Grounding Natural Language for Building Embodied Agents

Asli Celikyilmaz

Microsoft Research

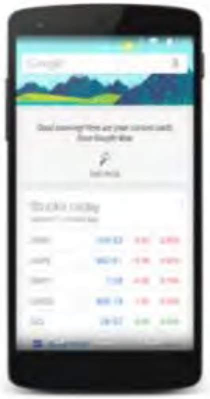
Acknowledgements:

Xin Wang (UCSB), Qiuyuan Huang (MSR), Dinghan Shen (Duke), Xiujun Li (MSR)
Jianfeng Gao (MSR), Lei Zhang (MSR), Yonatan Bisk (MSR), William Wang (UCSB)

Language Empowering Intelligent Agents

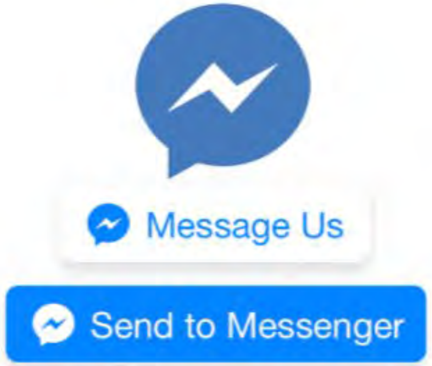
Microsoft Cortana

Apple Siri



Google Now
Google Assistant

Amazon Alexa/Echo



Facebook M & Bot

Google Home

Apple HomePod

Adapting Agents to Physical Environments

Image Source : Henderson Biomedical

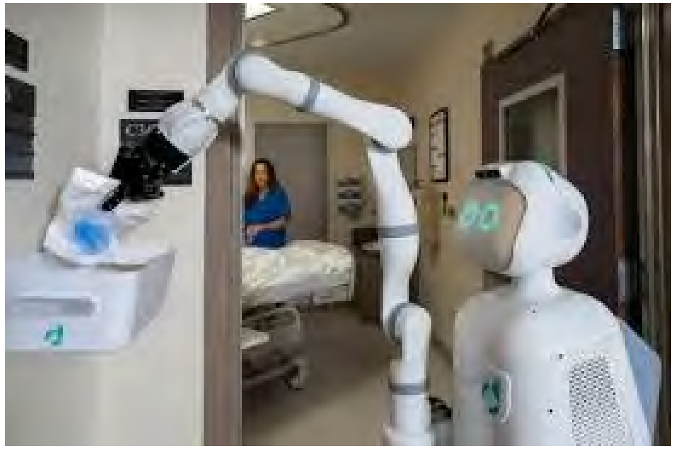


Image Source : boingboing.net

Image Source : Boston Dynamics

Outline

- Language grounding in visual environments
  - Visual Language Navigation Task
  - Self-supervised imitation learning [CVPR 2019]
- Ongoing Work
  - Navigation and Dialog
  - situated and bi-directional

Intelligent Agents Navigating Physical Environments

Our Goal → Build intelligent agents

- Communicate with people
  - Follow natural language instructions
- Understand the dynamics of the perceptual environment
- Alignment between the two !

Language Grounding in Situated Environments

Linguistic symbols ⟷ Perceptual experiences and actions

(Noun Phrase)
dog reading newspaper

(Verb)
sleeping

(Verb Phrase)
climb on chair to reach switch

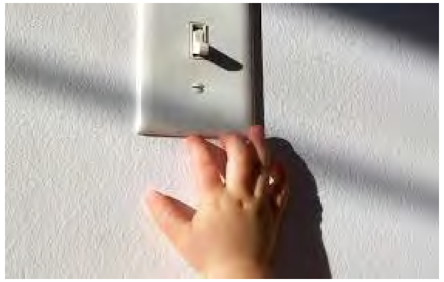


Image Source : bing.com

Understanding Visually Grounded Language

e : Microsoft/HoloLens web page

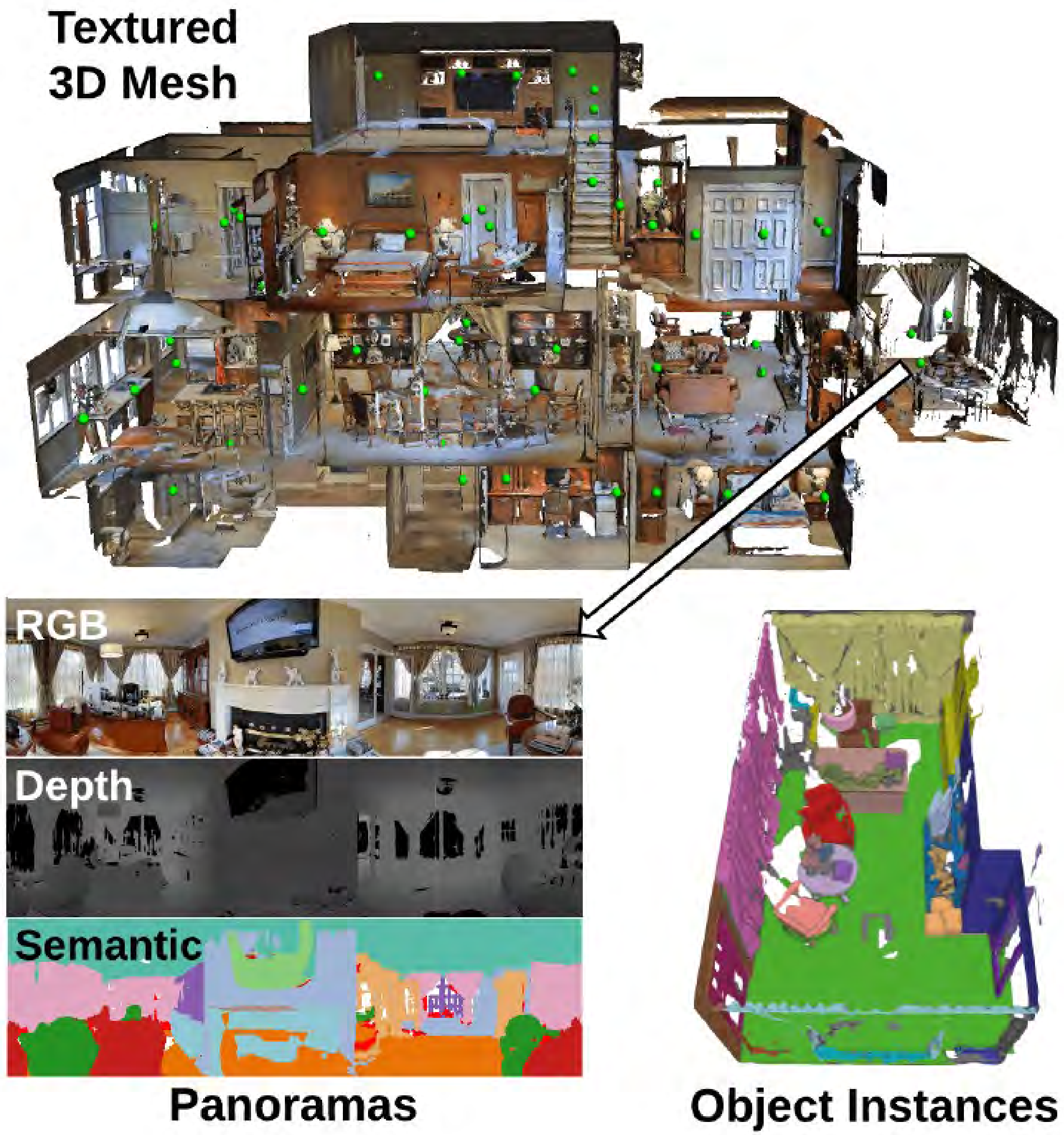
Understanding Visually Grounded Language

TASK: Vision & Language Navigation (VNL)

Navigating an agent inside real 3D environments by following natural language instructions.

