CMPT 983

Grounded Natural Language Understanding

March 16, 2022

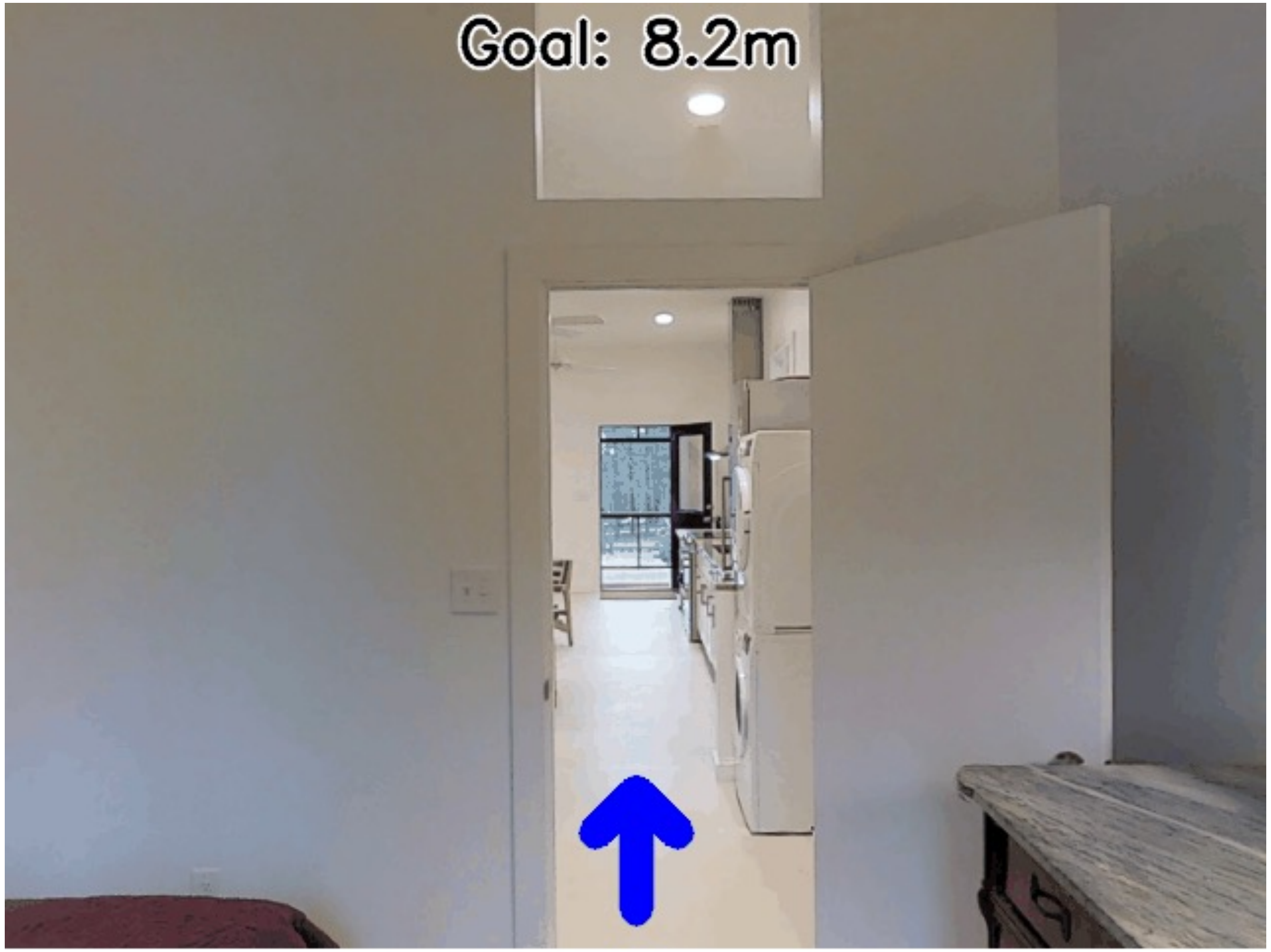
Instruction following for Visual Language Navigation

Task

Instruction-guided
Visual Navigation

Slide credit: Stefan Lee

Instruction-guided Visual Navigation



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

Slide credit: Stefan Lee

Instruction-guided Visual Navigation

Major Settings

Vision-and-Language Navigation

- **Indoor environments from the Matterport3D dataset + human directions**
- Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments arxiv.org/abs/1711.07280

StreetLearn

- **Google Street View + Google Maps directions**
- The StreetLearn Environment and Dataset arxiv.org/abs/1903.01292
- Learning To Follow Directions in Street View arxiv.org/abs/1903.00401
- **Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments** arxiv.org/abs/1811.12354

LANI

- **Simulated quadcopter in an open environment with landmark objects**
- Mapping Navigation Instructions to Continuous Control Actions with Position-Visitation Prediction arxiv.org/abs/1811.04179

Slide credit: Stefan Lee

Vision-and-Language Navigation (VLN)

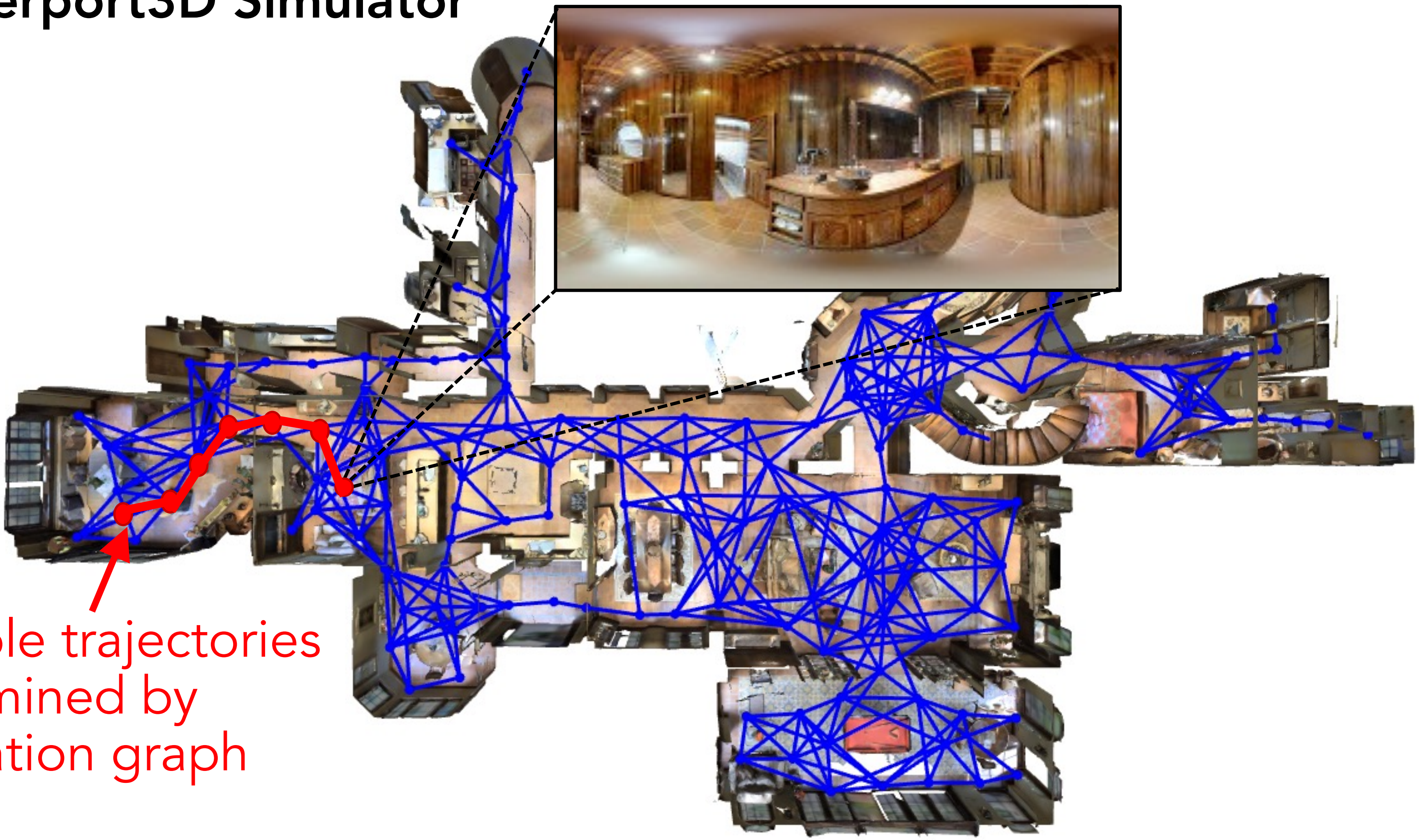
Matterport3D Simulator

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

Slide credit: Peter Anderson

Vision-and-Language Navigation (VLN)

Matterport3D Simulator



Feasible trajectories determined by navigation graph

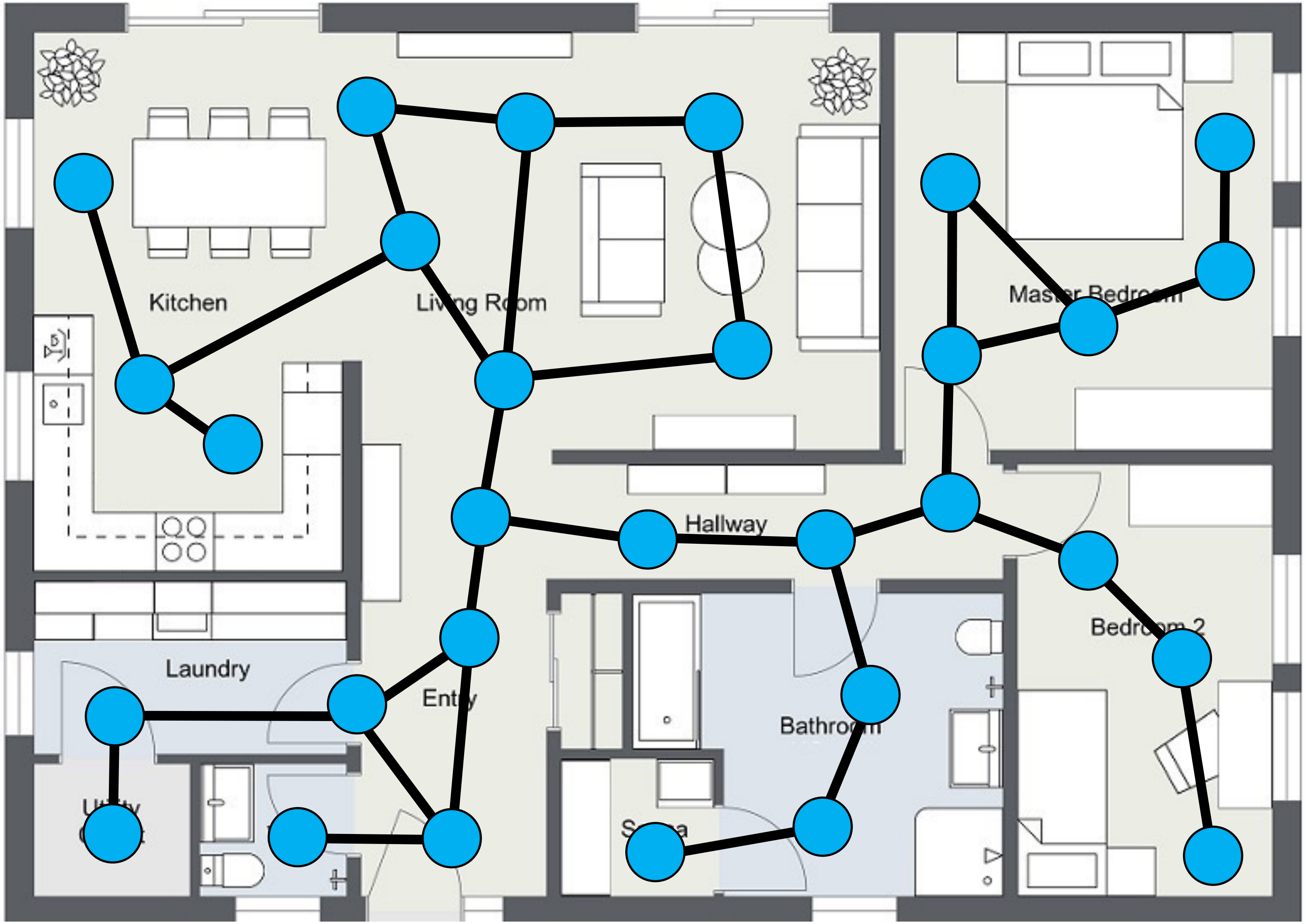
Slide credit: Peter Anderson

Vision-and-language Navigation (VLN)

