From VQA to VLN: Recent Advances in Vision-and-language Research

VLN Part

Qi Wu, Xin Wang, Yoav Artzi and Peter Anderson

VLN Tutorial

- See, communicate and act
- Embodied AI
- Vision-and-Language Navigation (VLN)
  - Proposed in 2018
  - Attracted a lot of attention
    - Methods
    - Datasets
    - Challenges
  - An emerging research area

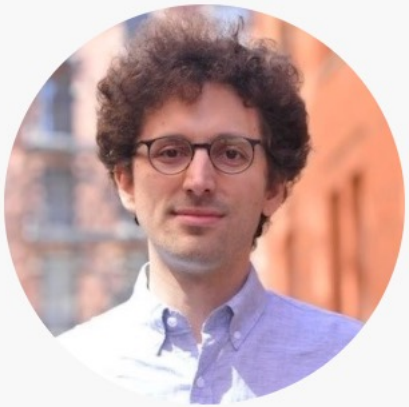
Tutorial Speakers

Qi Wu

University of Adelaide

Xin (Eric) Wang

UC Santa Cruz



Yoav Artzi

Cornell University

Peter Anderson

Google Research

Tutorial Schedule

1. VLN Overview
    1. VLN Tasks and Datasets (Qi Wu & Peter Anderson)
    2. Evaluation Metrics (Qi Wu)
2. Tackling Data Scarcity in VLN
    1. Data Augmentation (Xin Wang)
    2. Evaluation of Generated Navigation Instructions (Peter Anderson)
    3. Multitask Learning (Xin Wang)
3. Forward to Realistic VLN
    1. Real-life observations: Touchdown (Yoav Artzi)
    2. Real-life control: mapping instruction to continues control with a quadcopter drone. (Yoav Artzi)
    3. Sim-to-real transfer (Peter Anderson)

Part 1
Tasks, Datasets and Evaluations

Vision-and-Language

CV	+	NLP	=	Vision-to-Language (V2L)
Image Understanding	+	Language Generation	=	Image Captioning
Image Classification Object Detection Segmentation Object Counting Colour Analysis ….	+	Question Answering	=	Visual Question Answering
Image Understanding	+	Dialog	=	Visual Dialog

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

What color is the tablecloth?

Ours: white

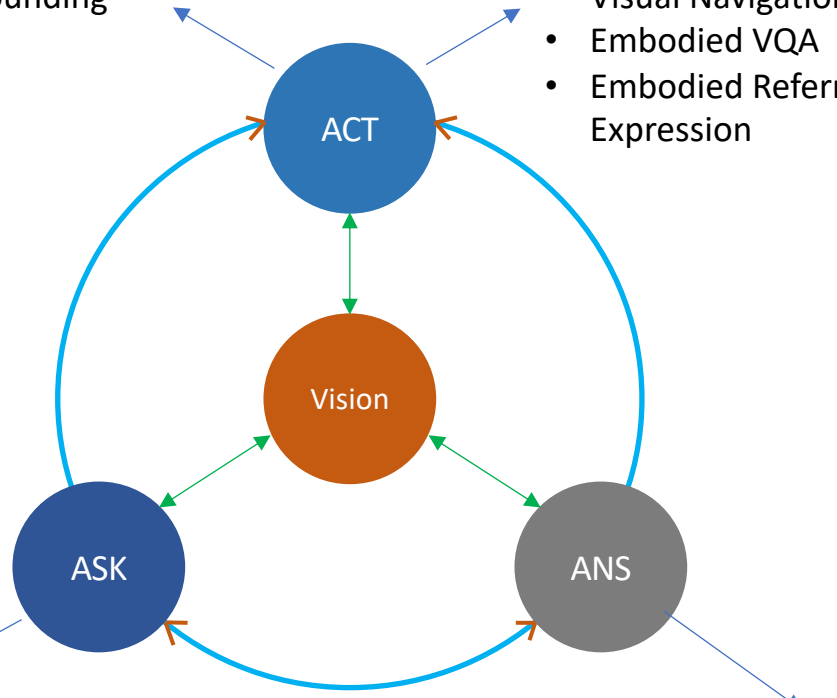
How many people in the photo?

Ours: 2

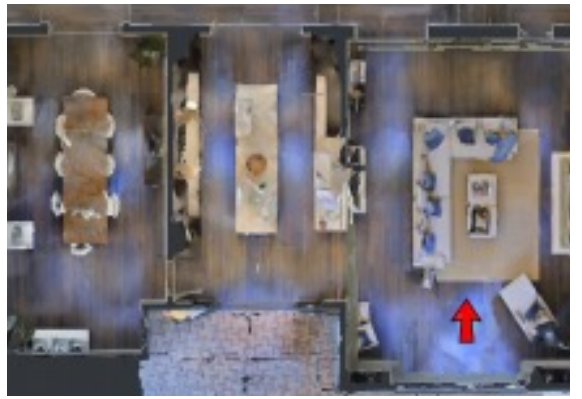
Connecting Vision and Language to Actions

- Referring Expression
- Visual Grounding

- Visual Question Generation (VQG)
- Question2Querry
- Image Captioning



- Language-guided Visual Navigation
- Embodied VQA
- Embodied Referring Expression



- VQA
- VisDial

Embodied AI

👁️ **See**: perceive their environment through vision or other senses.

🎤 **Talk**: hold a natural language dialog grounded in their environment.

👂 **Listen**: understand and react to audio input anywhere in a scene.

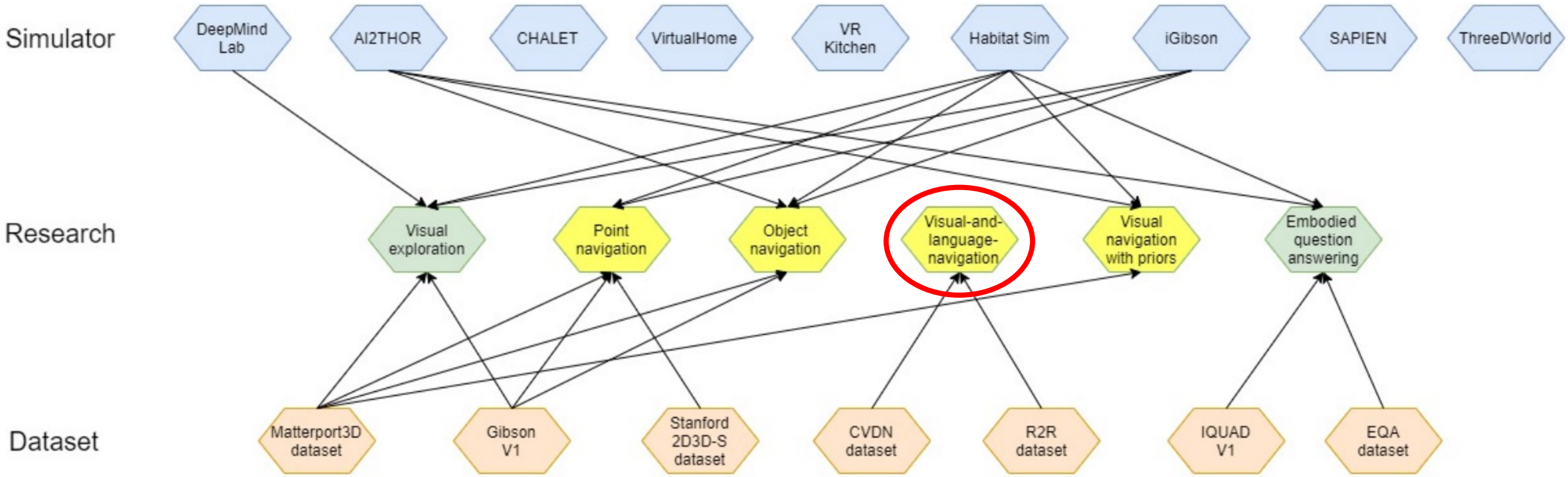
🕹️ **Act**: navigate and interact with their environment to accomplish goals.