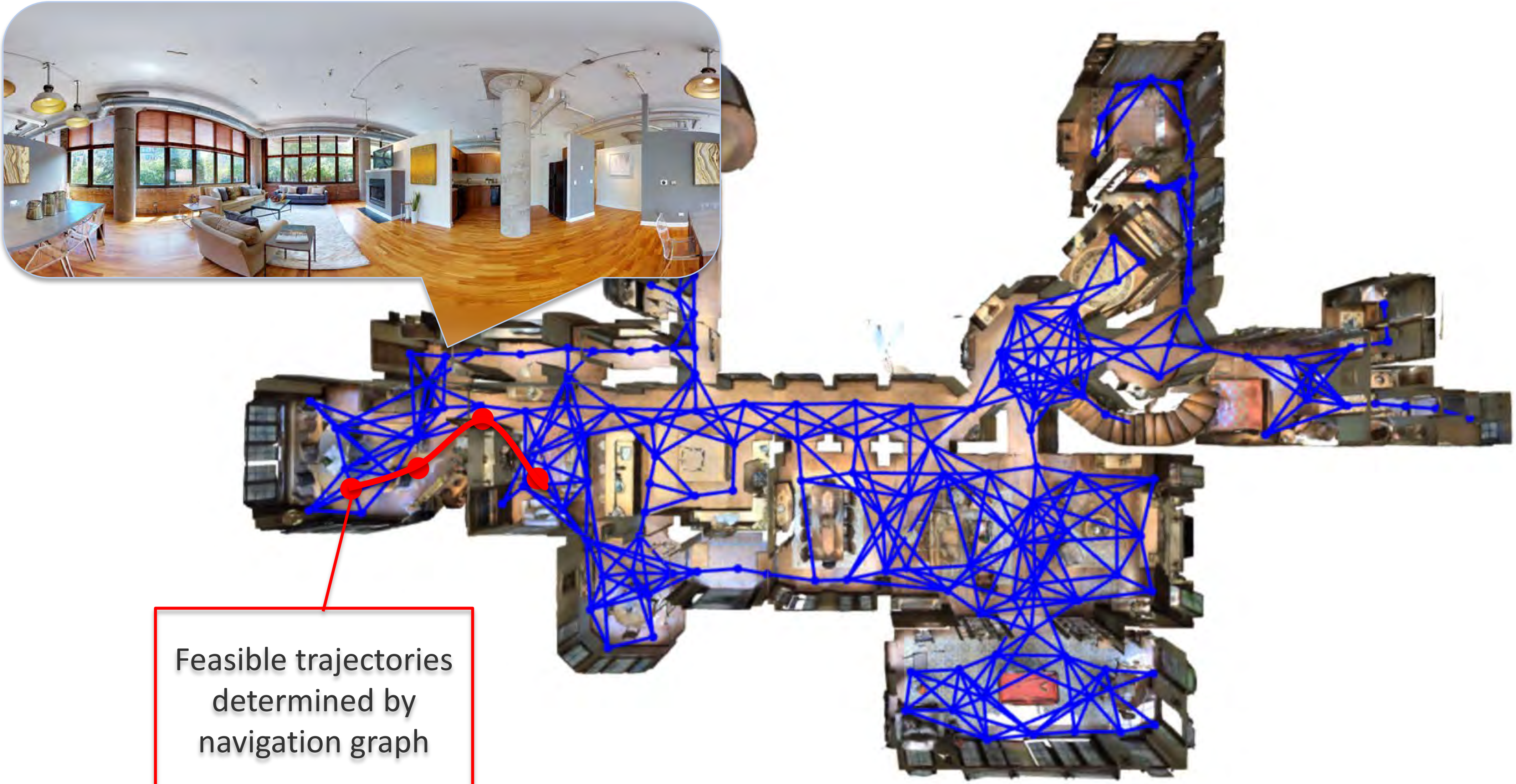
Matterport

Matterport 3D Dataset



- 10,800 panoramic views based on 194K RGB-D images
- 90 building-scale scenes (avg. 23 rooms each)
- Includes textured 3D meshes with object segmentations
- Largest RGB-D dataset

Matterport 3D Simulator for VLN Task



Feasible trajectories determined by navigation graph

Anderson, et.al., CVPR 2018

Matterport 3D Simulator for VLN Task

Room-to-Room (R2R) Dataset

- ~7K shortest paths
- 3 instructions for each path
  - Average instruction length 29 words
  - Average trajectory length is 10 meters
- **Task**: given natural language instructions, 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.

Anderson, et.al., CVPR 2018

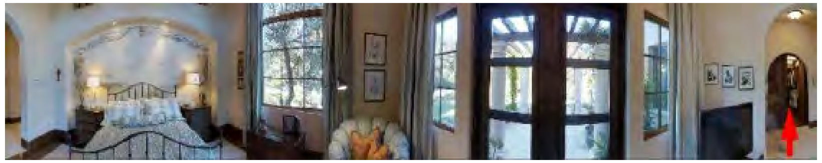
Room-to-Room Dataset Examples

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

Input: Instruction

turn completely around until you face an open door with a window to the left and a patio to the right, walk forward, … …