Room2Room Dataset

Slide credit: Stefan Lee

VLN: Room2Room Dataset



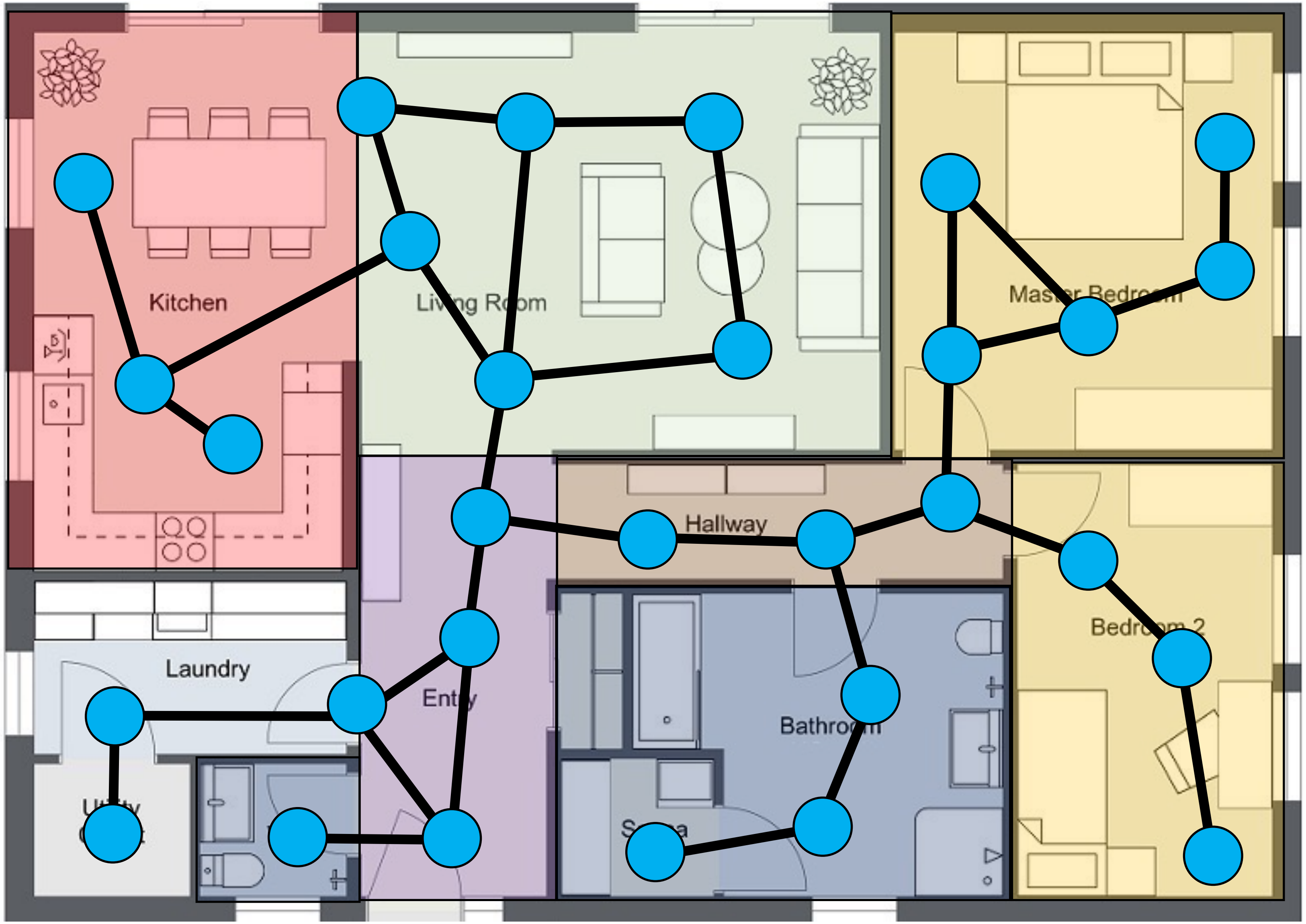
Nodes

- Panoramas
- 117 on average

Edges:

- Checks for clear ray-trace between nodes in the full mesh
- < 5 meters apart
- Manual cleaning
- Average degree 4.1

VLN: Room2Room Dataset



Nodes

- Panoramas
- 117 on average

Edges:

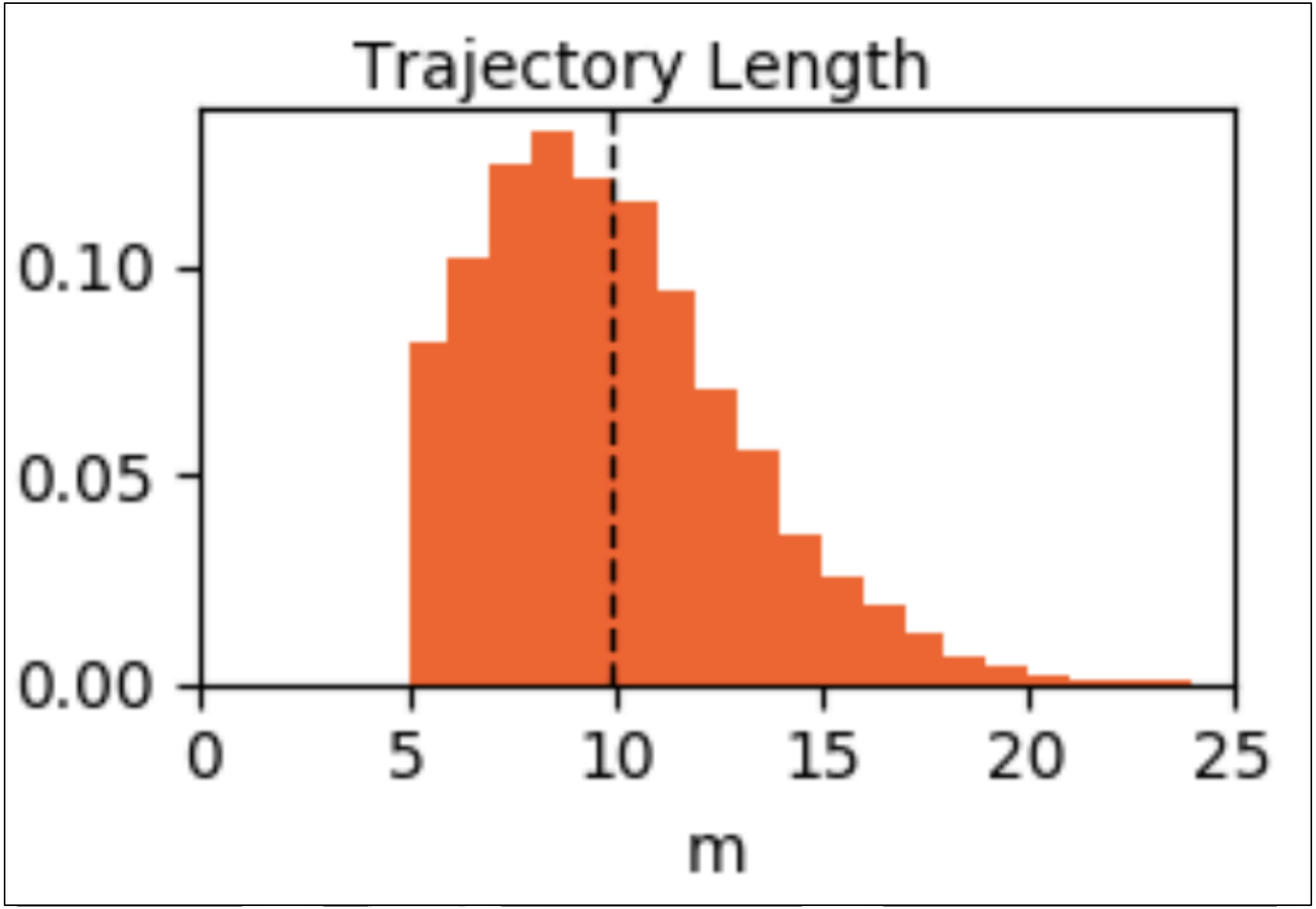
- Checks for clear ray-trace between nodes in the full mesh
- < 5 meters apart
- Manual cleaning
- Average degree 4.1

Paths:

- Two different rooms
- > 5 meters paths
- 4-6 edges

Slide credit: Stefan Lee

VLN: Room2Room Dataset



Nodes

- Panoramas
- 117 on average

Edges:

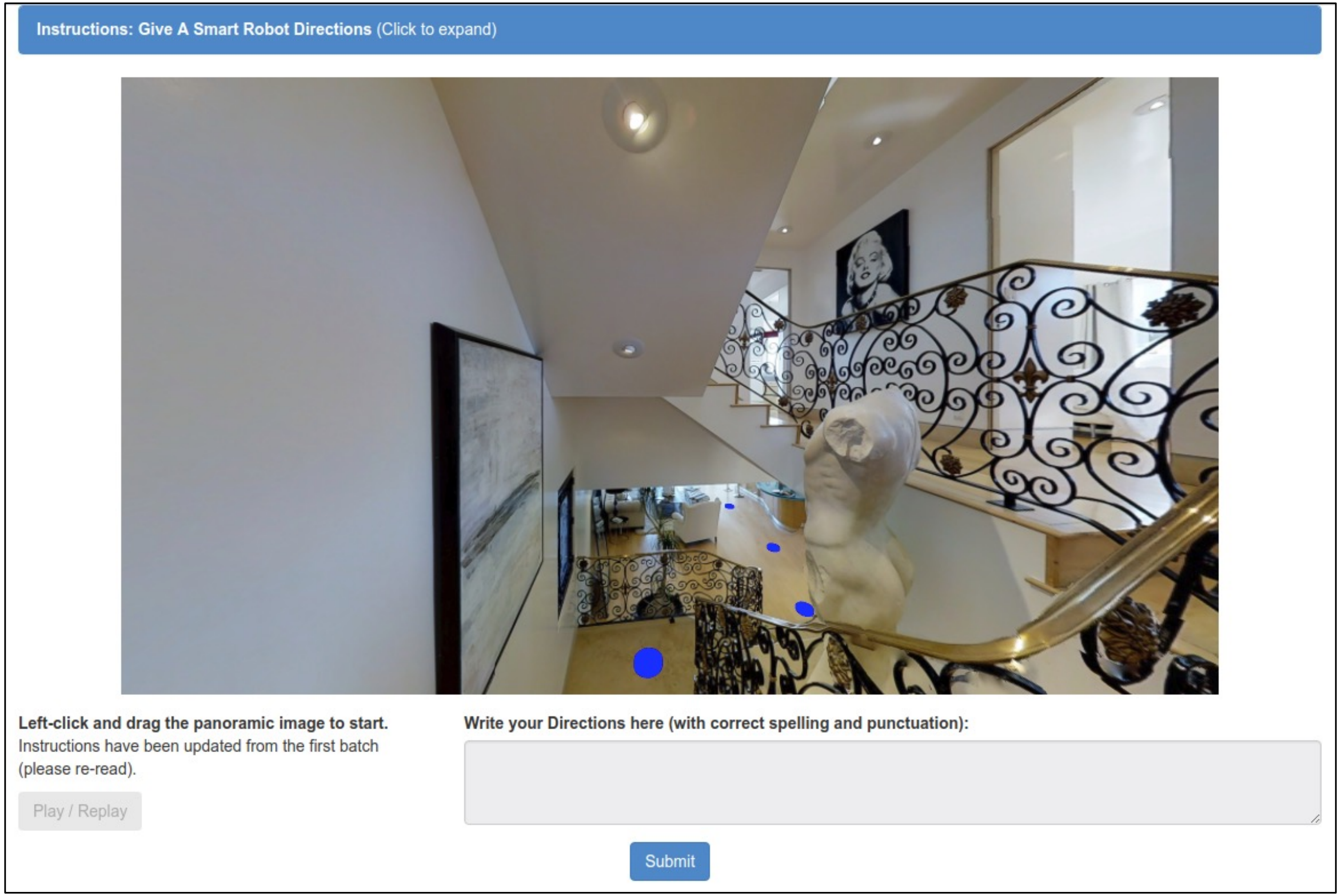
- Checks for clear ray-trace between nodes in the full mesh
- < 5 meters apart
- Manual cleaning
- Average degree 4.1

Paths:

- Two different rooms
- > 5 meters paths
- 4-6 edges

Slide credit: Stefan Lee

VLN: Room2Room Dataset



Instructions: Give A Smart Robot Directions (Click to expand)

Left-click and drag the panoramic image to start.
Instructions have been updated from the first batch (please re-read).