🤔 **Reason**: consider and plan for the long-term consequences of their actions.

Embodied AI is the field for solving AI problems for virtual robots that can move, see, speak, and interact in the virtual world and with other virtual robots — these simulated robot solutions are then transferred to real world robots.

--- Luis Bermudez, Overview of Embodied Artificial Intelligence

VLN in Embodied AI



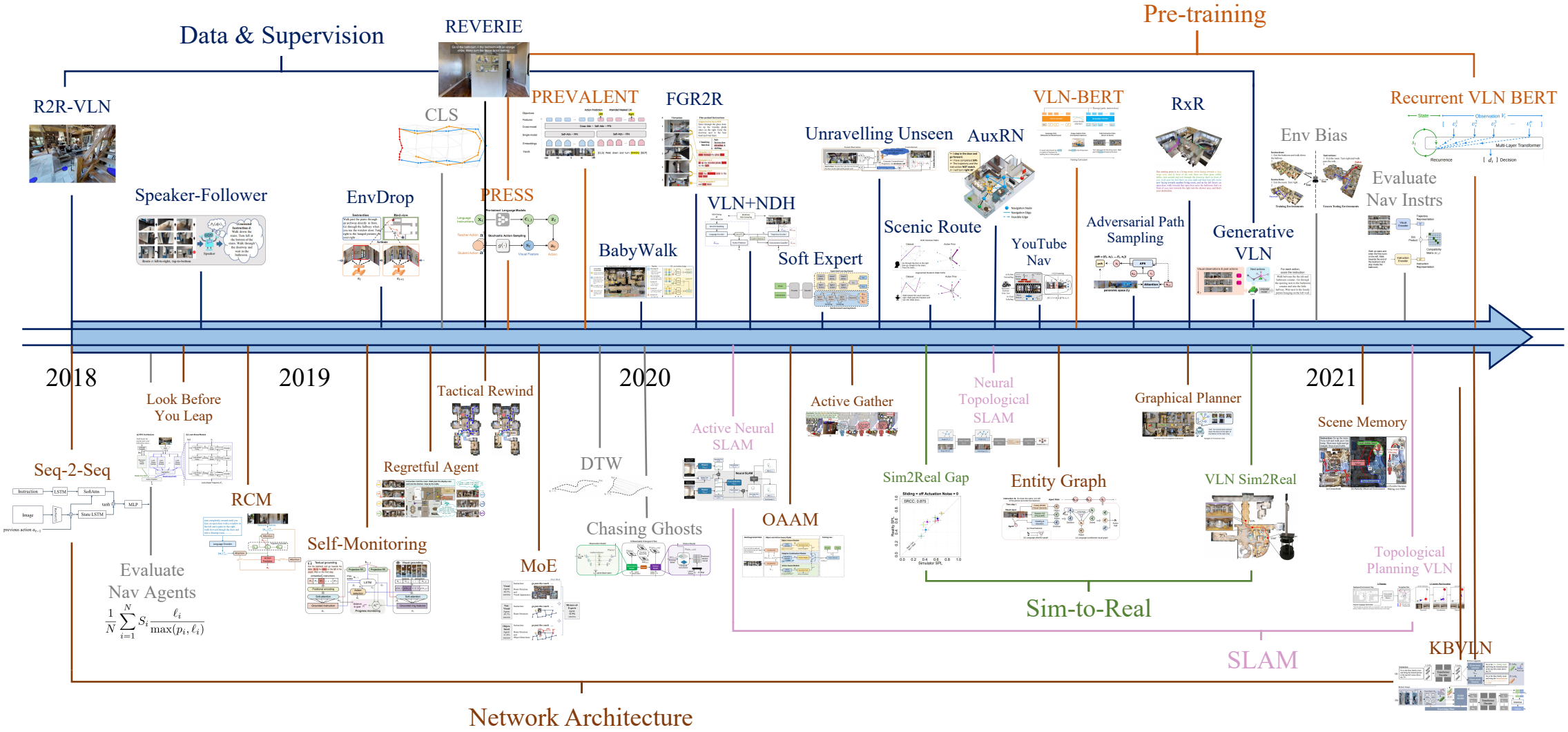
[Duan et al., A survey of Embodied AI: From simulators to Research Tasks, 2021]

Vision-and-Language Navigation (VLN)

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Vision-and-Language Navigation (VLN): an embodied agent is placed at a spot in a photo-realistic environment and the agent is called to navigate to a specific spot based on given natural language instructions.