Input: Panoramic View



Output

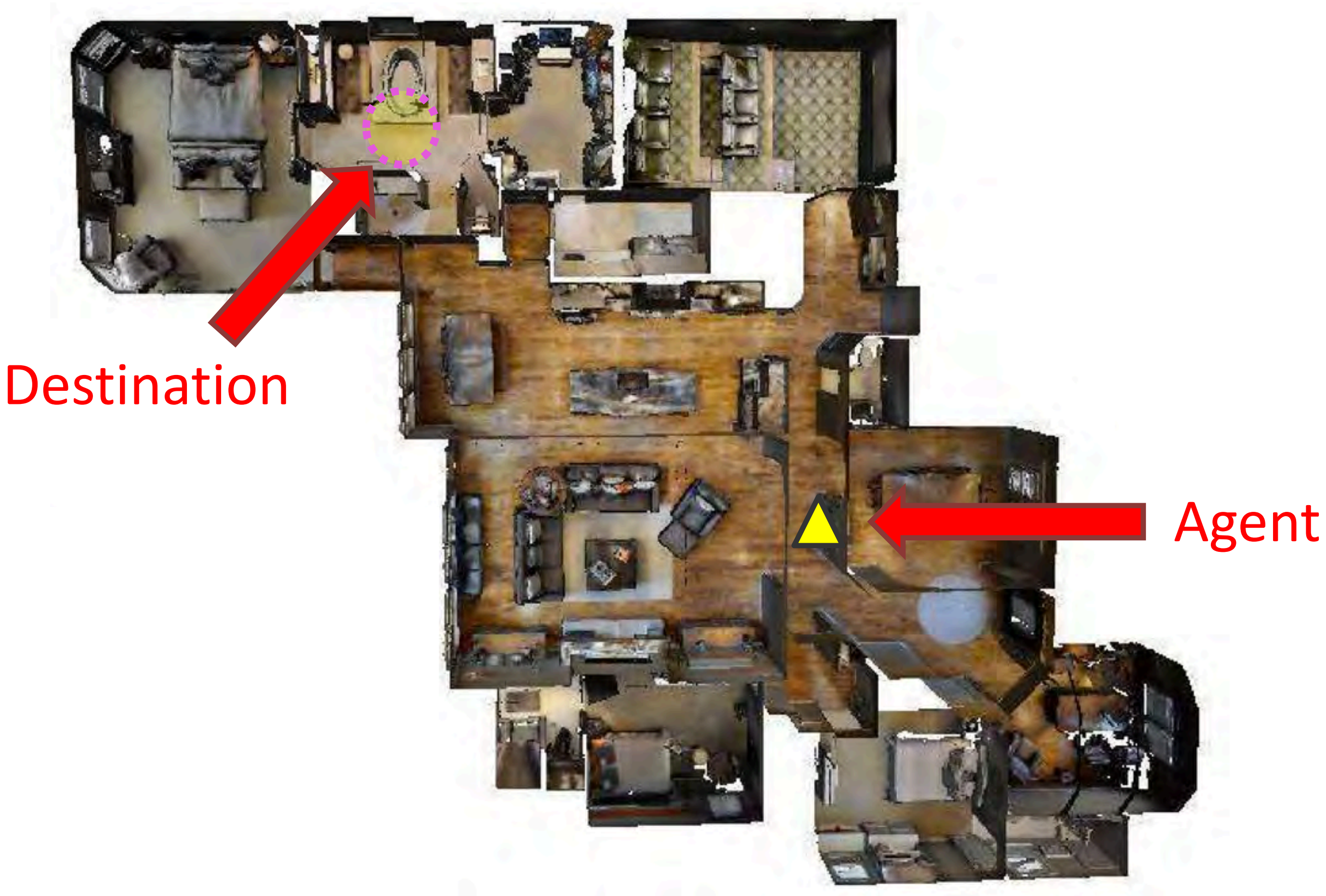
$$a_t \in A$$

Anderson, et.al., CVPR 2018

Visual-Language Navigation Task Challenges

(1) cross-modal grounding

Instruction: Go towards the *living room* and then turn right to the *kitchen*. Then turn left, pass a *table* and enter the *hallway*. Walk down the *hallway* and turn into the *entry way* to your right. Stop in front of the *toilet*.



Local visual scene

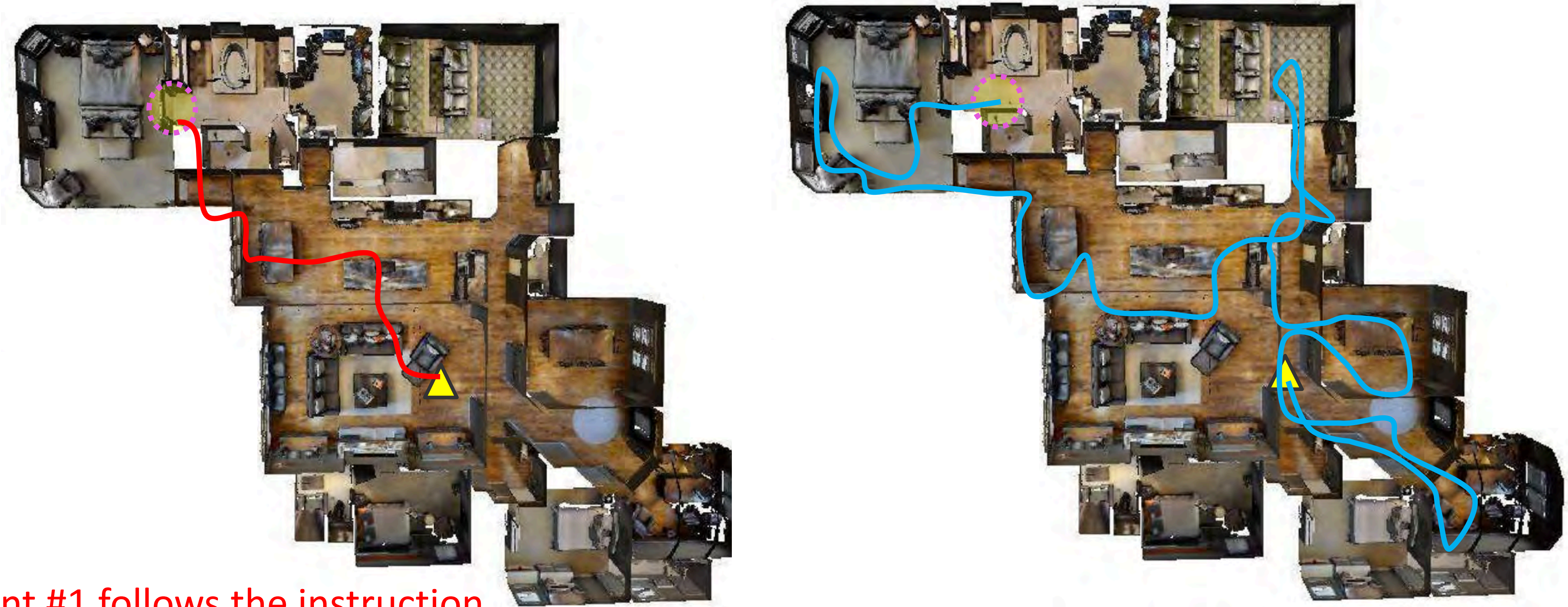
Global trajectory in top-down view (NOT visible to the agent)

Visual-Language Navigation Task Challenges

(1) cross-modal grounding

(2) ill-posed feedback

Instruction: Go towards the *living room* and then turn right to the *kitchen*. Then turn left, pass a *table* and enter the *hallway*. Walk down the *hallway* and turn into the *entry way* to your right. Stop in front of the *toilet*.



Agent #1 follows the instruction and reaches the destination.

Agent #2 randomly walks insides the house and reaches the destination.

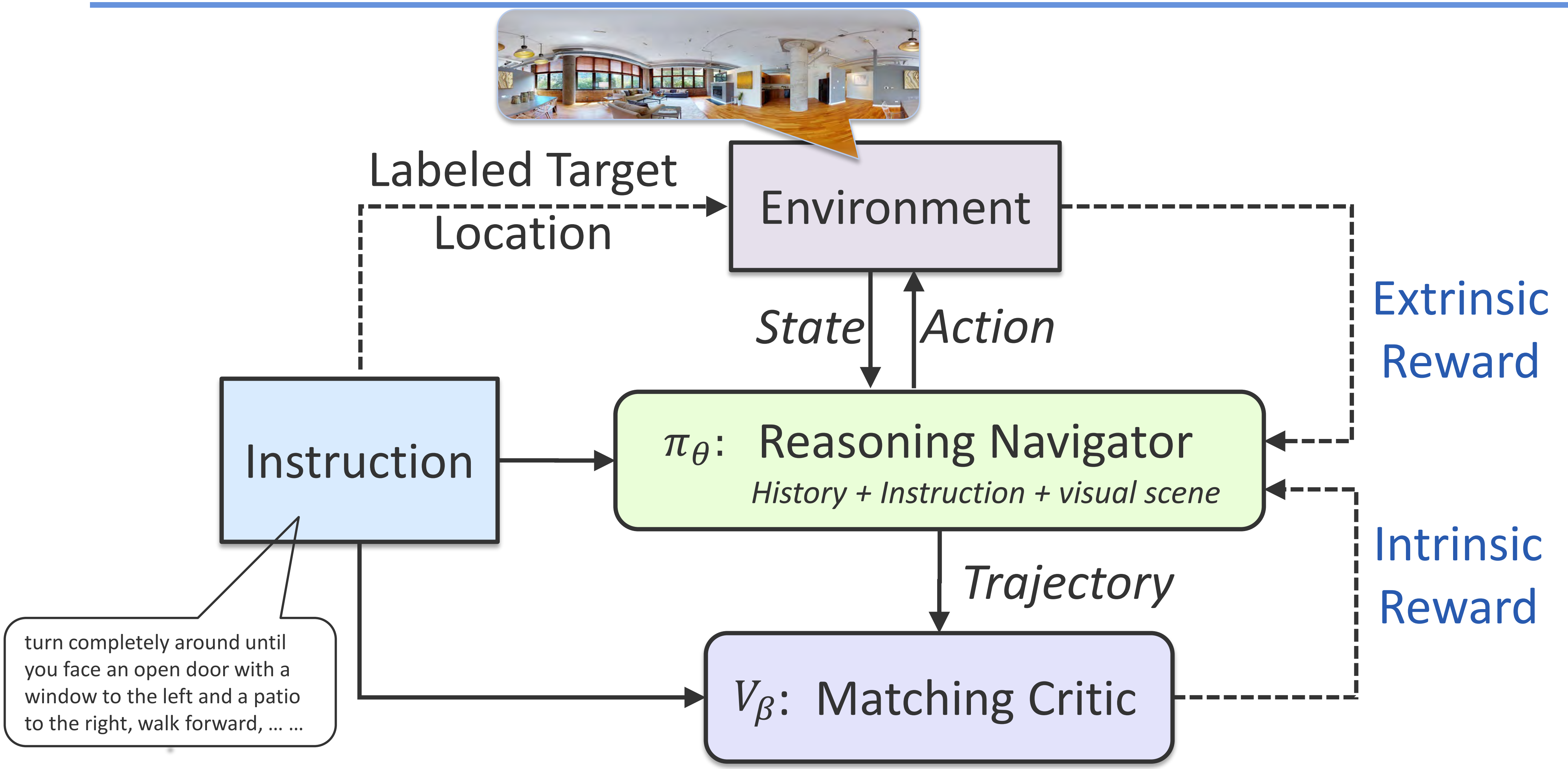
Both trajectories are considered same in terms of the success signal.