Write your Directions here (with correct spelling and punctuation):

Annotation Task:

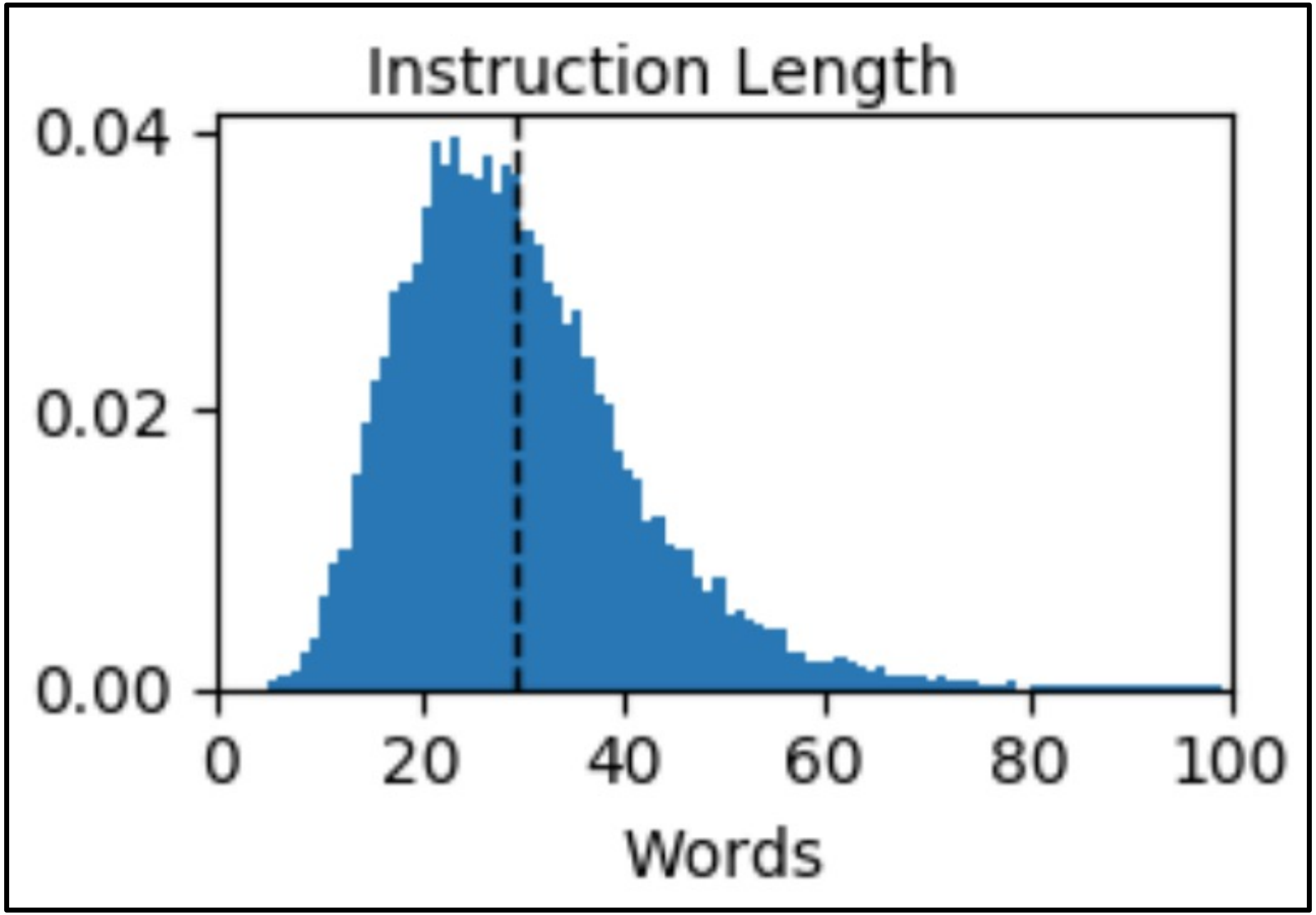
- Given a fly-through and pan/tilt controls, give natural language instruction to get to goal
- 3 workers per trajectory

Amazon Mechanical Turk:

- >400 US-based workers with strong HIT history
- 1600 hours of effort
- 21,567 instructions

Slide credit: Stefan Lee

VLN: Room2Room Dataset



Annotation Task:

- Given a fly-through and pan/tilt controls, give natural language instruction to get to goal
- 3 workers per trajectory

Amazon Mechanical Turk:

- >400 US-based workers with strong HIT history
- 1600 hours of effort
- 21,567 instructions

Slide credit: Stefan Lee

VLN: Room2Room Dataset

Instruction for same trajectory:

- Go past the ovens and the counter and wait just before you go outside.

- Walk through the kitchen towards the living room. Walk around the island and step onto the patio near the two chairs and stop in the patio doorway.

- Exit the kitchen by walking past the ovens and then head right, stopping just at the doorway leading to the patio outside.

Slide credit: Stefan Lee

VLN: Room2Room Dataset

Instruction for same trajectory:

- Turn and enter the living room area. Go past the table and sofas and stop in the foyer in front of the front door.

- Turn around and exit the room. Walk around the sofa and enter the hallway. Wait by the side table.

- Exit the room through the doorway nearest you, and continue into the adjacent room, exiting the room via the exit to your left.

Slide credit: Stefan Lee

VLN: Room2Room Dataset

Instruction for same trajectory:

- Walk along the insulated bare walls towards the window ahead in the next room. Walk through the unfinished room and through the door on the other side of the room that leads to a finished hallway. Walk into the first open door in the hall that leads to a bedroom with photo art on the wall near the entrance of classic black and white scenes.

- Walk forward past the window then turn right and enter the hallway. Enter the first bedroom on your right. wait near the bed.

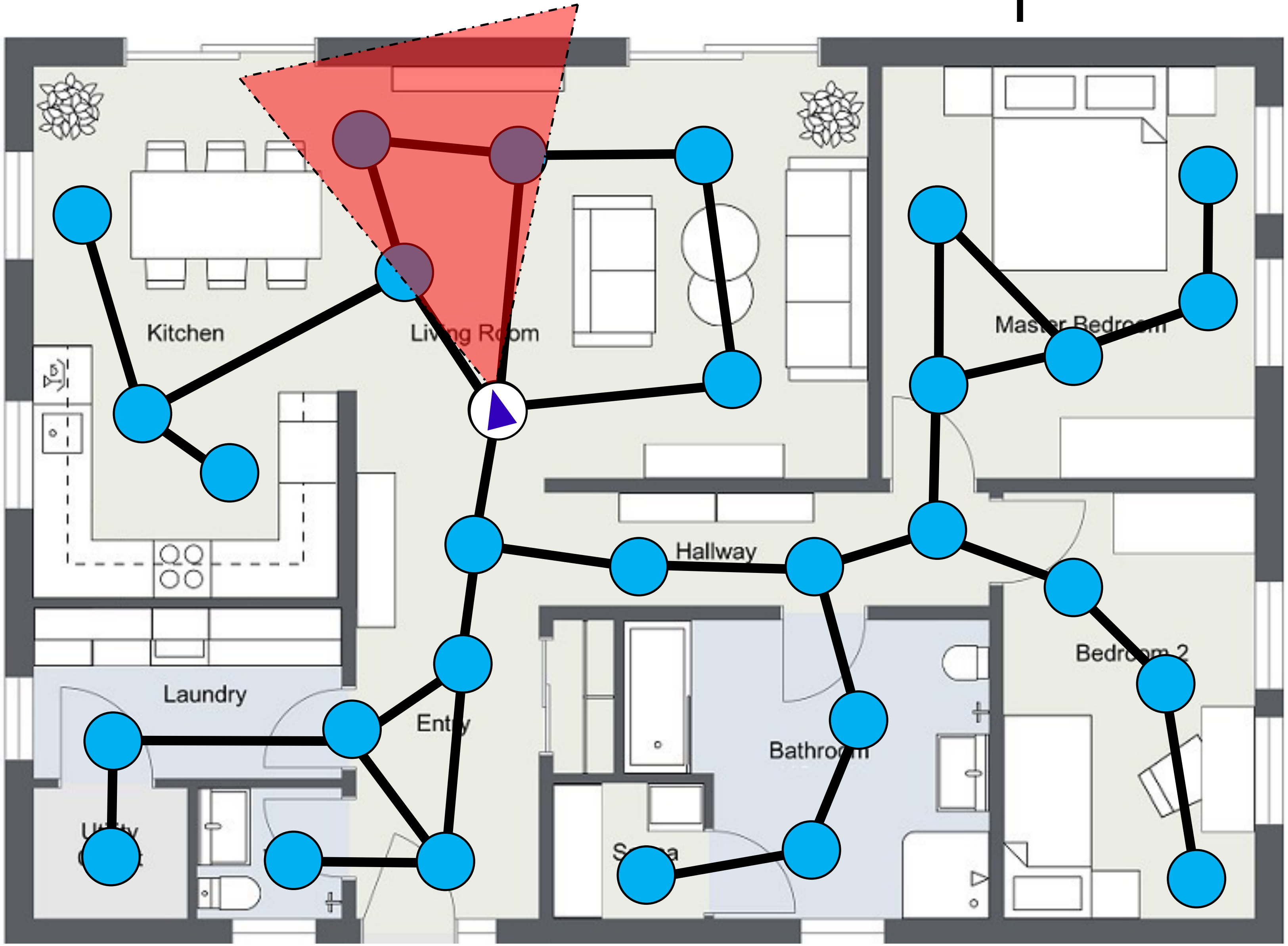
- Walk forward and take a right. Enter the hallway through the door on the right. Take the first left into a bedroom. Stop once you are in the bedroom.