VLN Timeline



- Evaluation methods, Analysis, and others

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

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, Anton van den Hengel

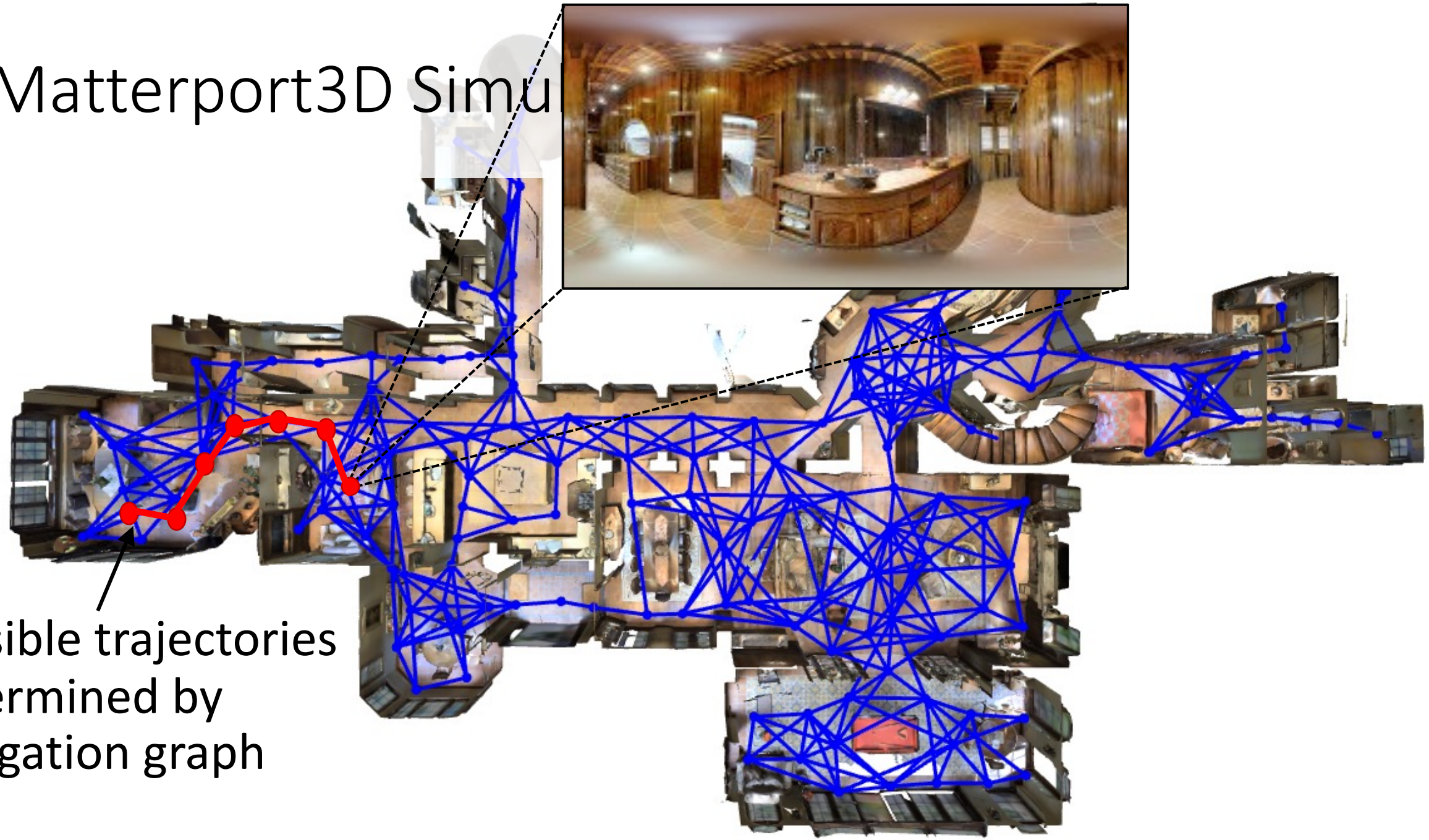
Project url: https://bringmeaspoon.org/

CVPR 2018
SALT LAKE CITY • JUNE 18-22

Matterport3D Simulator

- Simulator for **embodied visual agents**, based on Matterport3D dataset (Chang et. al. 2017)
- Contains 10,800 panoramic images / 90 buildings
- High visual diversity

Matterport3D Simu



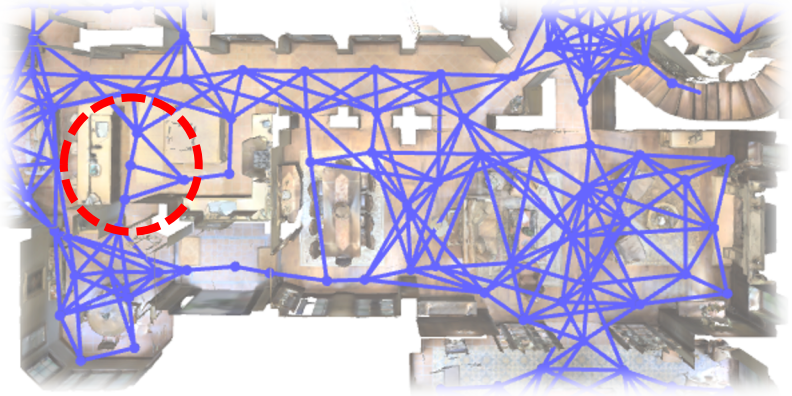
Feasible trajectories determined by navigation graph

Room-to-Room (R2R) Navigation dataset

Task: Given some natural language instructions, navigate through the environment to find the goal location.

Clear Evaluation Protocol:

- Single test run
- Agent must stop
- Success if <3m from goal



Metrics:

- Success / Oracle Success Rate (%)
- Navigation Error (m)
- SPL [Anderson et al. arxiv 1807.06757, 2018]
- CLS [Jain et al. arxiv 1905.12255, 2019]
- nDTW [Magalhaes et al. arxiv 1907.05446, 2019]