Our Recent “Explanatory” Work on VLN

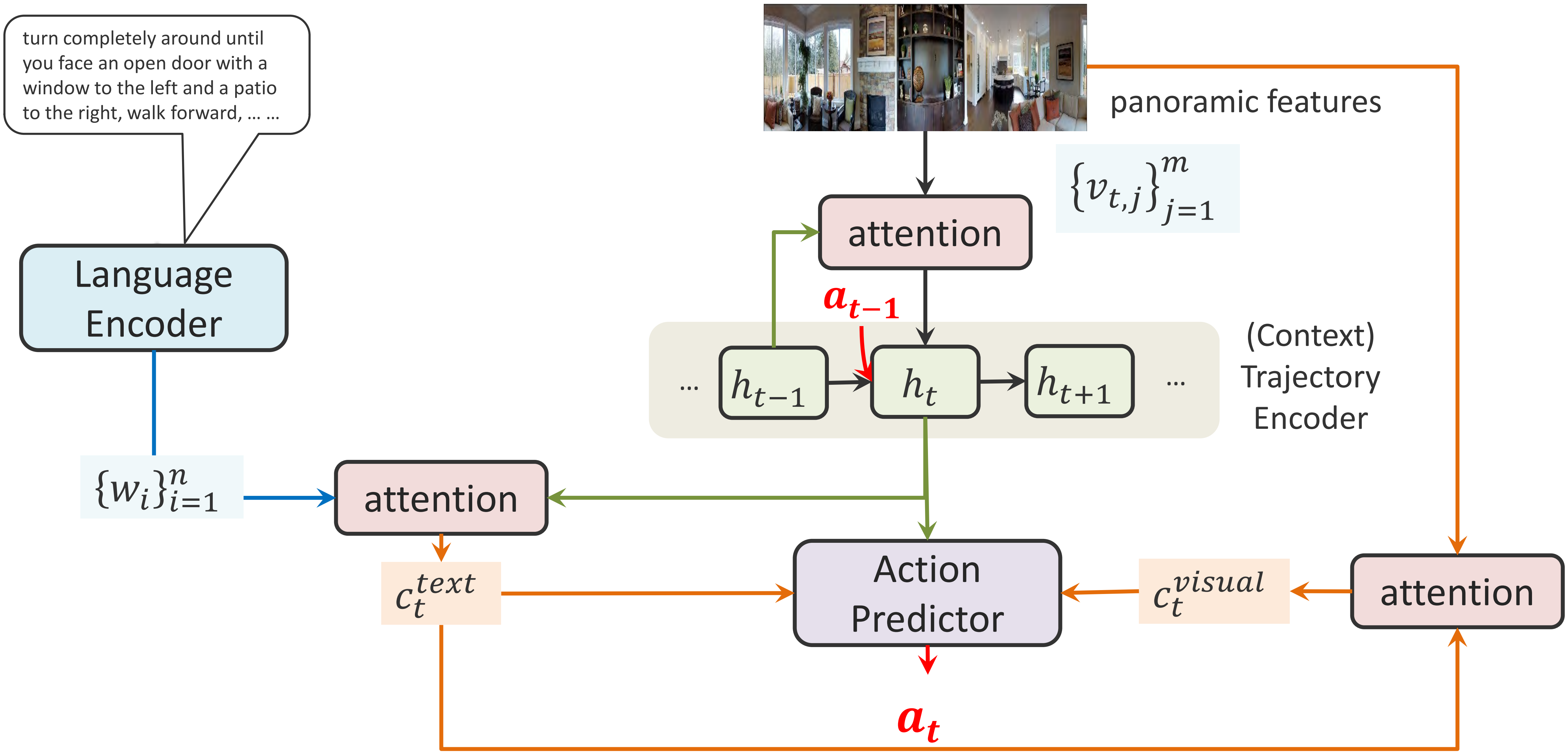
Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation, CVPR 2019

Learns to ground language in visual context using RL and Self-Supervised Imitation Learning

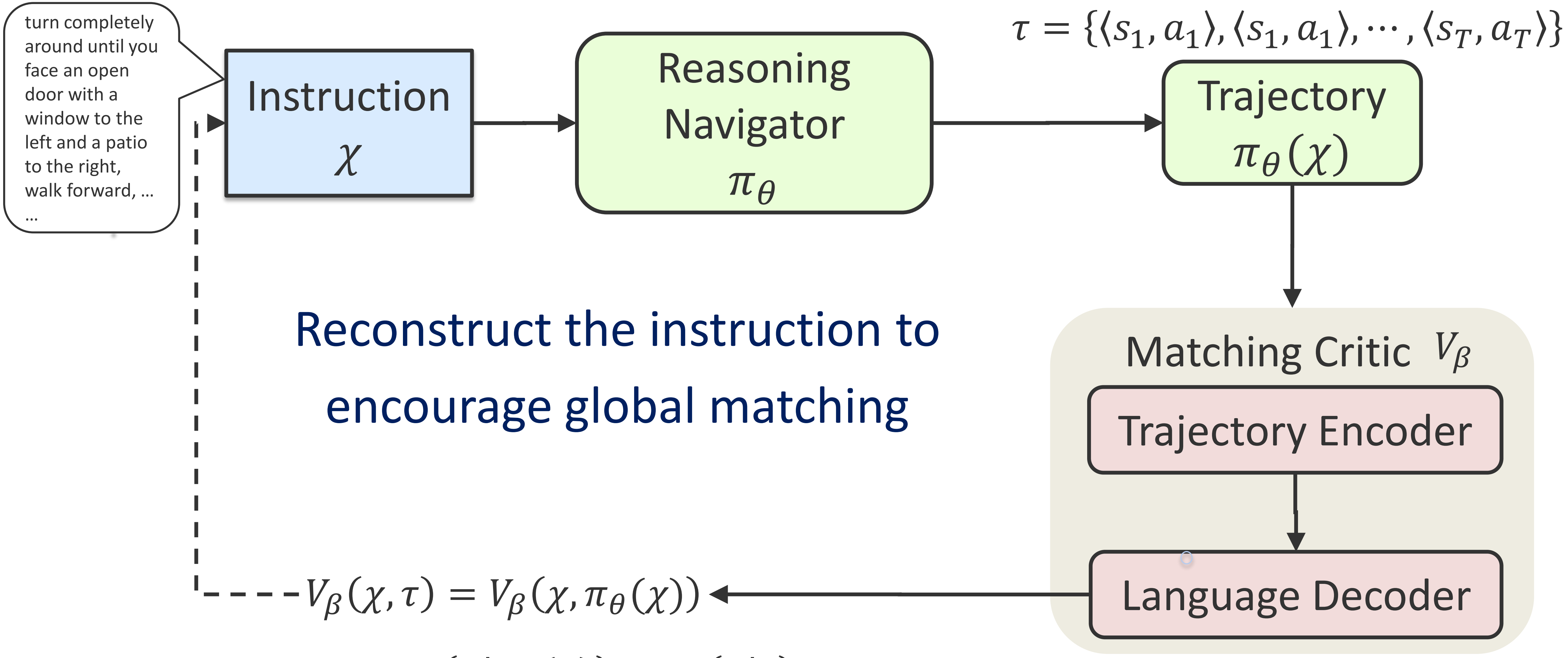
Reinforced Cross-modal Matching (RCM)