Slide credit: Stefan Lee

Vision-and-language Navigation (VLN)
State and Action Space

Slide credit: Stefan Lee

VLN: State and Action Space



Agent

- Egocentric camera

Actions:

- Turn: left/right 30
- Tilt: up/down 30
- Forward (?)
- Stop

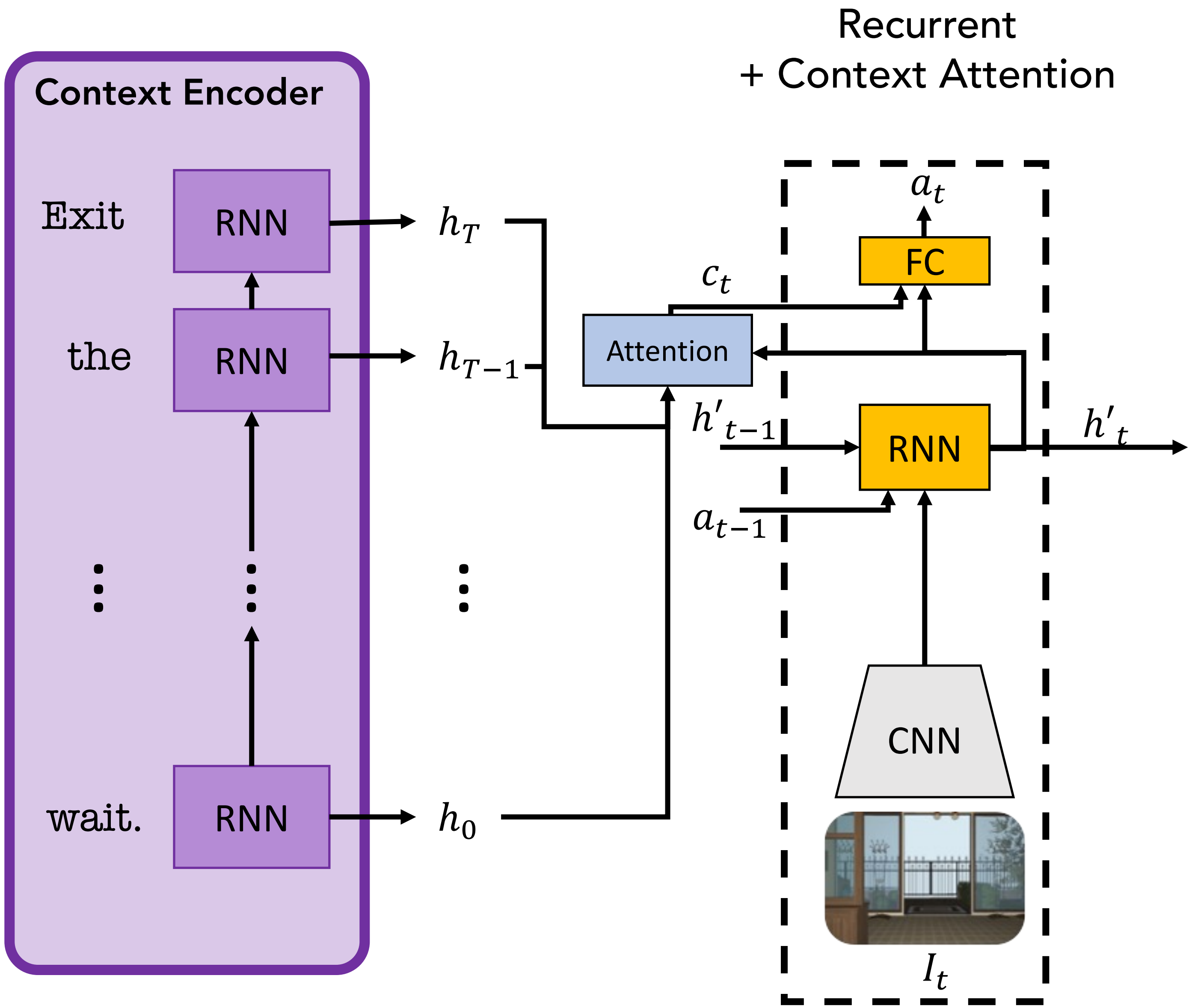
Slide credit: Stefan Lee

Vision-and-language Navigation (VLN)