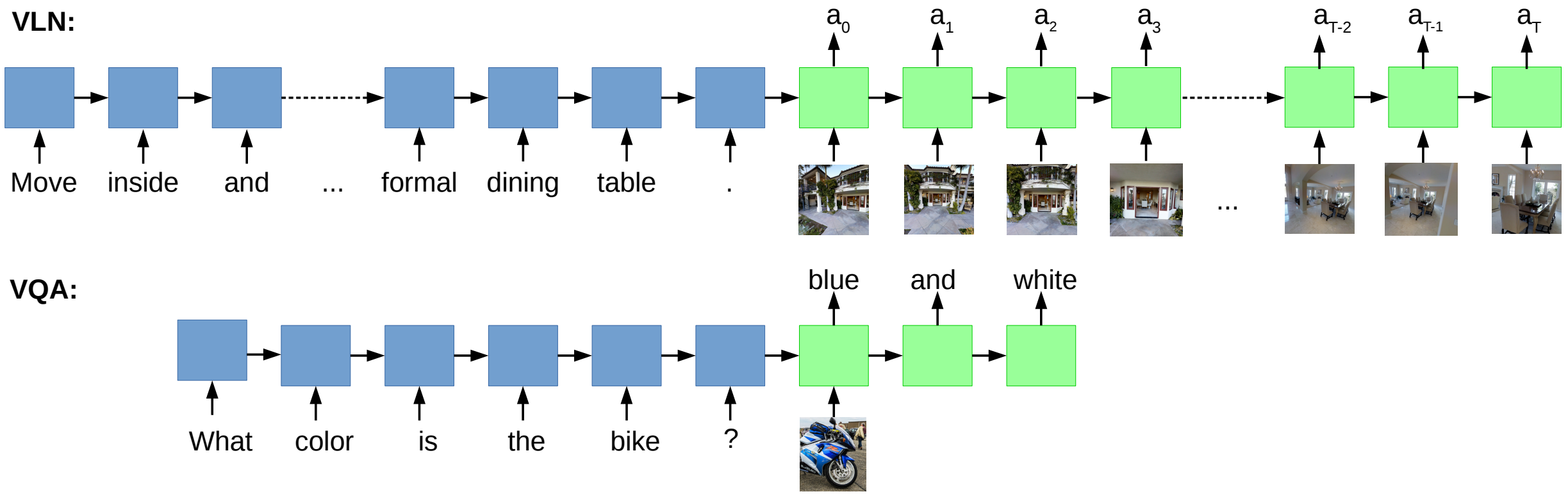
Room-to-Room (R2R) Dataset

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

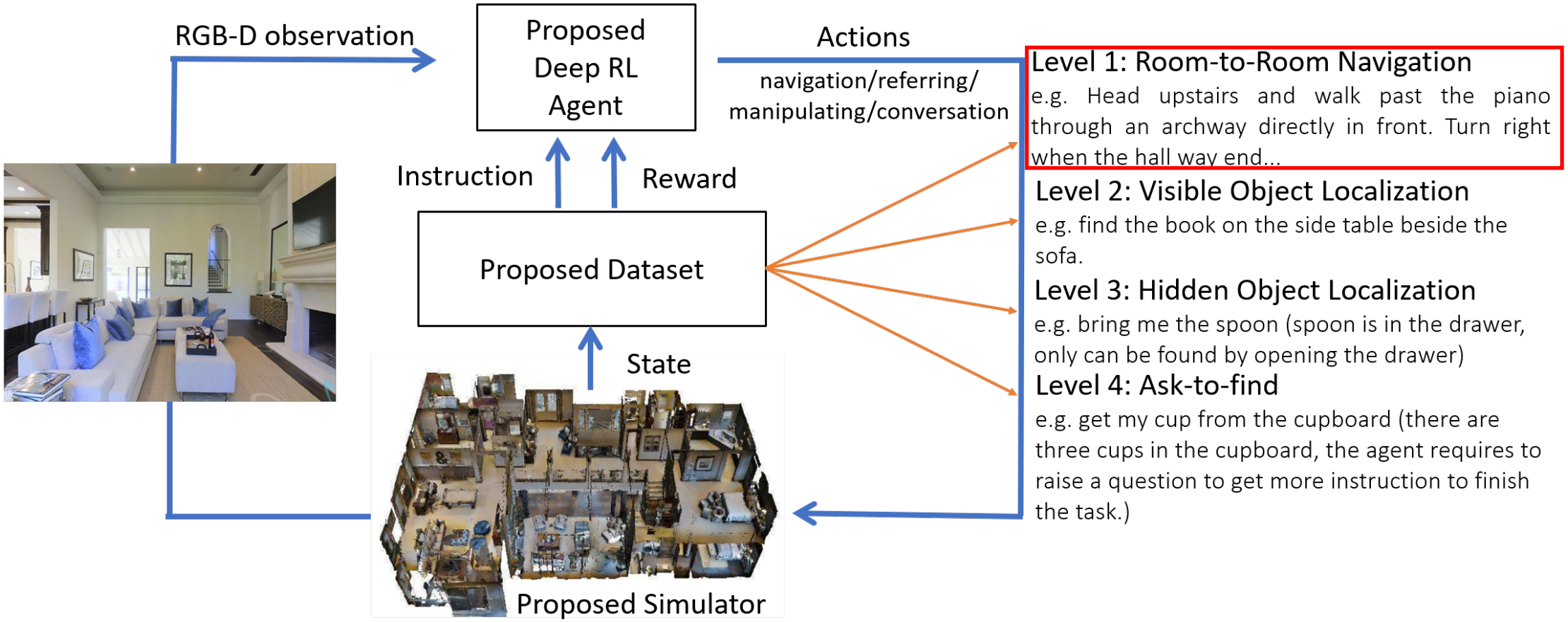
- Sampled 7,189 shortest paths between locations (mostly) in different rooms
- Collected 21,567 instructions using AMT and a WebGL simulator interface
- 400 people spent 1,600 hours annotating

Comparison of VLN to VQA

- VLN adds active vision, longer sequences, domain gap between train and test



Indoor VLN in 4 levels



Indoor VLN in 4 levels

RGB-D observation

Proposed Deep RL Agent

Actions

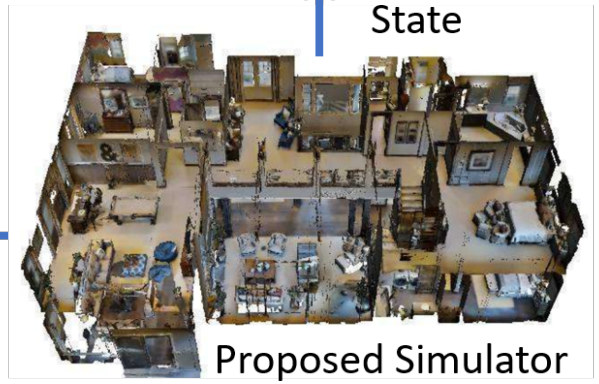
navigation/referring/ manipulating/conversation

Instruction

Reward

Proposed Dataset

State



Proposed Simulator

Level 1: Room-to-Room Navigation

e.g. Head upstairs and walk past the piano through an archway directly in front. Turn right when the hall way end...

Level 2: Visible Object Localization

e.g. find the book on the side table beside the sofa.

Level 3: Hidden Object Localization

e.g. bring me the spoon (spoon is in the drawer, only can be found by opening the drawer)

Level 4: Ask-to-find

e.g. get my cup from the cupboard (there are three cups in the cupboard, the agent requires to raise a question to get more instruction to finish the task.)

Example:

"Bring me the blue cushion from the living room"

REVERIE: Remote Embodied Visual Referring Expressions in Real Indoor Environments

Yuankai Qi[1], **Qi Wu**[1], Peter Anderson[2], Xin Wang[3], William Yang Wang[3], Chunhua Shen[1], Anton van den Hengel[1]

[1]*Australian Centre for Robotic Vision, The University of Adelaide, Australia*

[2]*Georgia Institute of Technology, USA* [3]*University of California, Santa Barbara, USA*

CVPR 2020

The REVERIE Task

R2R vs. REVERIE

Two key difference:

- Fine-grained instructions vs. High-level instruction

  R2R: ‘Go to the top of the stairs then turn left and walk along the hallway and stop at the first bedroom on your right’

  REVERIE: ‘the cold tap in the first bedroom on level two’

- Point navigation vs. Remote object grounding

Challenges

(1/3) Significant Appearance Variation

Challenges

(2/3) Rich Linguistic Phenomena

Dangling modifiers (e.g. 1), spatial relations (e.g. 3), imperatives (e.g. 4), co-references (e.g. 5)

1. Fold the towel in the bathroom with the fishing theme
2. Push in the bar chair, in the kitchen, by the oven.
3. Go to the blue family room and bring the framed picture of a person on a horse at the top left corner above the TV.
4. Could you please dust the light above the toilet in the bathroom that is near the entry way?
5. There is a bottle in the office alcove next to the piano. It is on the shelf above the sink on the extreme right. Please bring it here.