Cross-Modal Reasoning Navigator



Matching Critic → Intrinsic Reward



Reconstruct the instruction to encourage global matching

$$R_{intr} = p_\beta(\chi|\pi_\theta(\chi)) = p_\beta(\chi|\tau)$$

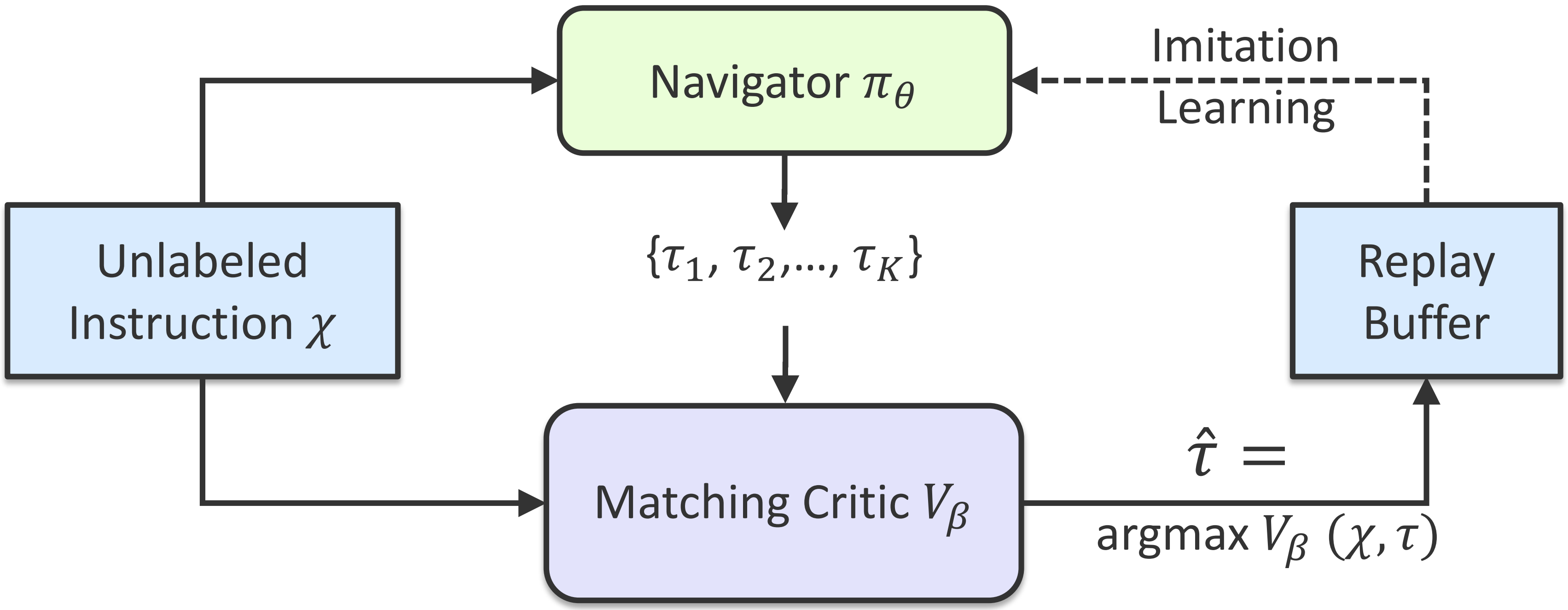
Visual-Language Navigation Task Challenges

(1) cross-modal grounding

(2) ill-posed feedback

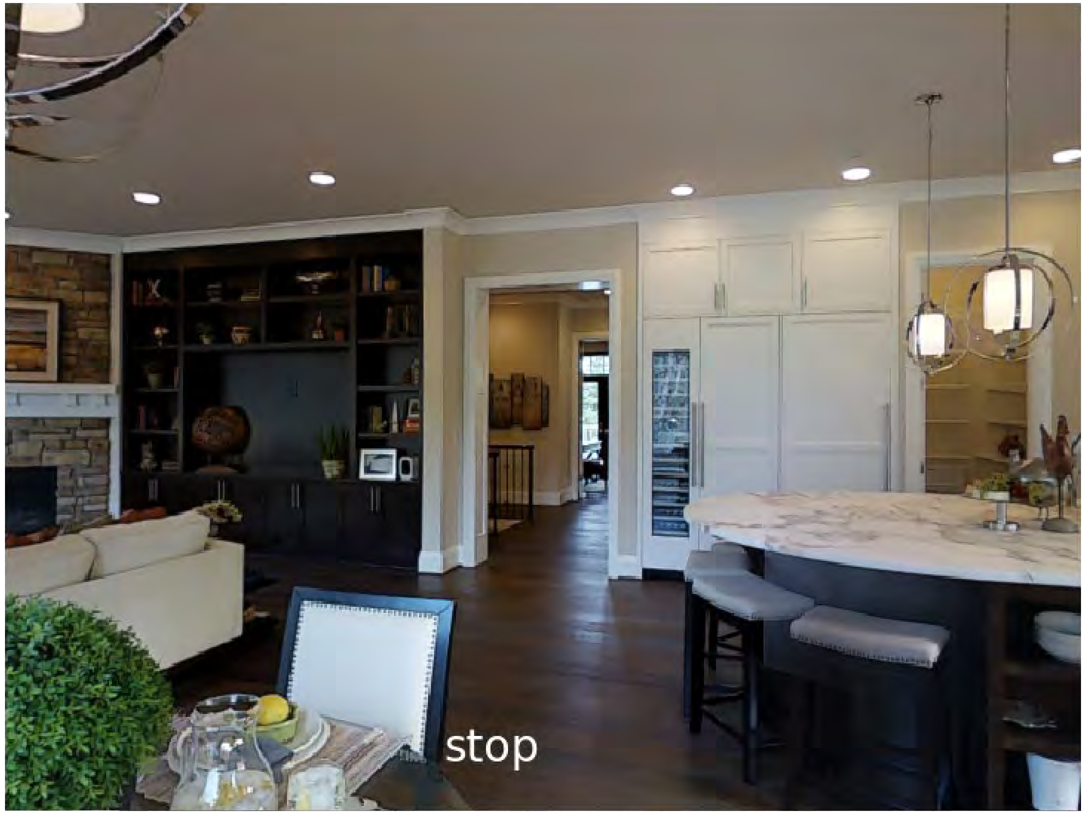
(3) generalization

Self-supervised Imitation Learning (SIL)