Model and Training

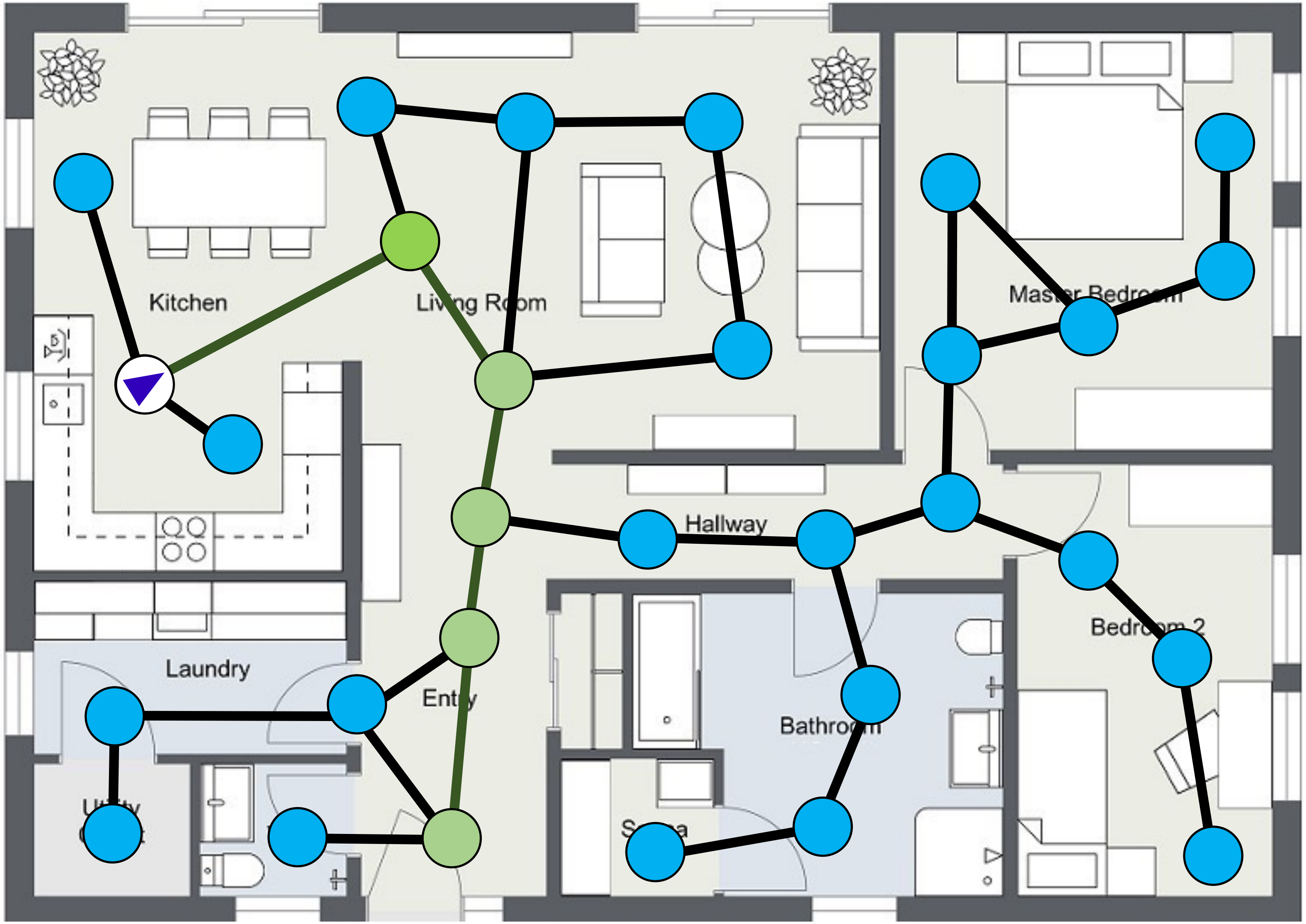
Slide credit: Stefan Lee

Our Attentive Recurrent Agent: Context Attention



Slide credit: Stefan Lee

VLN: State and Action Space

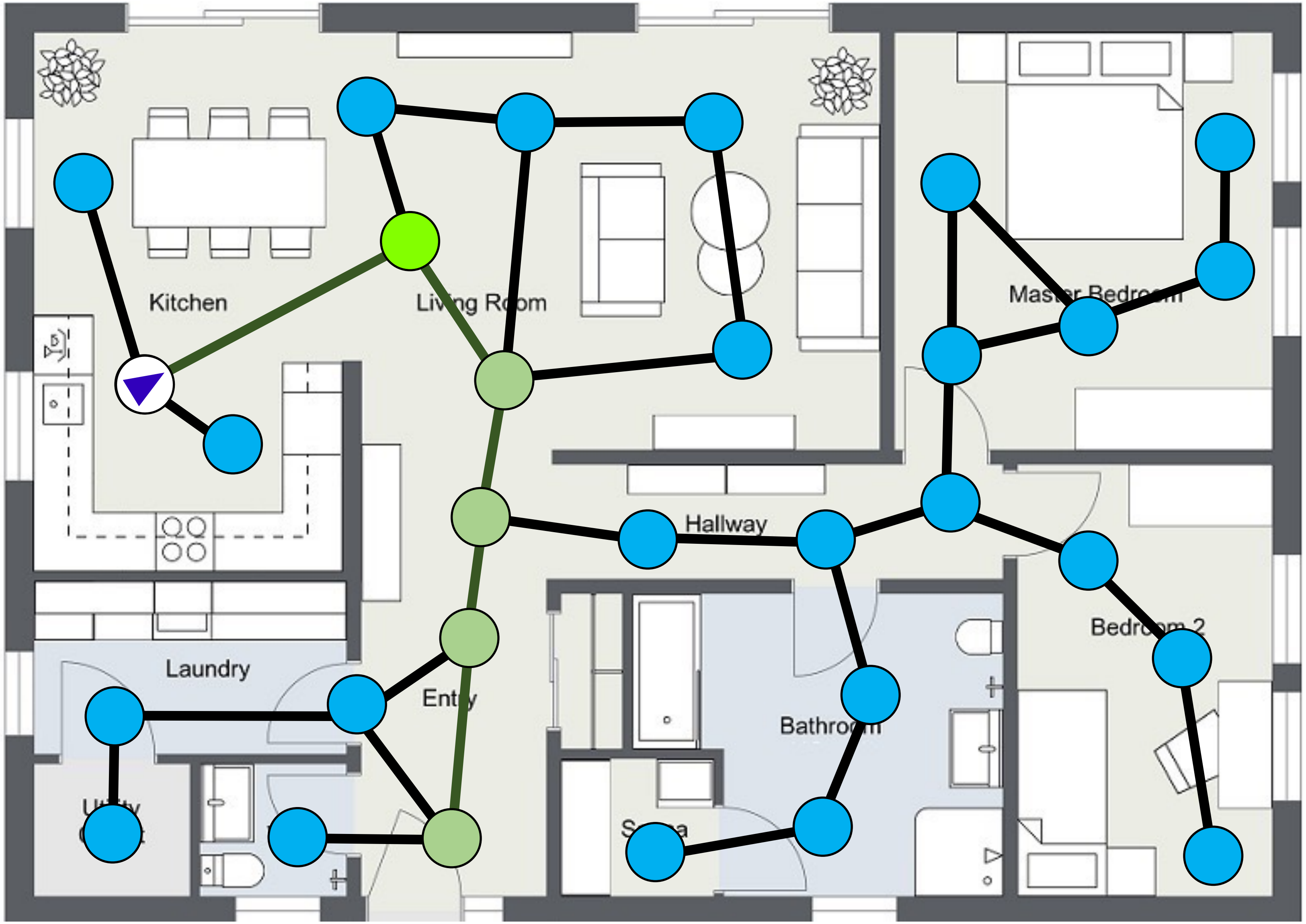


Teacher Forcing

- Ignore agent action, continue on GT path
- Just behavior cloning

Slide credit: Stefan Lee

VLN: State and Action Space

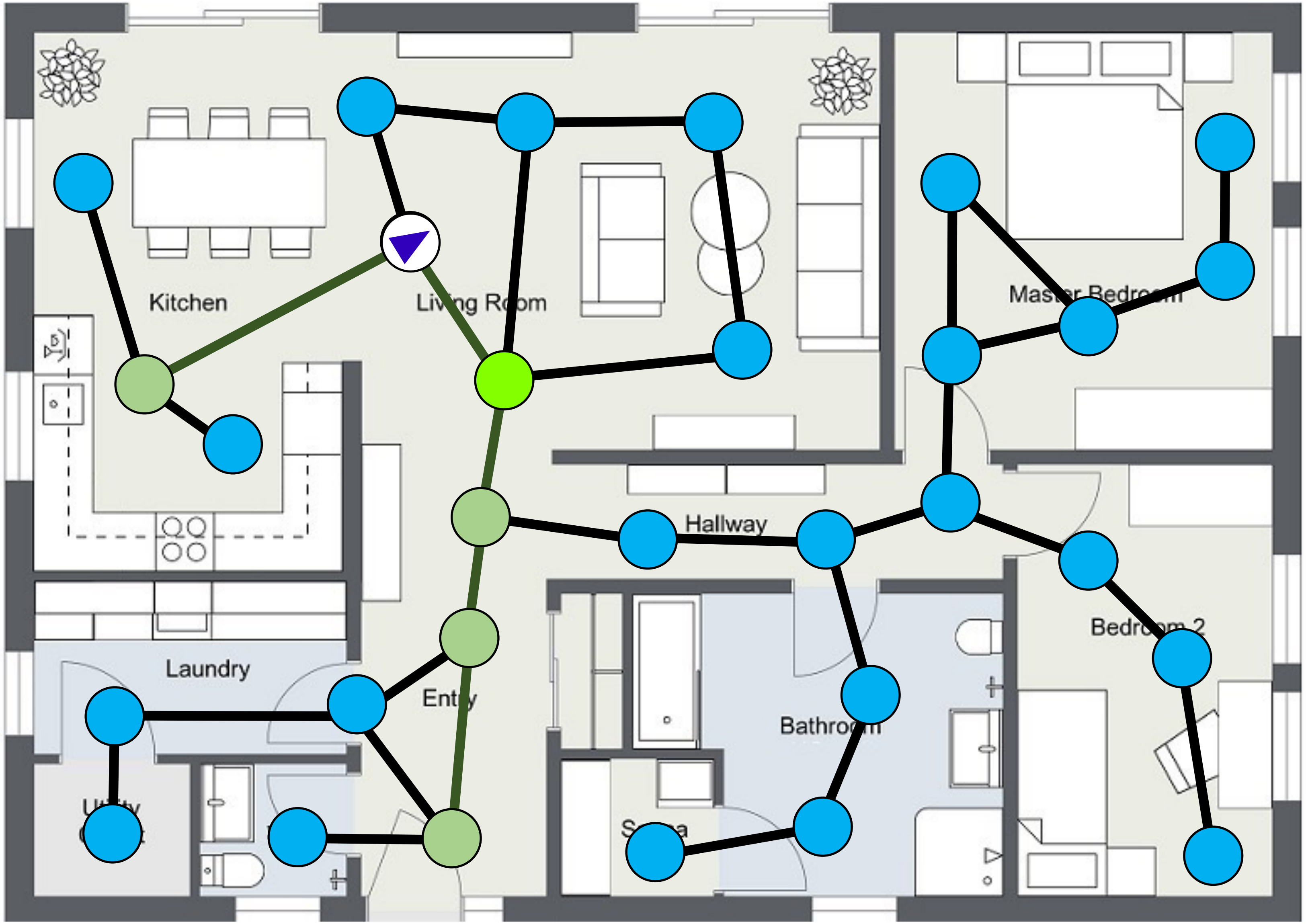


Teacher Forcing

- Ignore agent action, continue on GT path
- Just behavior cloning

Slide credit: Stefan Lee

VLN: State and Action Space

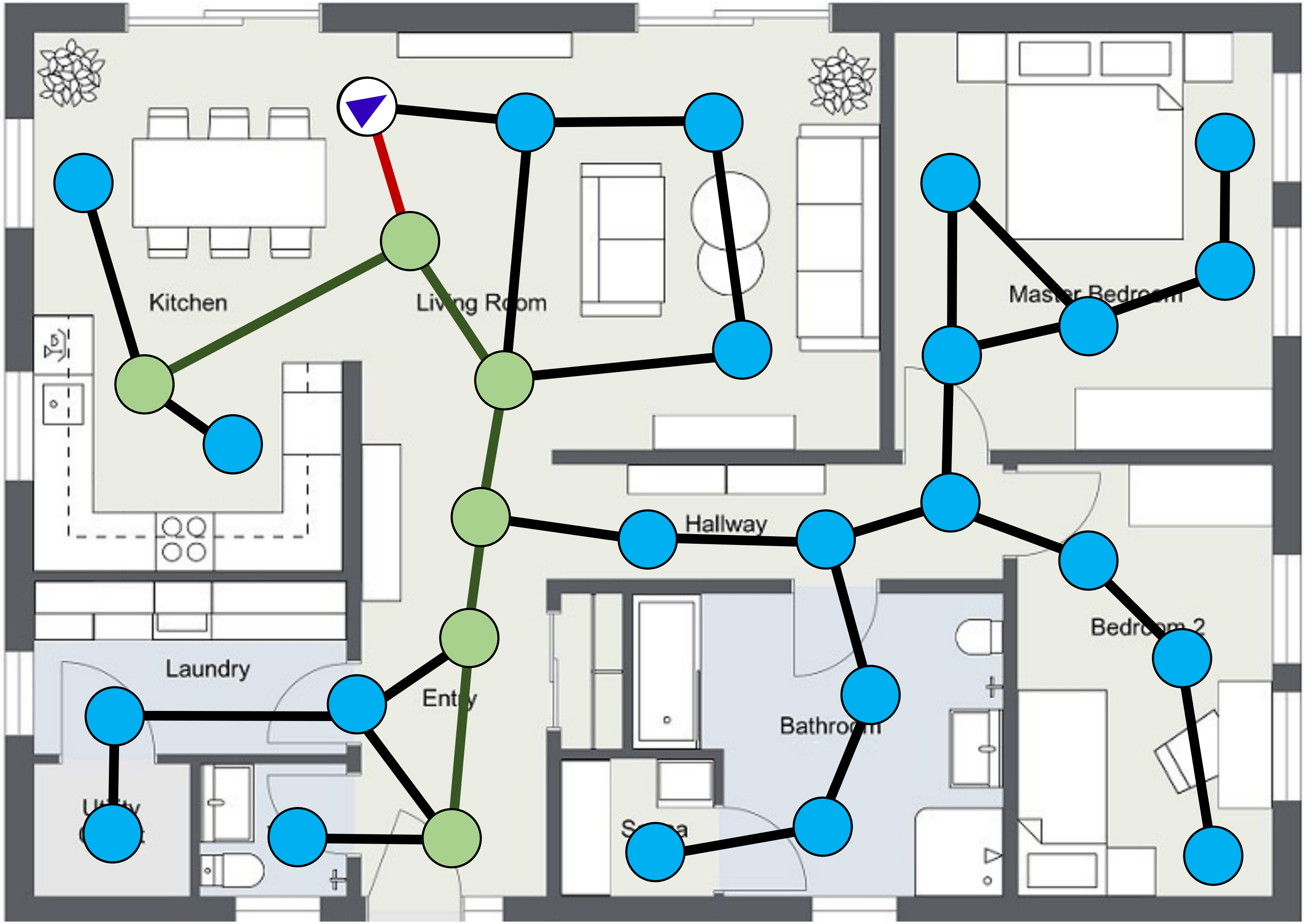


Teacher Forcing

- Ignore agent action, continue on GT path
- Just behavior cloning

Slide credit: Stefan Lee

VLN: State and Action Space

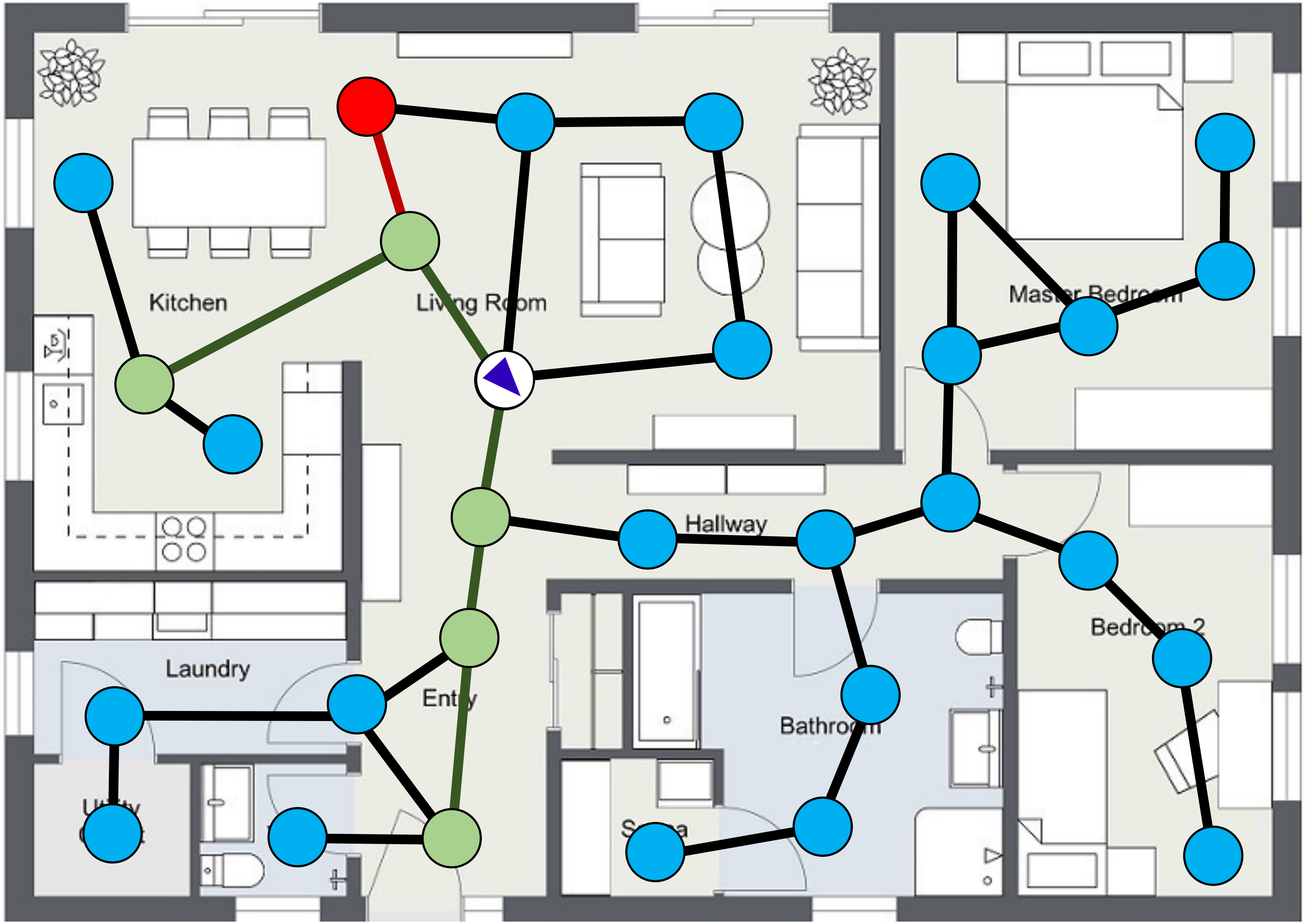


Teacher Forcing

- Ignore agent action, continue on GT path
- Just behavior cloning

Slide credit: Stefan Lee

VLN: State and Action Space