Challenges

(3/3) Less Words, More Contents

- Instruction length: 18 vs 29 words (Room-to-Room dataset)
- 56% instructions mention 3 or more objects, 28% mention 2 objects
- Involve 4,140 objects, falling into 489 categories vs 80 categories in ReferCOO

Indoor VLN in 4 levels

RGB-D observation

Proposed Deep RL Agent

Actions

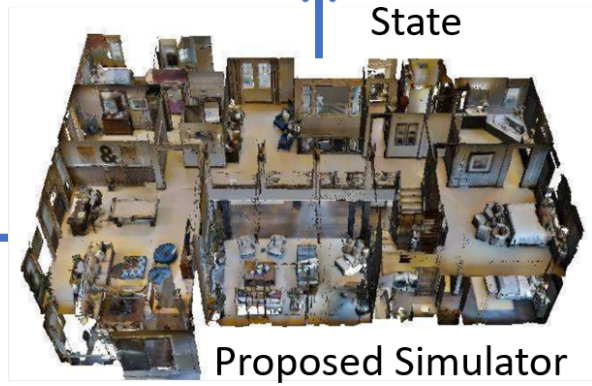
navigation/referring/ manipulating/conversation

Instruction

Reward

Proposed Dataset

State



Proposed Simulator

Level 1: Room-to-Room Navigation

e.g. Head upstairs and walk past the piano through an archway directly in front. Turn right when the hall way end...

Level 2: Visible Object Localization

e.g. find the book on the side table beside the sofa.

Level 3: Hidden Object Localization

e.g. bring me the spoon (spoon is in the drawer, only can be found by opening the drawer)

Level 4: Ask-to-find

e.g. get my cup from the cupboard (there are three cups in the cupboard, the agent requires to raise a question to get more instruction to finish the task.)

ALFRED

- A Benchmark for Interpreting Grounded Instructions for Everyday Tasks

chill the lettuce and then put the lettuce on the table

[Shridhar et al., ALFRED A Benchmark for Interpreting Grounded Instructions for Everyday Tasks, 2020]

Indoor VLN in 4 levels

RGB-D observation

Proposed Deep RL Agent

Actions

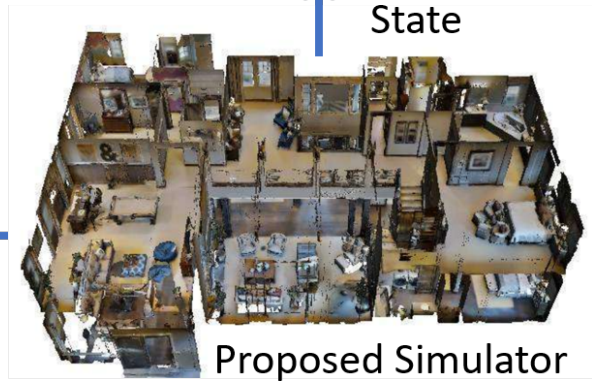
navigation/referring/ manipulating/conversation

Instruction

Reward

Proposed Dataset

State



Proposed Simulator

Level 1: Room-to-Room Navigation

e.g. Head upstairs and walk past the piano through an archway directly in front. Turn right when the hall way end...

Level 2: Visible Object Localization

e.g. find the book on the side table beside the sofa.

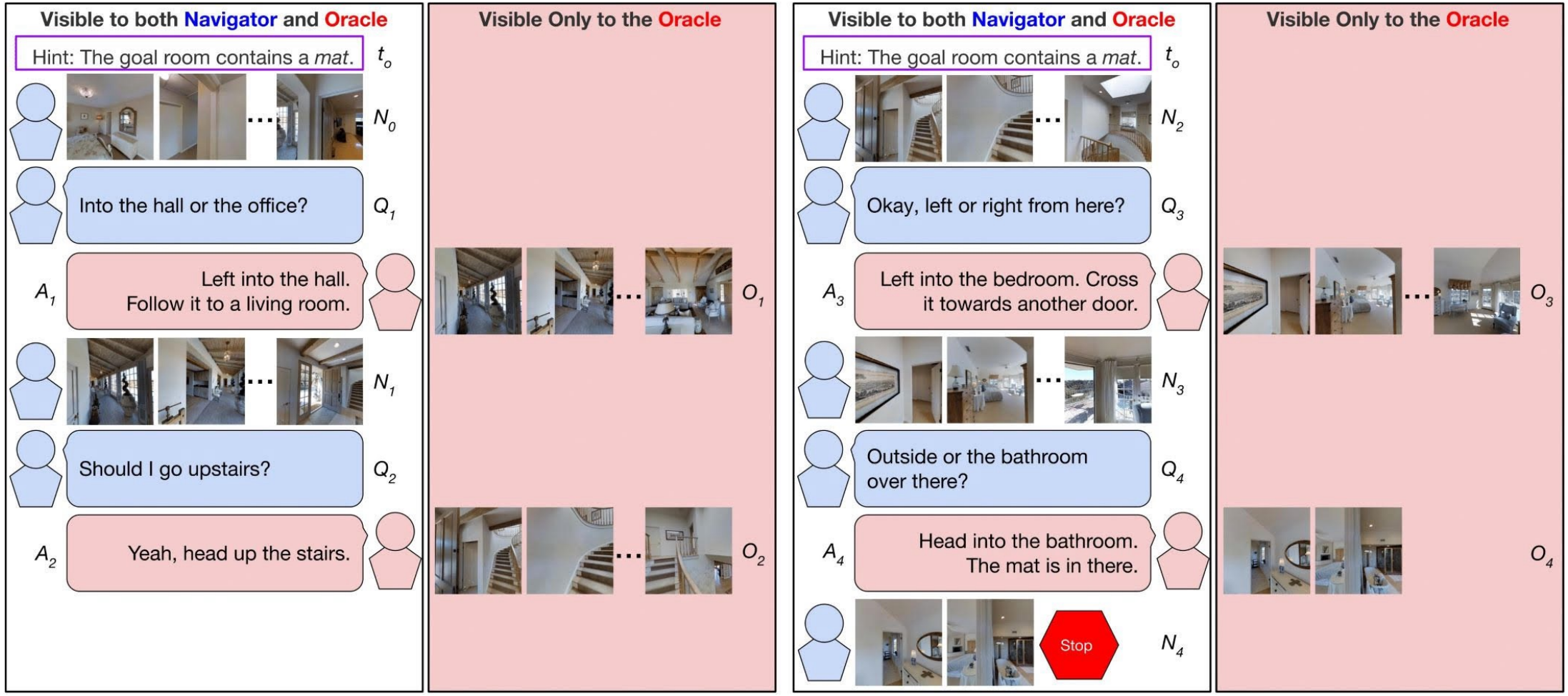
Level 3: Hidden Object Localization

e.g. bring me the spoon (spoon is in the drawer, only can be found by opening the drawer)