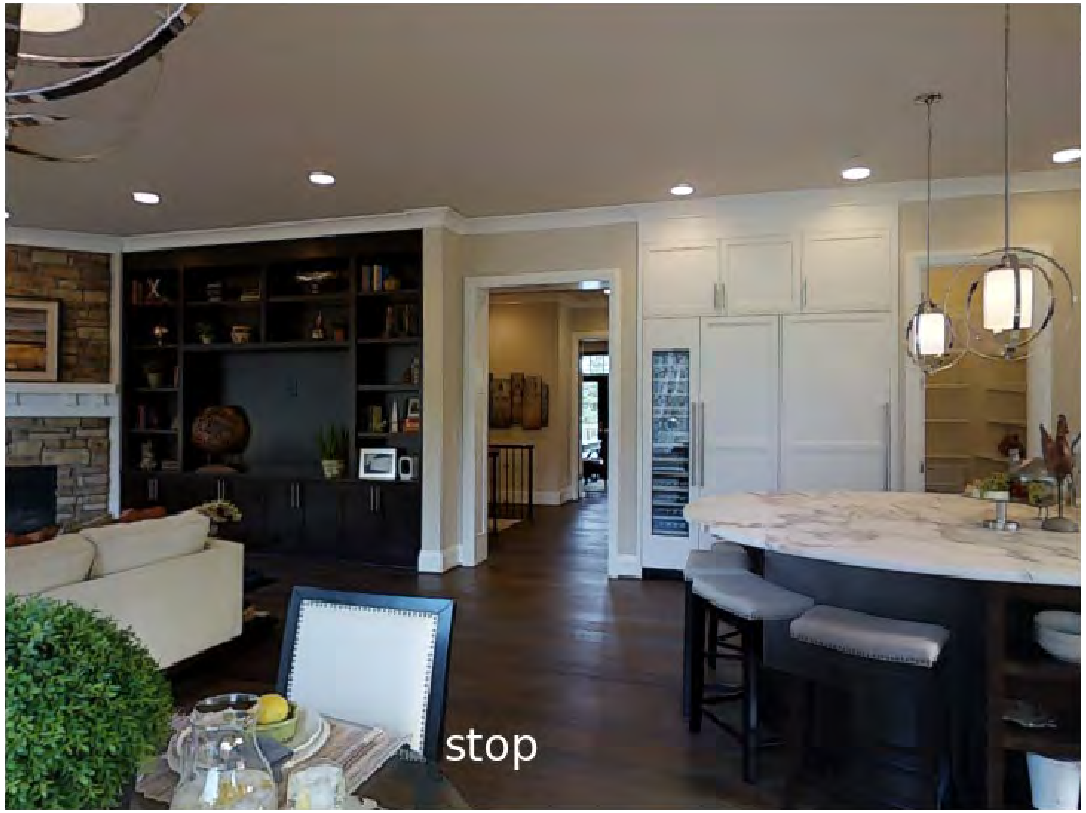
Learning from its previous good behaviors → better policy that adapts to new environments

Instruction: Turn right and head towards the kitchen. Before you get to the kitchen, turn left and enter the hallway. ... Walk forward and stop beside the bottom of the steps ***facing the double white doors***.