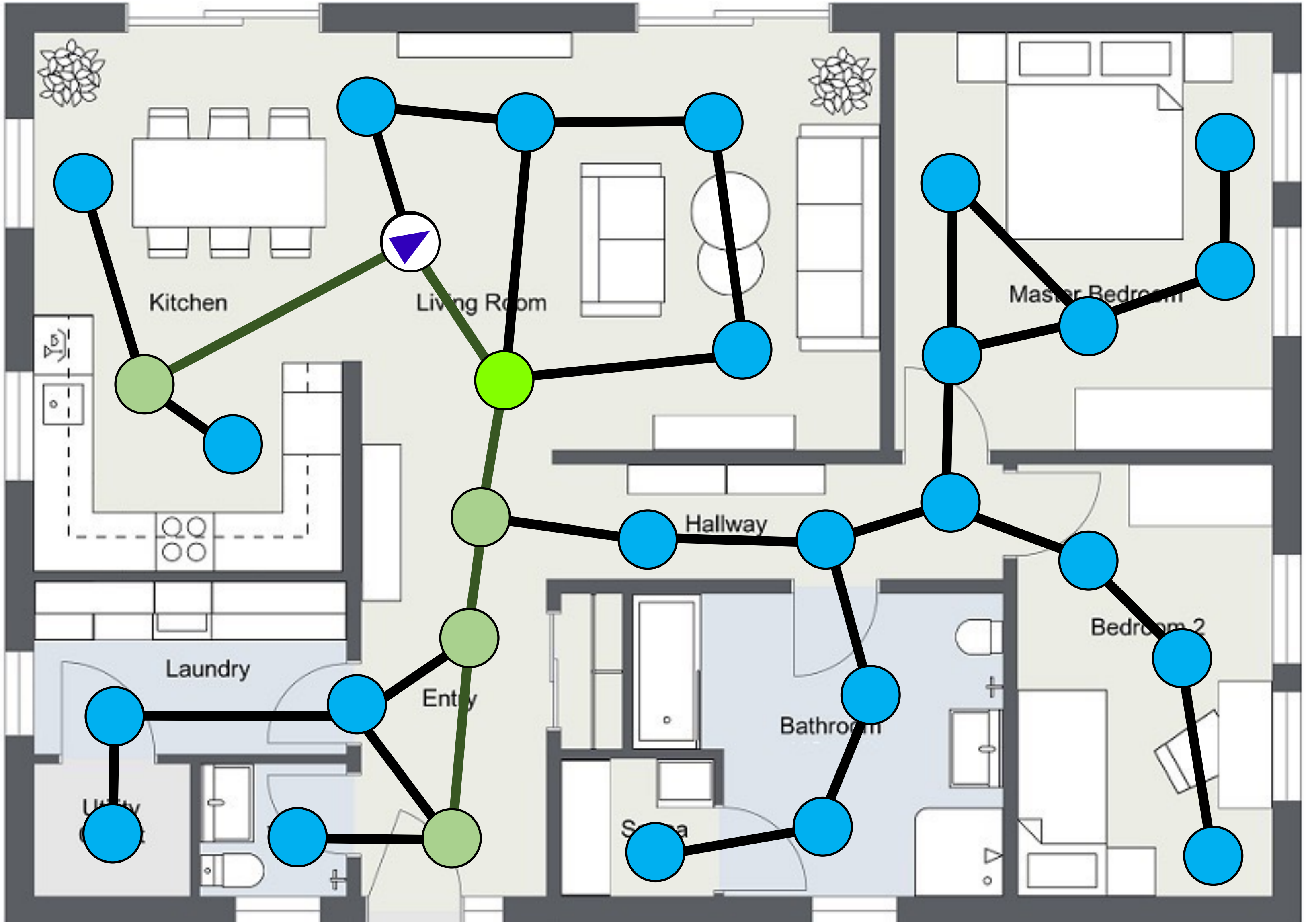


Teacher Forcing

- Ignore agent action, continue on GT path
- Just behavior cloning

Slide credit: Stefan Lee

VLN: State and Action Space



Teacher Forcing

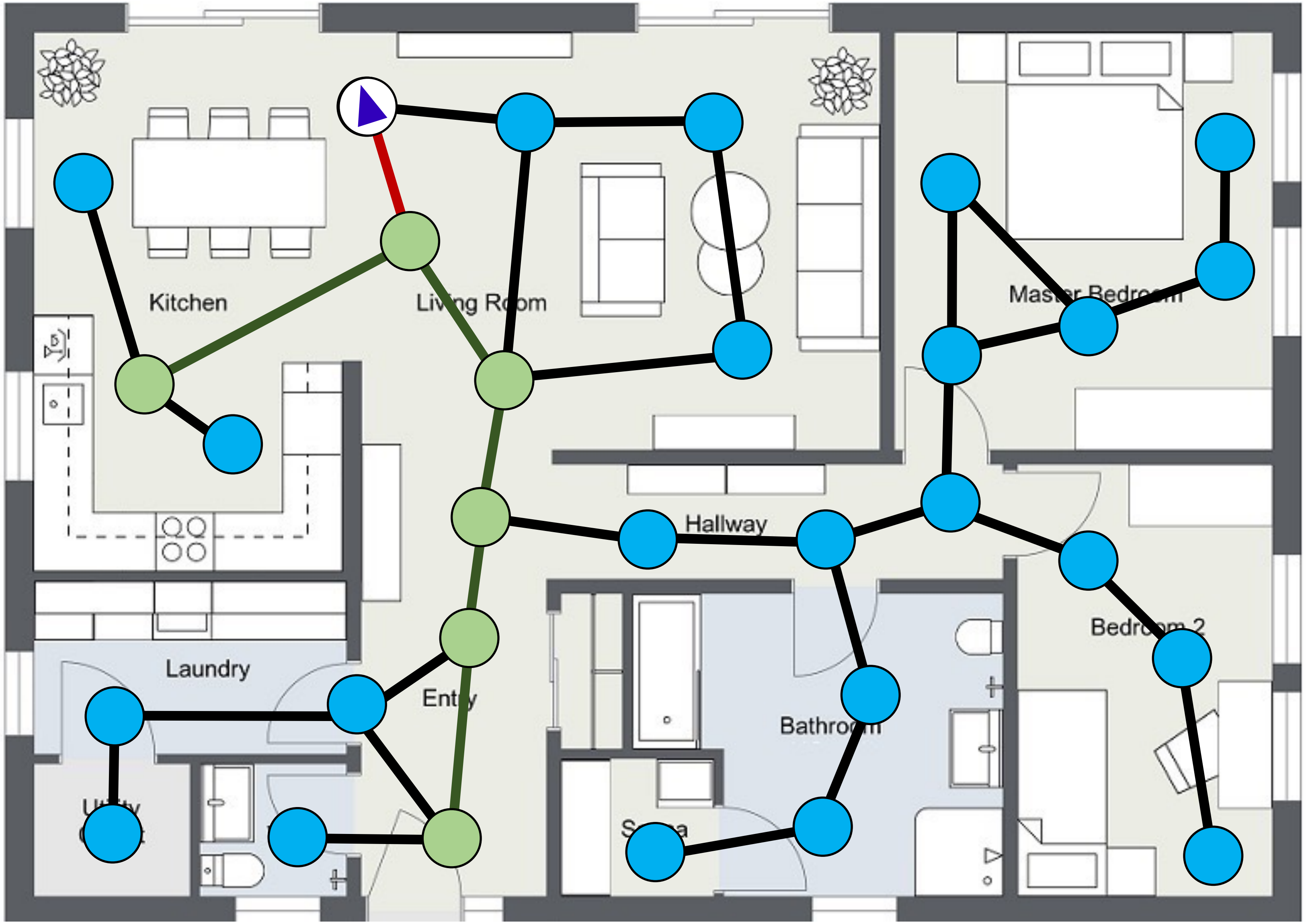
- Ignore agent action, continue on GT path
- Just behavior cloning

Student Forcing:

- Agent acts, oracle is queried to find next step
- Online DAGGER

Slide credit: Stefan Lee

VLN: State and Action Space



Teacher Forcing

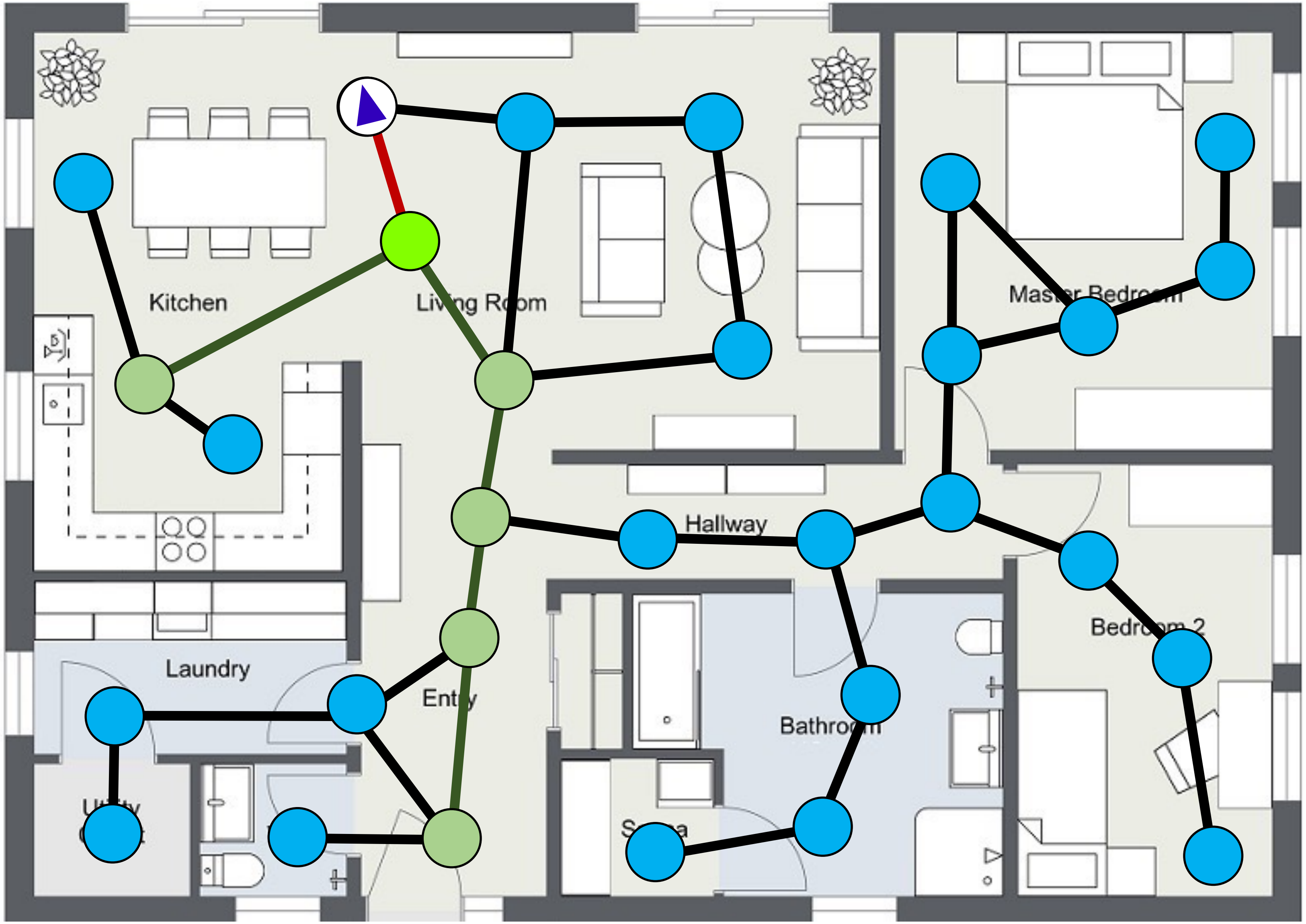
- Ignore agent action, continue on GT path
- Just behavior cloning

Student Forcing:

- Agent acts, oracle is queried to find next step
- Online DAGGER

Slide credit: Stefan Lee

VLN: State and Action Space



Teacher Forcing

- Ignore agent action, continue on GT path
- Just behavior cloning

Student Forcing:

- Agent acts according to its policy, oracle is queried to find next step back to path
- Online DAGGER