Level 4: Ask-to-find

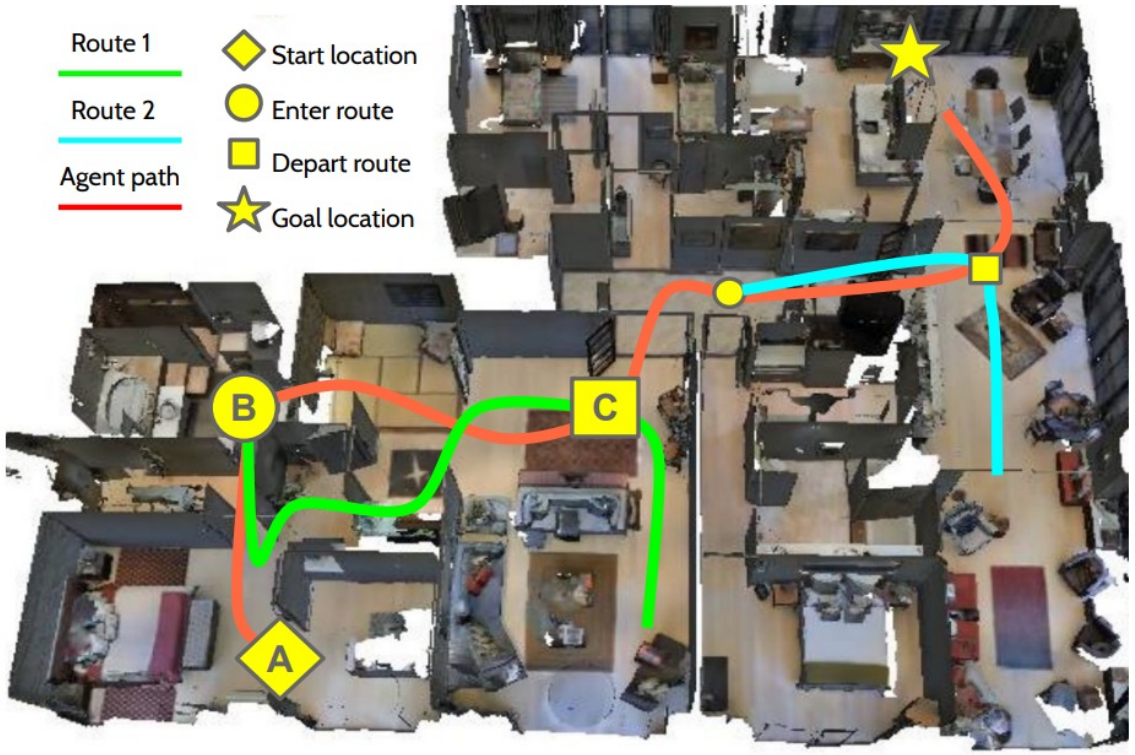
e.g. get my cup from the cupboard (there are three cups in the cupboard, the agent requires to raise a question to get more instruction to finish the task.)

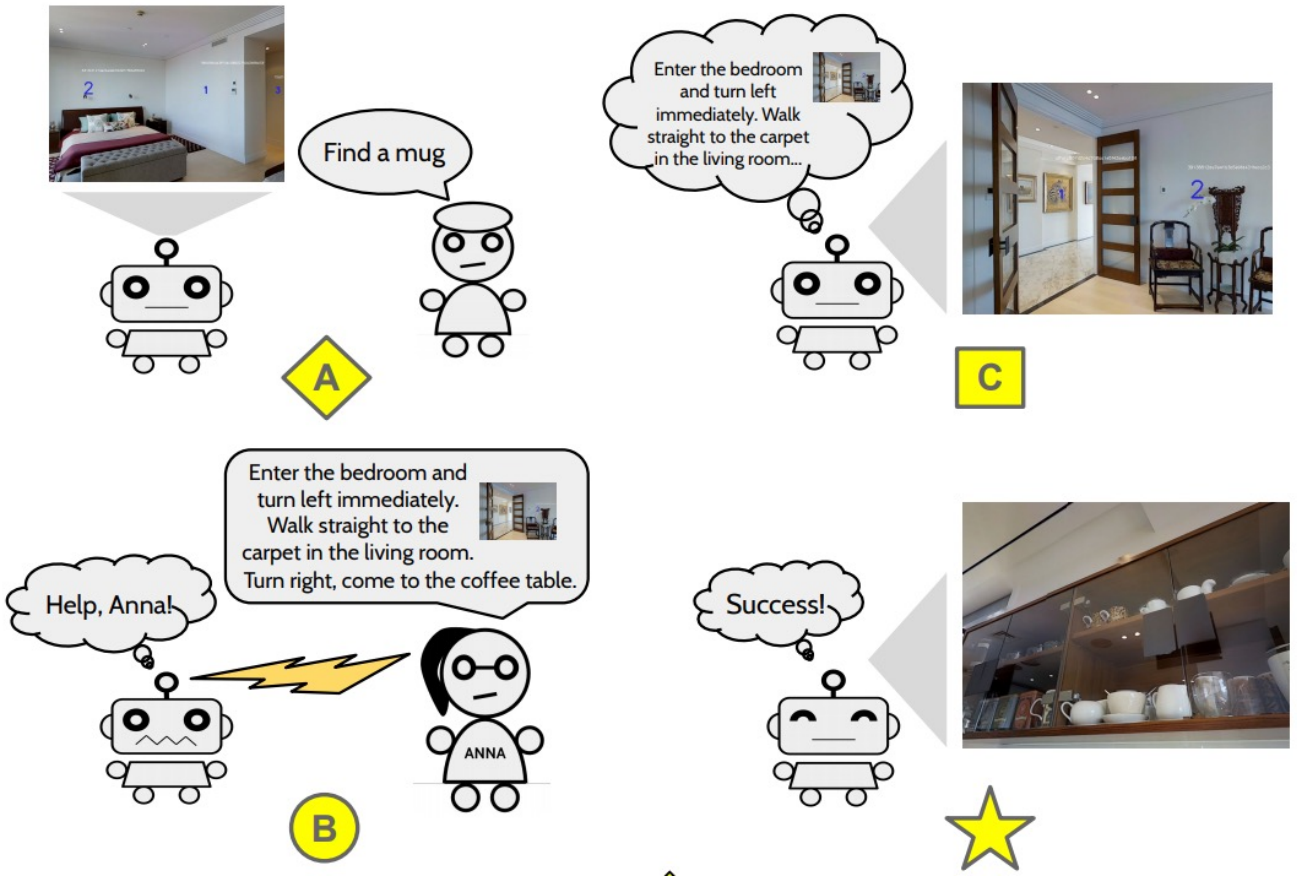
Cooperative Vision-and-Dialog Navigation



[Thomason et al., Vision-and-Dialog Navigation, 2019]