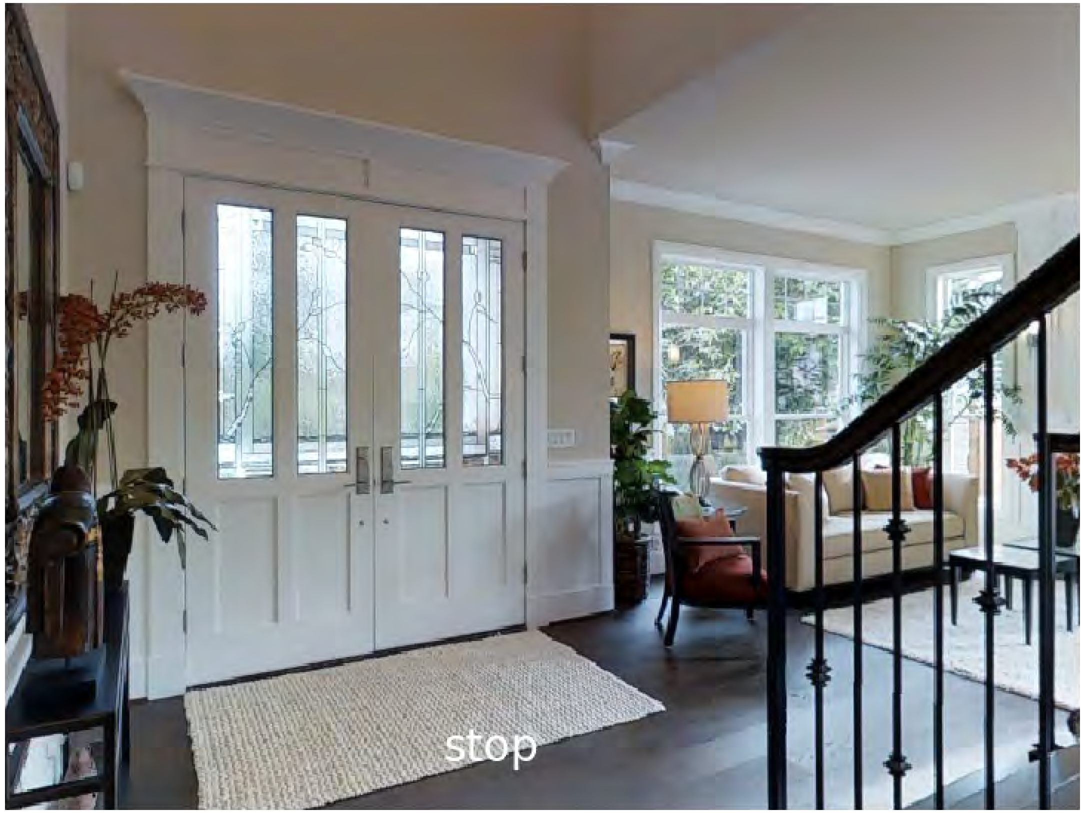


Reinforcement Learning Only

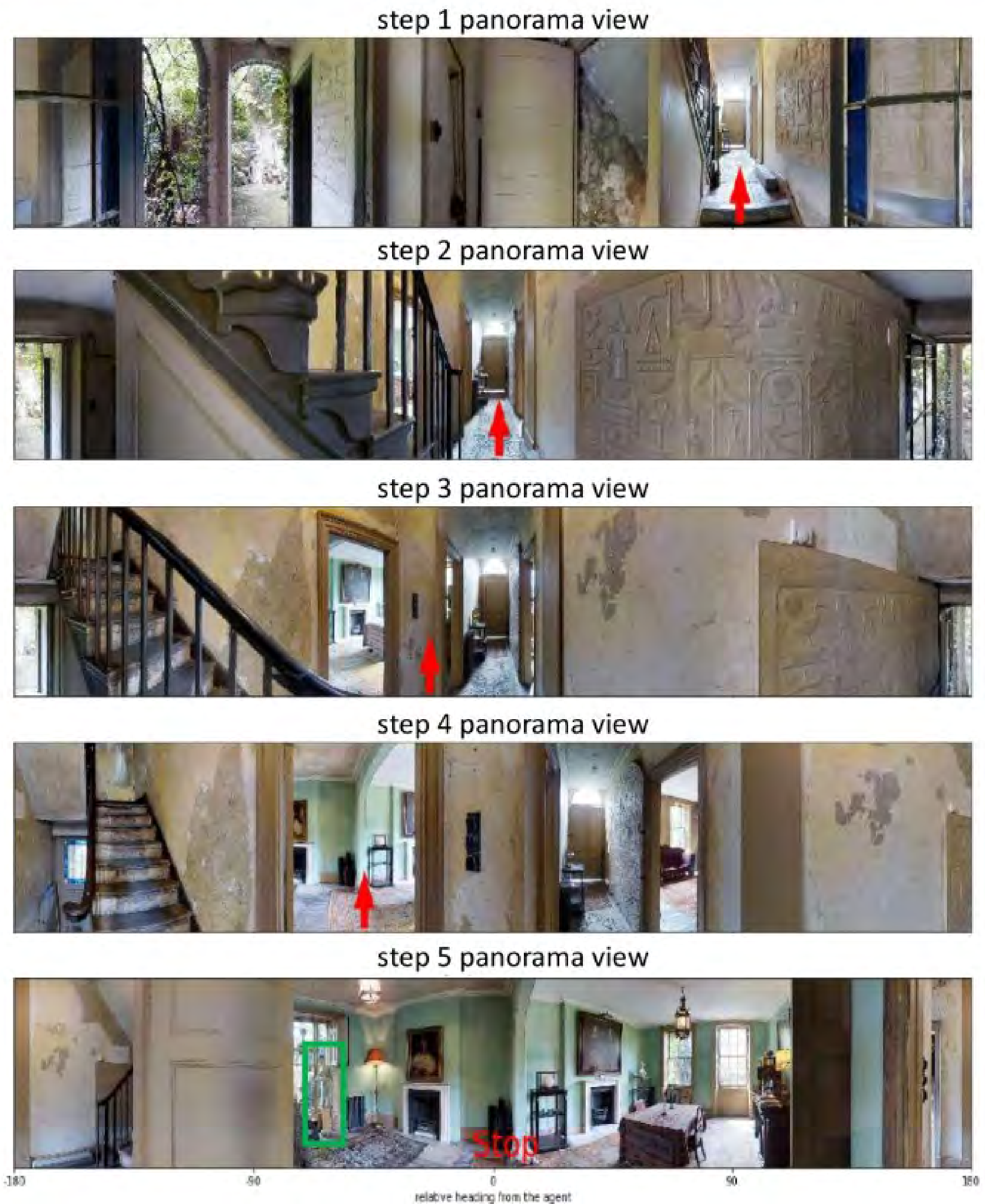
Instruction: Turn right and head towards the kitchen. Before you get to the kitchen, turn left and enter the hallway. ... Walk forward and stop beside the bottom of the steps ***facing the double white doors***.



Reinforcement Learning Only



RL + Self-Supervised Imitation Learning



Instruction: Go up the stairs to the right, turn left and go into the room on the left. Turn left and stop near the **mannequins**.

Intrinsic Reward : 0.51

Result : Failure (error = 3.1m)

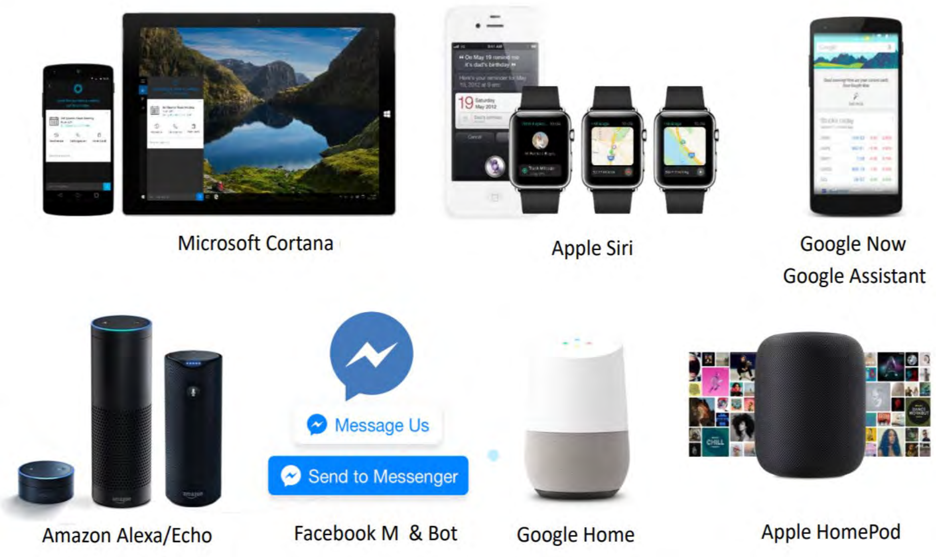

Best Student Paper Award
Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation
Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, Lei Zhang

What’s next ?

Image Source : Microsoft/HoloLens web page

Image Source : mirror.co.uk

Situated Reasoning Machines



Language Empowered Agents

Bi-directional but not situated!

Situated Language Empowered Agents

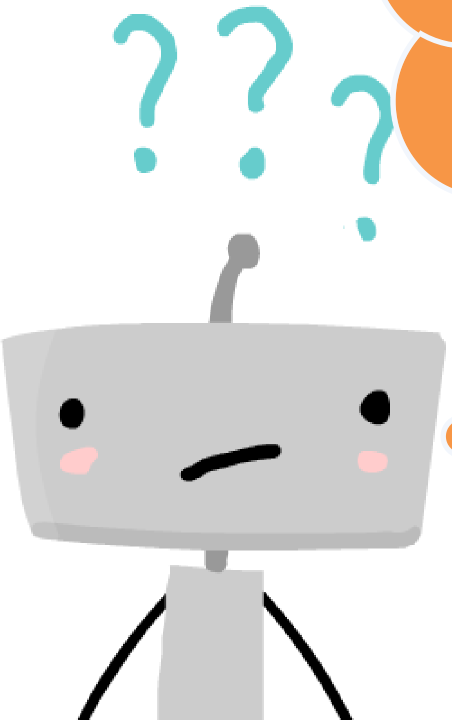
Situated but uni-directional !

Situated Reasoning Machines

Bi-directional and situated!

