26.6
Human	11.90	1.61	86.4	90.2
Shortest Path	9.93	0.0	100	100

- Unseen environments prove very challenging for Seq2Seq.
- Test server available
- More data coming soon

Simulator, dataset, models & test server available via: https://bringmeaspoon.org