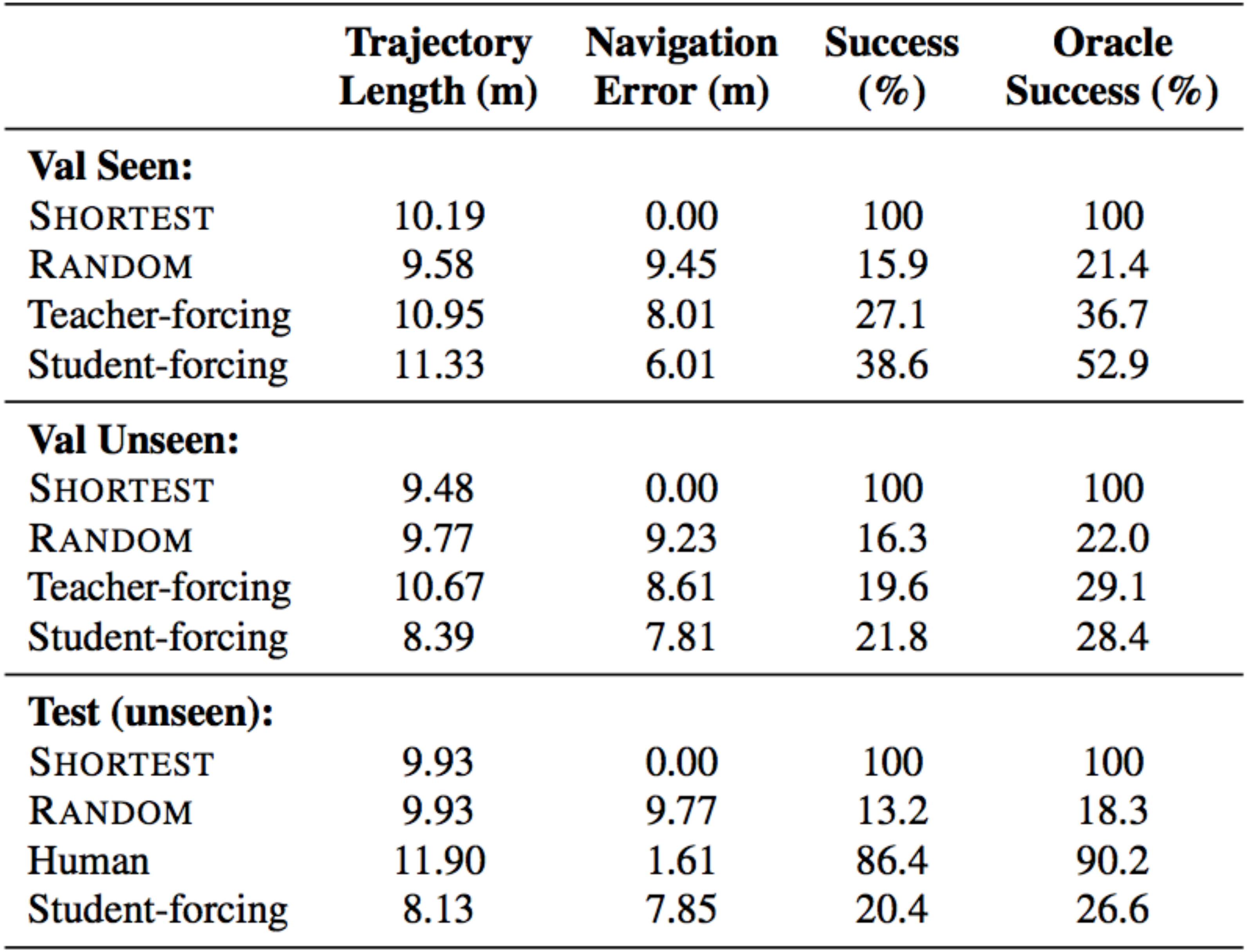
Slide credit: Stefan Lee

Vision-and-language Navigation (VLN)

Results

Slide credit: Stefan Lee

VLN: Results



	Trajectory Length (m)	Navigation Error (m)	Success (%)	Oracle Success (%)
Val Seen:				
SHORTEST	10.19	0.00	100	100
RANDOM	9.58	9.45	15.9	21.4
Teacher-forcing	10.95	8.01	27.1	36.7
Student-forcing	11.33	6.01	38.6	52.9
Val Unseen:				
SHORTEST	9.48	0.00	100	100
RANDOM	9.77	9.23	16.3	22.0
Teacher-forcing	10.67	8.61	19.6	29.1
Student-forcing	8.39	7.81	21.8	28.4
Test (unseen):				
SHORTEST	9.93	0.00	100	100
RANDOM	9.93	9.77	13.2	18.3
Human	11.90	1.61	86.4	90.2
Student-forcing	8.13	7.85	20.4	26.6

Slide credit: Stefan Lee

Vision-and-language Navigation (VLN)

Evaluation

Slide credit: Stefan Lee

Vision-and-Language Navigation Evaluation

Initial Metrics:

- Trajectory Length (m)
- Navigation Error (m)
- Success (%)
- Oracle Success (%)

Standard metrics for navigation tasks

Not the best for visual language navigation

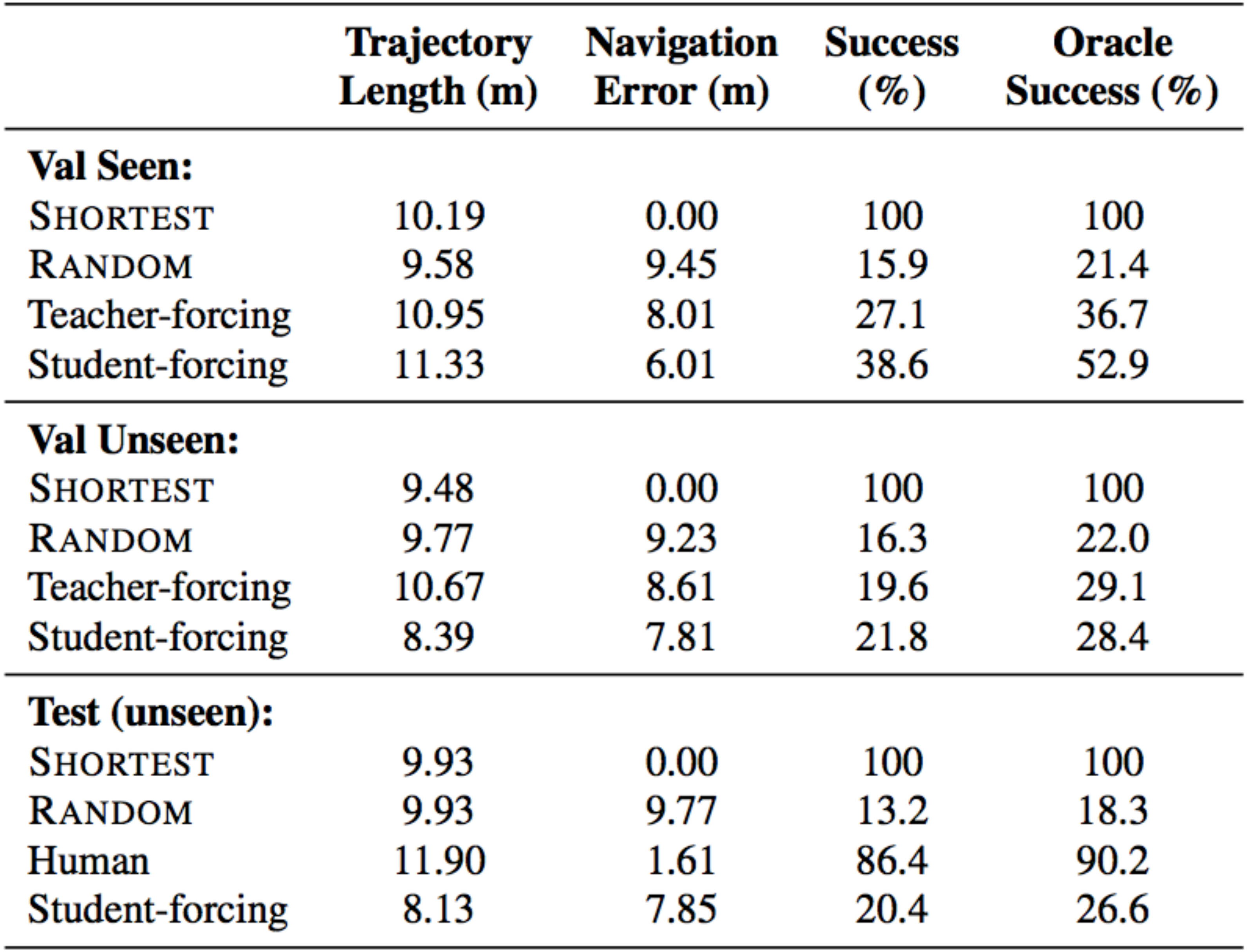
Slide credit: Stefan Lee

Vision-and-Language Navigation Evaluation

train	val-seen	val-unseen	test
61 Environments		11 Environments	18 Environments
14,025 Instructions	1020 Instructions	2349 Instructions	4173 Instructions
4675 Trajectories	340 Trajectories	783 Trajectories	1391 Trajectories

Slide credit: Stefan Lee

VLN: Results



	Trajectory Length (m)	Navigation Error (m)	Success (%)	Oracle Success (%)
Val Seen:				
SHORTEST	10.19	0.00	100	100
RANDOM	9.58	9.45	15.9	21.4
Teacher-forcing	10.95	8.01	27.1	36.7
Student-forcing	11.33	6.01	38.6	52.9
Val Unseen:				
SHORTEST	9.48	0.00	100	100
RANDOM	9.77	9.23	16.3	22.0
Teacher-forcing	10.67	8.61	19.6	29.1
Student-forcing	8.39	7.81	21.8	28.4
Test (unseen):				
SHORTEST	9.93	0.00	100	100
RANDOM	9.93	9.77	13.2	18.3
Human	11.90	1.61	86.4	90.2
Student-forcing	8.13	7.85	20.4	26.6

Slide credit: Stefan Lee

VLN: Improved navigation evaluation

Success weighted by Path Length

- Cares not only about success, but also efficiency

Binary Success

Shortest Path Length

$$\frac{1}{N}\sum_{i=1}^{N} S_i \frac{\ell_i}{\max(p_i, \ell_i)}.$$

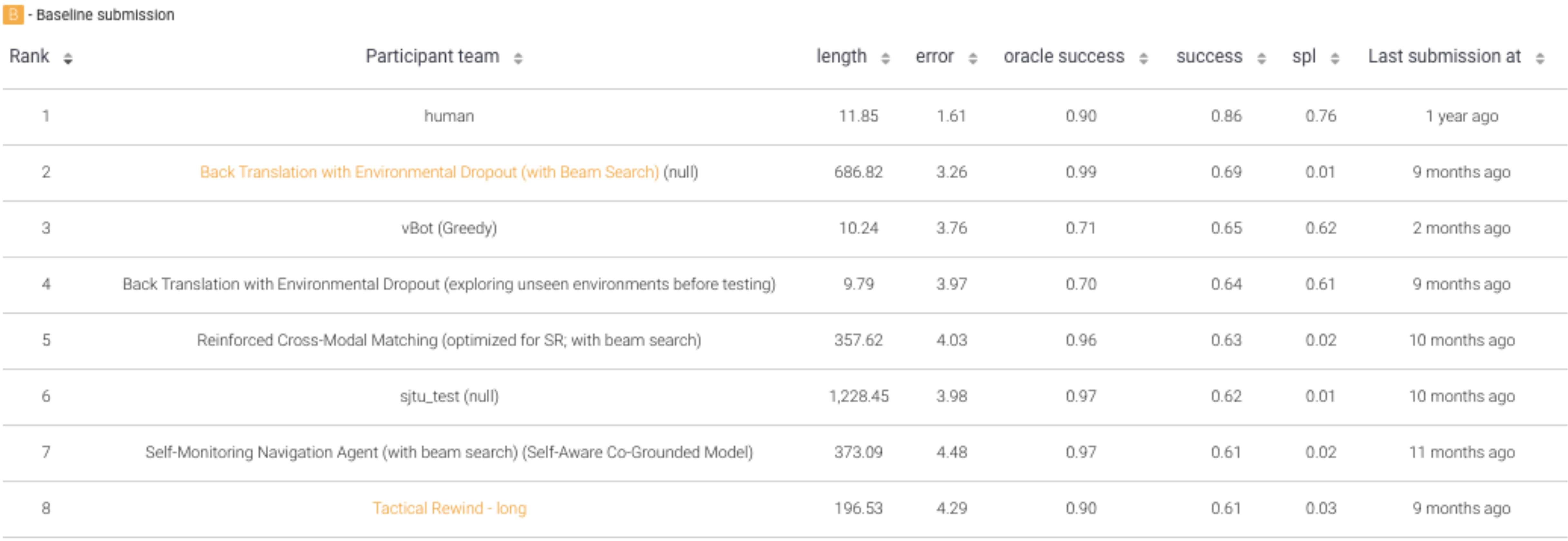
Average Over Episodes

Agent Path Length

Slide credit: Stefan Lee

Vision-and-Language Navigation Evaluation

Leaderboard hosted on EvalAI (fall 2019)