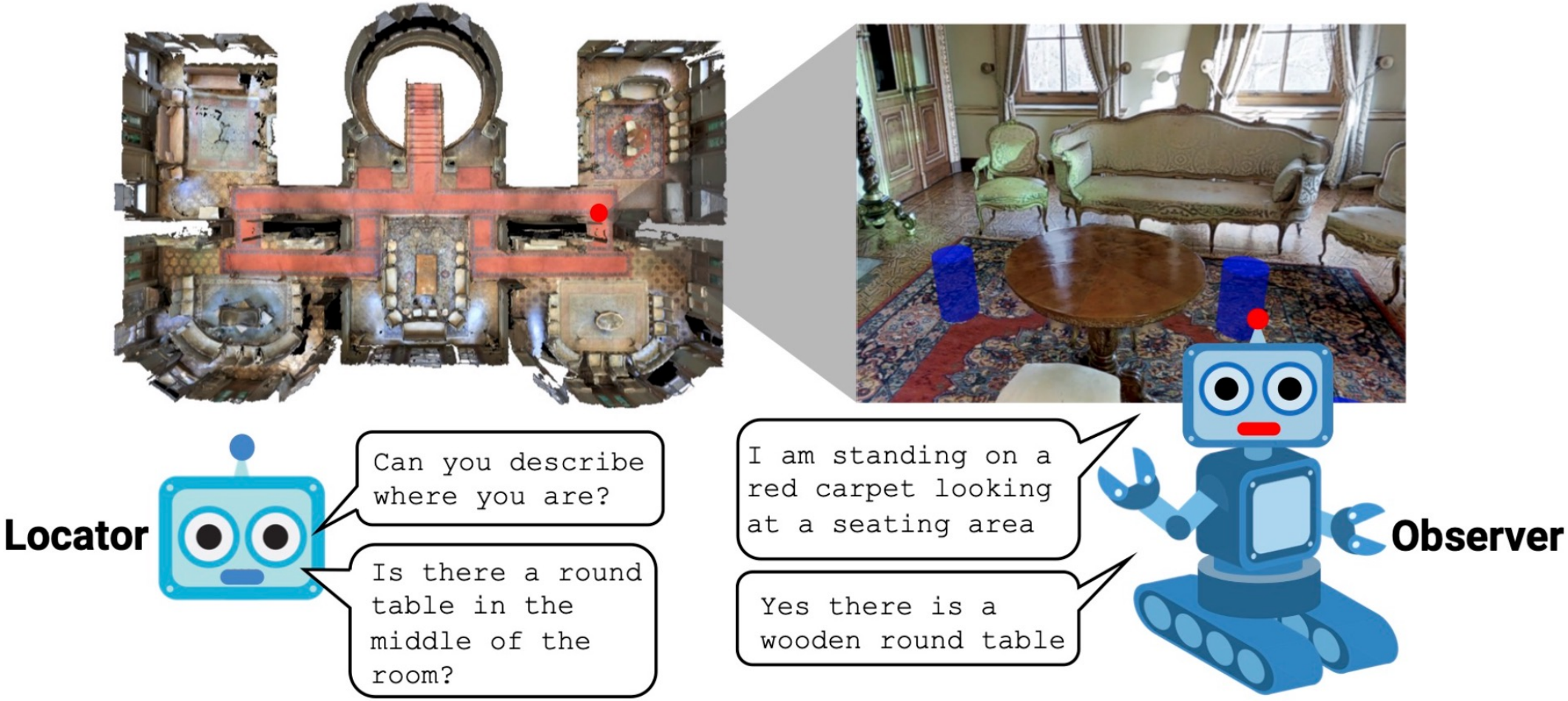
HANNA: Visual Navigation with Natural Multimodal Assistance





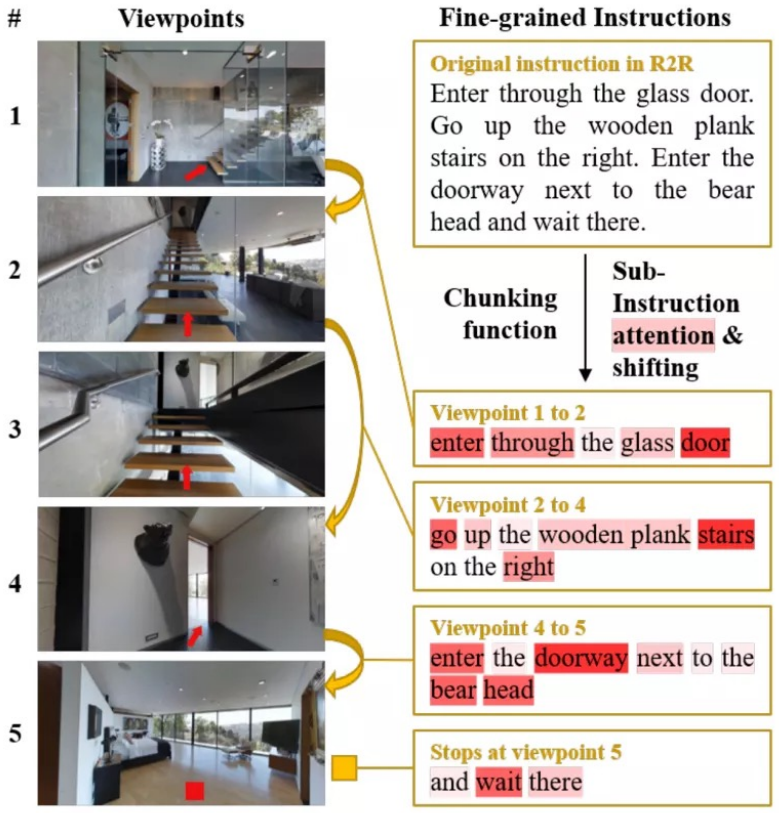
[Nguyen et al., Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning, 2019]