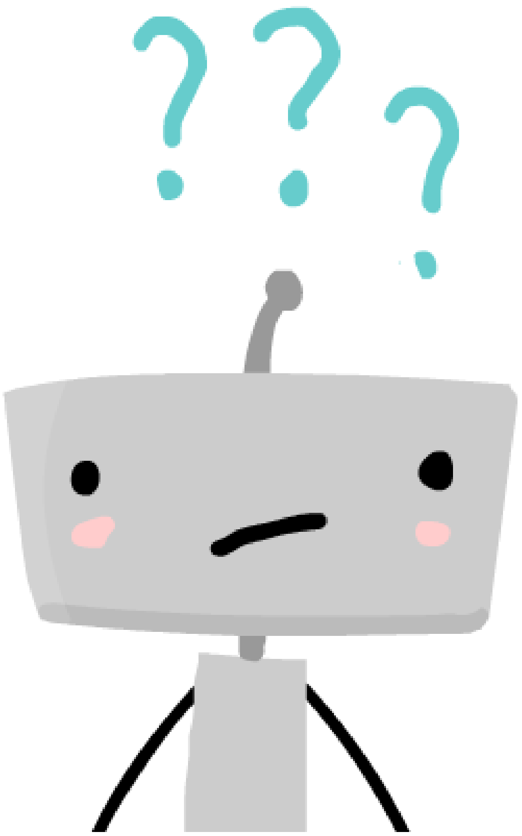
Grounding via Interaction

Make sure to clean behind the couch



Cool, wait! Which? Where? Hellllllllpppppp Humans are the worst

Vision and Dialog Navigation

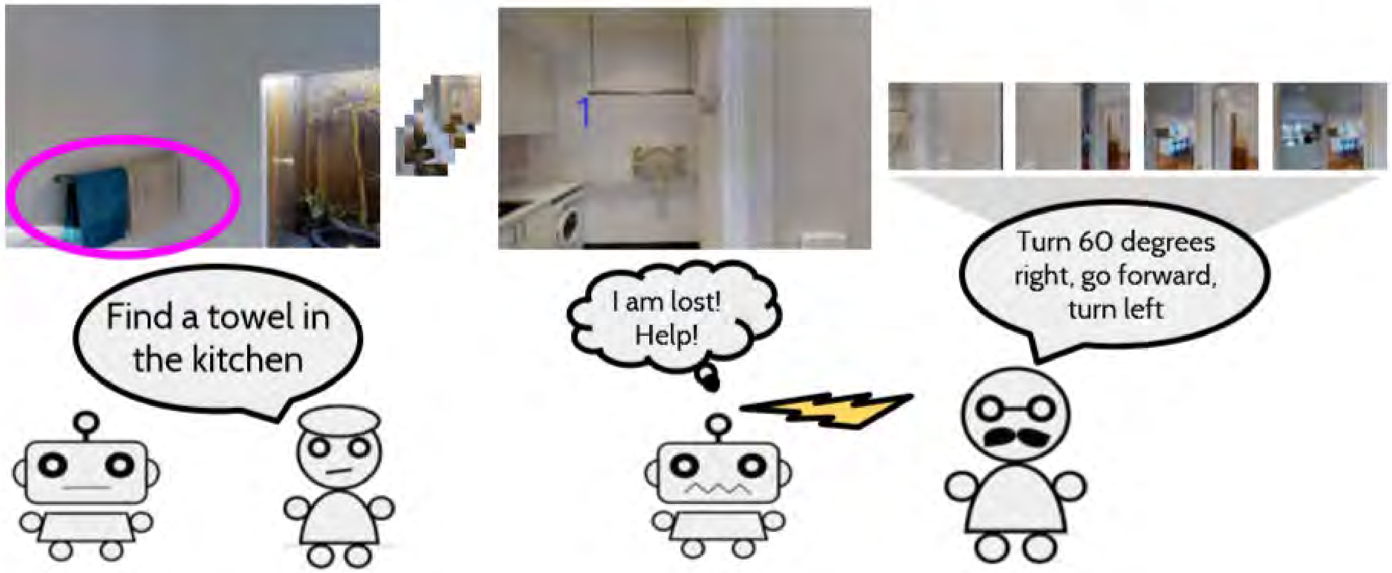


- **Connecting Language and Vision**

  What's the meaning of "the second door on the right?"
- **Modeling uncertainty**

  How does an agent know if it's lost or confused?
- **NL Question and Answer generation**

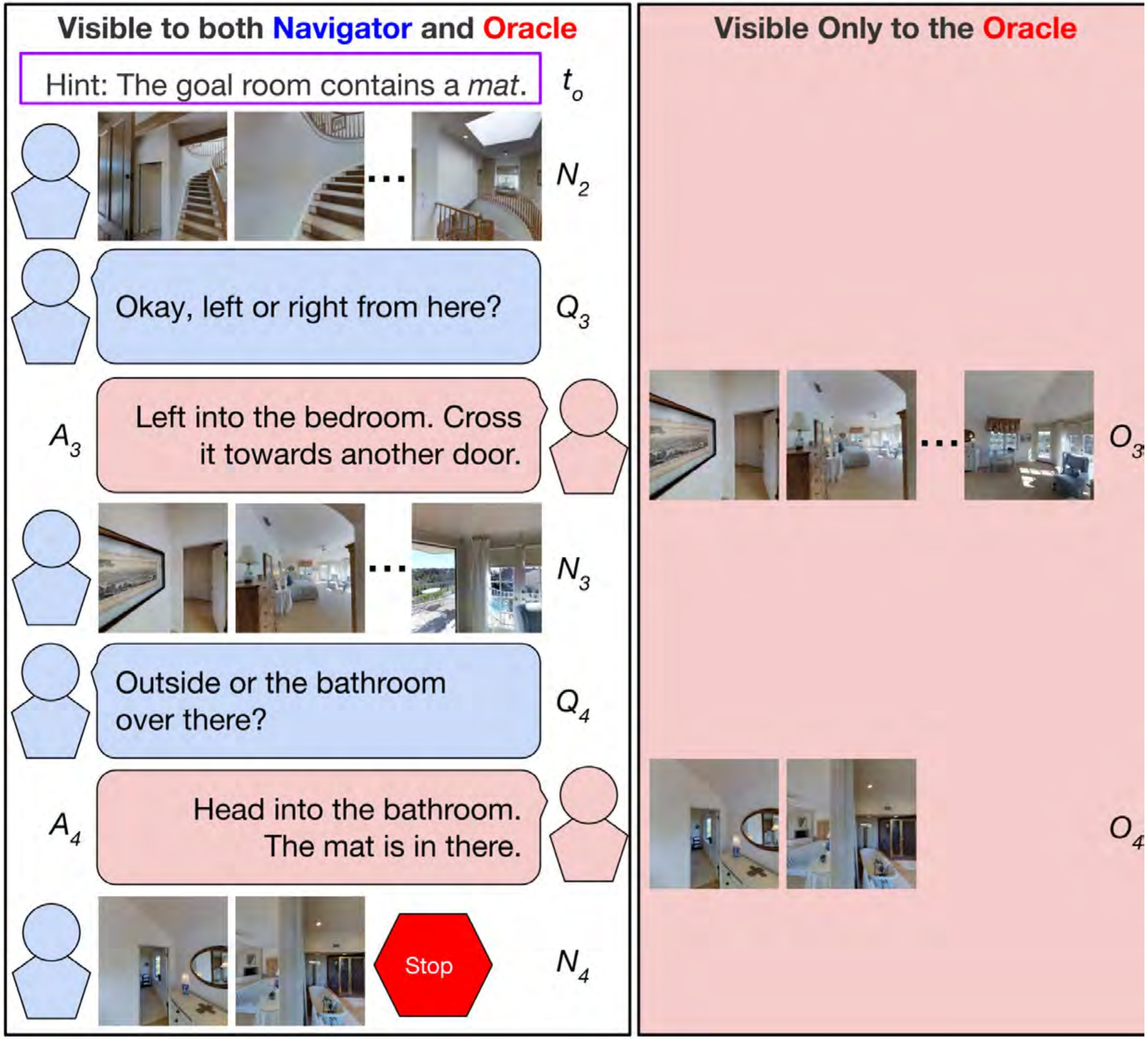
  Provide targeted feedback



Vision-based Navigation with Language-based Assistance via Imitation Learning with Indirect Intervention
CVPR 2019

Data + Model

K. Nyugen (UMD), **D. Dey (MSR), C. Brocket (MSR), B. Dolan (MSR)**



Interaction Snapshot

Goal: Build both the Navigator and the Oracle systems

Current SOTA? 0%
Brand new dataset!

Briefly …

- Situated Unidirectional Task: Visual Language Navigation
  - **RL agent** navigating **3D** environment
  - **Cycle loss** to evaluate local and global path behavior
  - Imitation learning via **self supervision**

- Situated Bi-directional Task : Visual+Dialog Navigation (VDN)
  - Learn to ask questions
  - Transfer from previous tasks : Unimodal Dialog, Visual Dialog, VLN, etc.
  - Meta-learn !

Thank you !

QUESTIONS