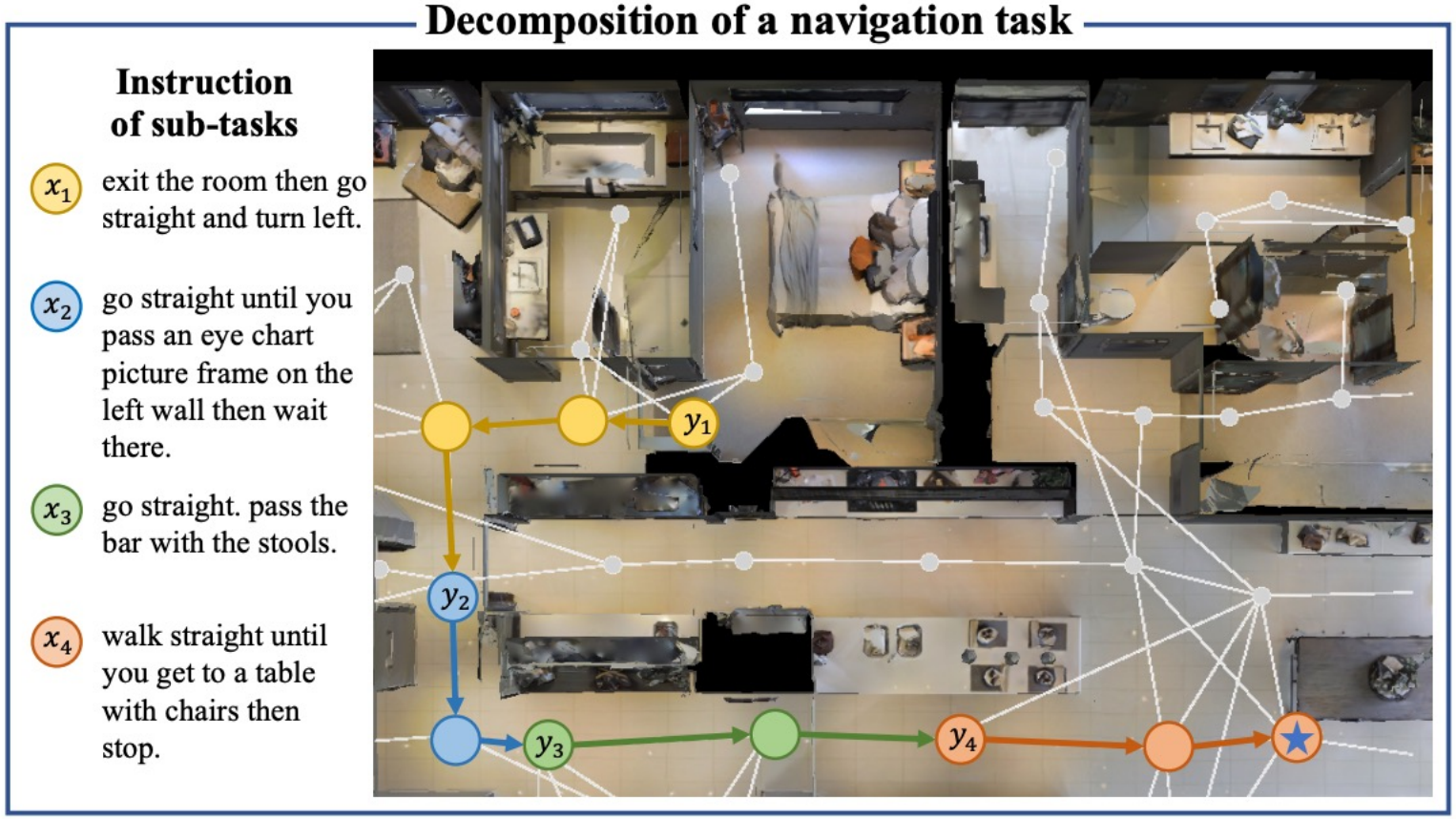
Where Are You?



[Hahn et al., Where Are You? Localizaiton from Embodied Dialog, 2020]

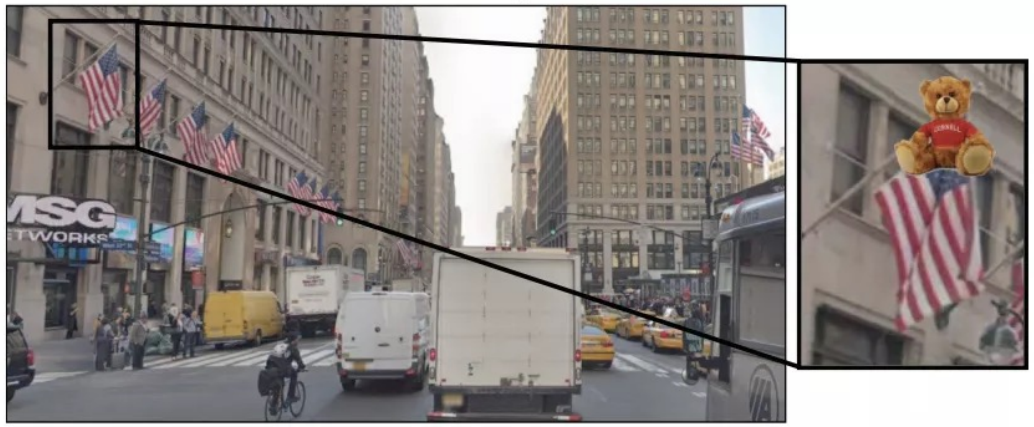
Fine-grained Instructions





[Hong et al., Sub-Instruction Aware Vision-and-Language Navigation, 2020]
[Zhu et al., BabyWalk: Going Farther in Vision-and-Language Navigation by Taking Baby Steps, 2020]

Outdoor VLN



Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

[Chen et al., Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments, 2019]

[Vries et al., Talk the Walk: Navigating New York City through Grounded Dialogue, 2019]

Tourist

Should I go to pub corner? Or one of the restaurant corners?

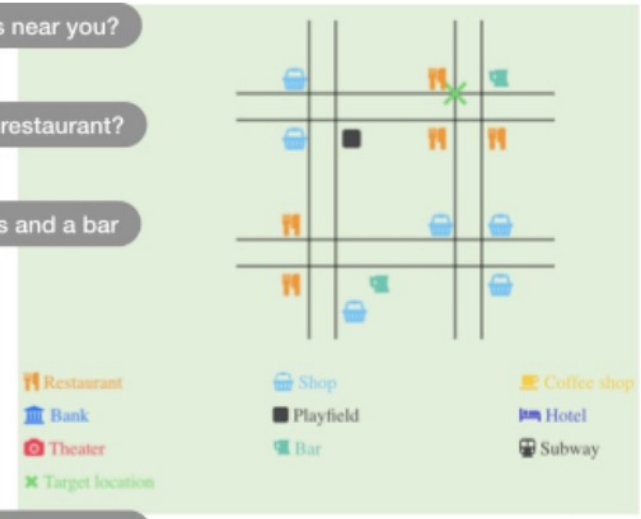
I'm there

Hey, what is near you?

Brook Brothers

Is that a shop or restaurant?

Your target intersection has three restaurants and a bar



Go to a restaurant corner facing the pub.

Evaluate

Guide

Others (Peter Anderson)

- RxR/ PanGEA
- RxR-Habitat

Evaluation Metrics

- Success / Oracle Success Rate (%)
- Navigation Error (m)
- SPL (Success weighted by Path Length)
- CLS (Coverage weighted by Length Score)
  - Measuring fidelity to the reference path
- nDTW (normalized Dynamic Time Warping)
- SDTW (Success weighted by normalized Dynamic Time Warping)

VLN Models

- Seq2seq (a golden baseline)
  - Speaker-follower
- Attention Mechanism (something must try)
  - EnvDrop, Self-monitoring, OAAM
- Transformer (this is all you need)
  - PREVALENT, Recurrent-Bert
- Reinforcement Learning (Add-on)
  - RCM, Soft Expert