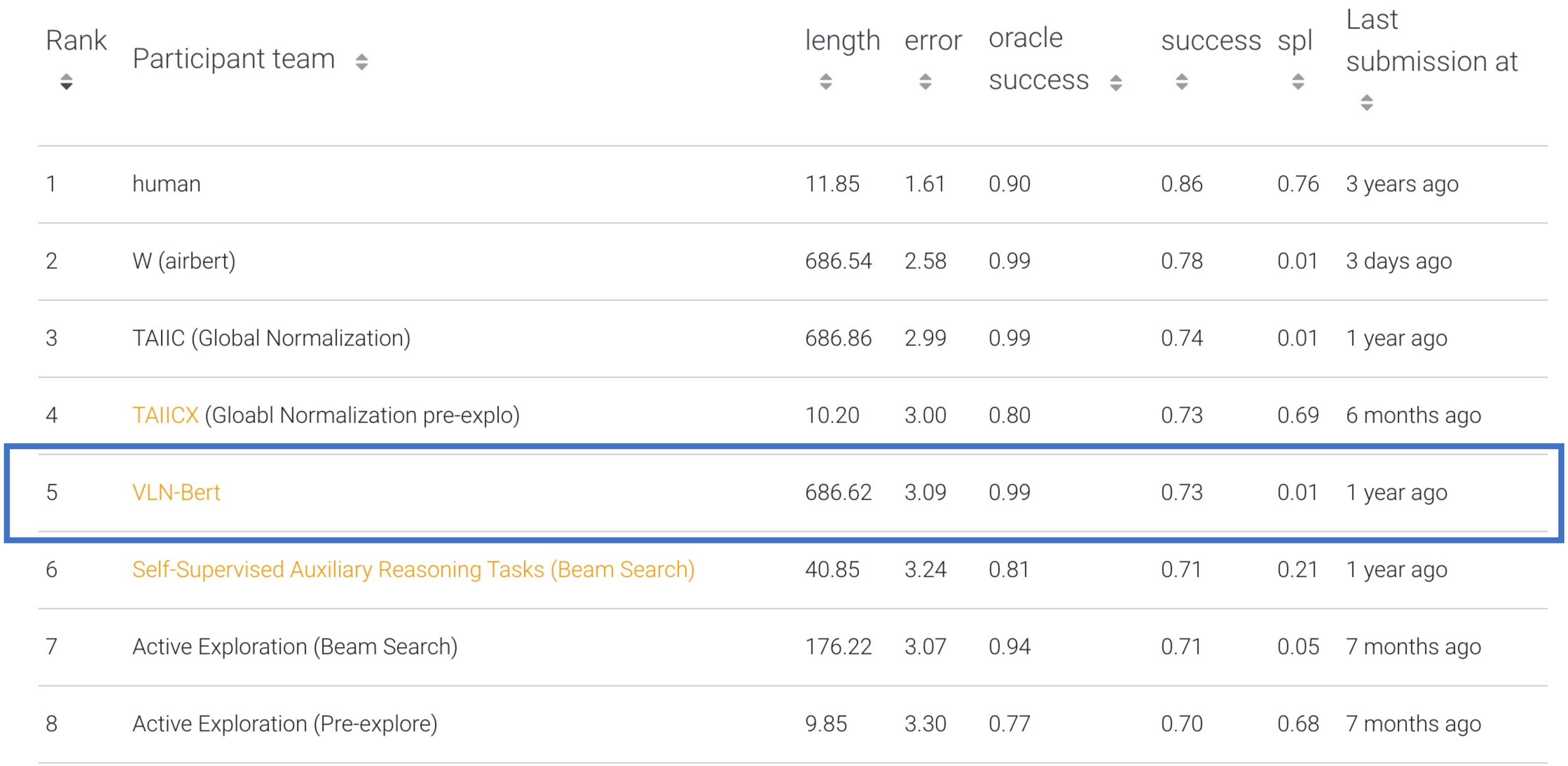
B - Baseline submission

Rank	Participant team	length	error	oracle success	success	spl	Last submission at
1	human	11.85	1.61	0.90	0.86	0.76	1 year ago
2	Back Translation with Environmental Dropout (with Beam Search) (null)	686.82	3.26	0.99	0.69	0.01	9 months ago
3	vBot (Greedy)	10.24	3.76	0.71	0.65	0.62	2 months ago
4	Back Translation with Environmental Dropout (exploring unseen environments before testing)	9.79	3.97	0.70	0.64	0.61	9 months ago
5	Reinforced Cross-Modal Matching (optimized for SR; with beam search)	357.62	4.03	0.96	0.63	0.02	10 months ago
6	sjtu_test (null)	1,228.45	3.98	0.97	0.62	0.01	10 months ago
7	Self-Monitoring Navigation Agent (with beam search) (Self-Aware Co-Grounded Model)	373.09	4.48	0.97	0.61	0.02	11 months ago
8	Tactical Rewind - long	196.53	4.29	0.90	0.61	0.03	9 months ago

Slide credit: Stefan Lee

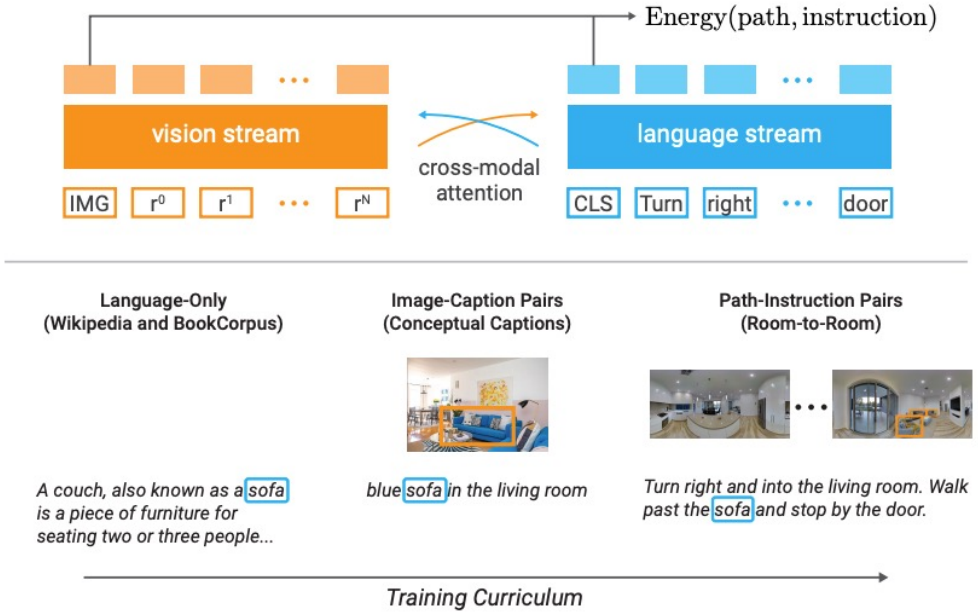
Vision-and-Language Navigation Evaluation

Leaderboard hosted on EvalAI (spring 2021)



Rank	Participant team	length	error	oracle success	success	spl	Last submission at
1	human	11.85	1.61	0.90	0.86	0.76	3 years ago
2	W (airbert)	686.54	2.58	0.99	0.78	0.01	3 days ago
3	TAIIC (Global Normalization)	686.86	2.99	0.99	0.74	0.01	1 year ago
4	TAIICX (Gloabl Normalization pre-explo)	10.20	3.00	0.80	0.73	0.69	6 months ago
5	VLN-Bert	686.62	3.09	0.99	0.73	0.01	1 year ago
6	Self-Supervised Auxiliary Reasoning Tasks (Beam Search)	40.85	3.24	0.81	0.71	0.21	1 year ago
7	Active Exploration (Beam Search)	176.22	3.07	0.94	0.71	0.05	7 months ago
8	Active Exploration (Pre-explore)	9.85	3.30	0.77	0.70	0.68	7 months ago

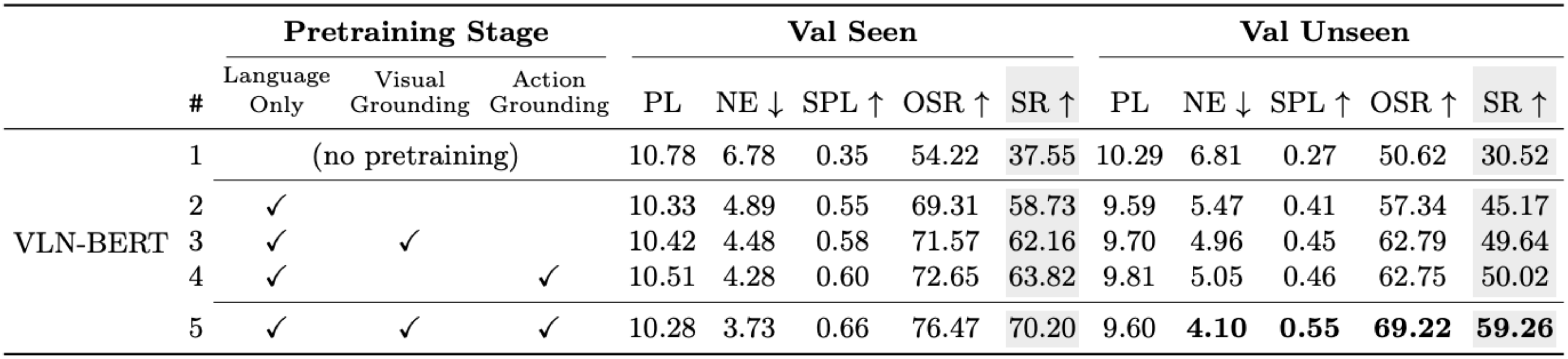
Pretraining for VLN: VLN-BERT


Energy(path, instruction)
vision stream
cross-modal attention
language stream
IMG
r⁰
r¹
...
rᴺ
CLS
Turn
right
...
door
Language-Only (Wikipedia and BookCorpus)
Image-Caption Pairs (Conceptual Captions)
Path-Instruction Pairs (Room-to-Room)
Training Curriculum

A couch, also known as a sofa is a piece of furniture for seating two or three people...

blue sofa in the living room

Turn right and into the living room. Walk past the sofa and stop by the door.

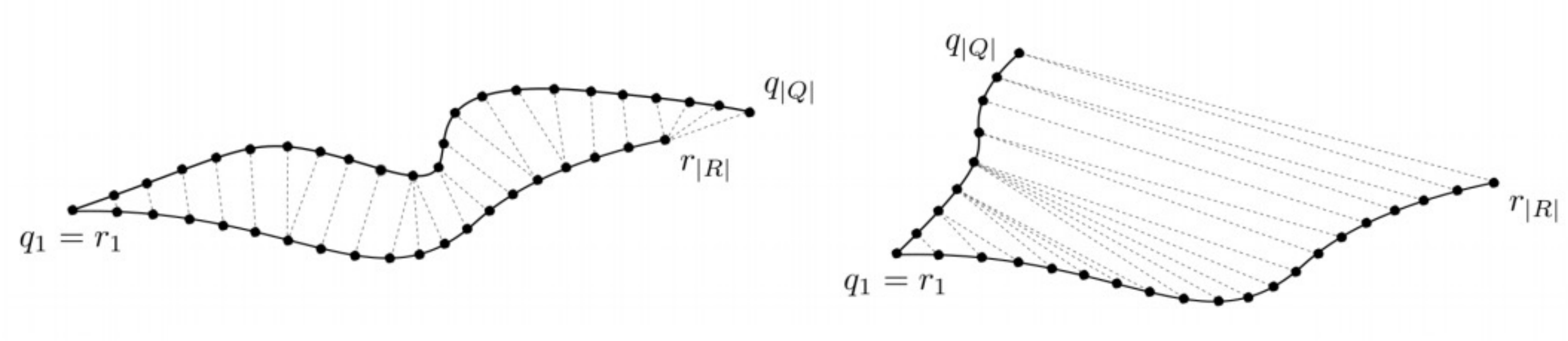


		Pretraining Stage			Val Seen					Val Unseen				
	#	Language Only	Visual Grounding	Action Grounding	PL	NE ↓	SPL ↑	OSR ↑	SR ↑	PL	NE ↓	SPL ↑	OSR ↑	SR ↑
VLN-BERT	1	(no pretraining)			10.78	6.78	0.35	54.22	37.55	10.29	6.81	0.27	50.62	30.52
	2	✓			10.33	4.89	0.55	69.31	58.73	9.59	5.47	0.41	57.34	45.17
	3	✓	✓		10.42	4.48	0.58	71.57	62.16	9.70	4.96	0.45	62.79	49.64
	4	✓		✓	10.51	4.28	0.60	72.65	63.82	9.81	5.05	0.46	62.75	50.02
	5	✓	✓	✓	10.28	3.73	0.66	76.47	70.20	9.60	**4.10**	**0.55**	**69.22**	**59.26**

Improving Vision-and-Language Navigation with Image-Text Pairs from the Web
https://arxiv.org/pdf/2004.14973.pdf
Majumdar et al, ECCV 2020

Vision-and-Language Navigation Evaluation

But… path matters when following instructions!

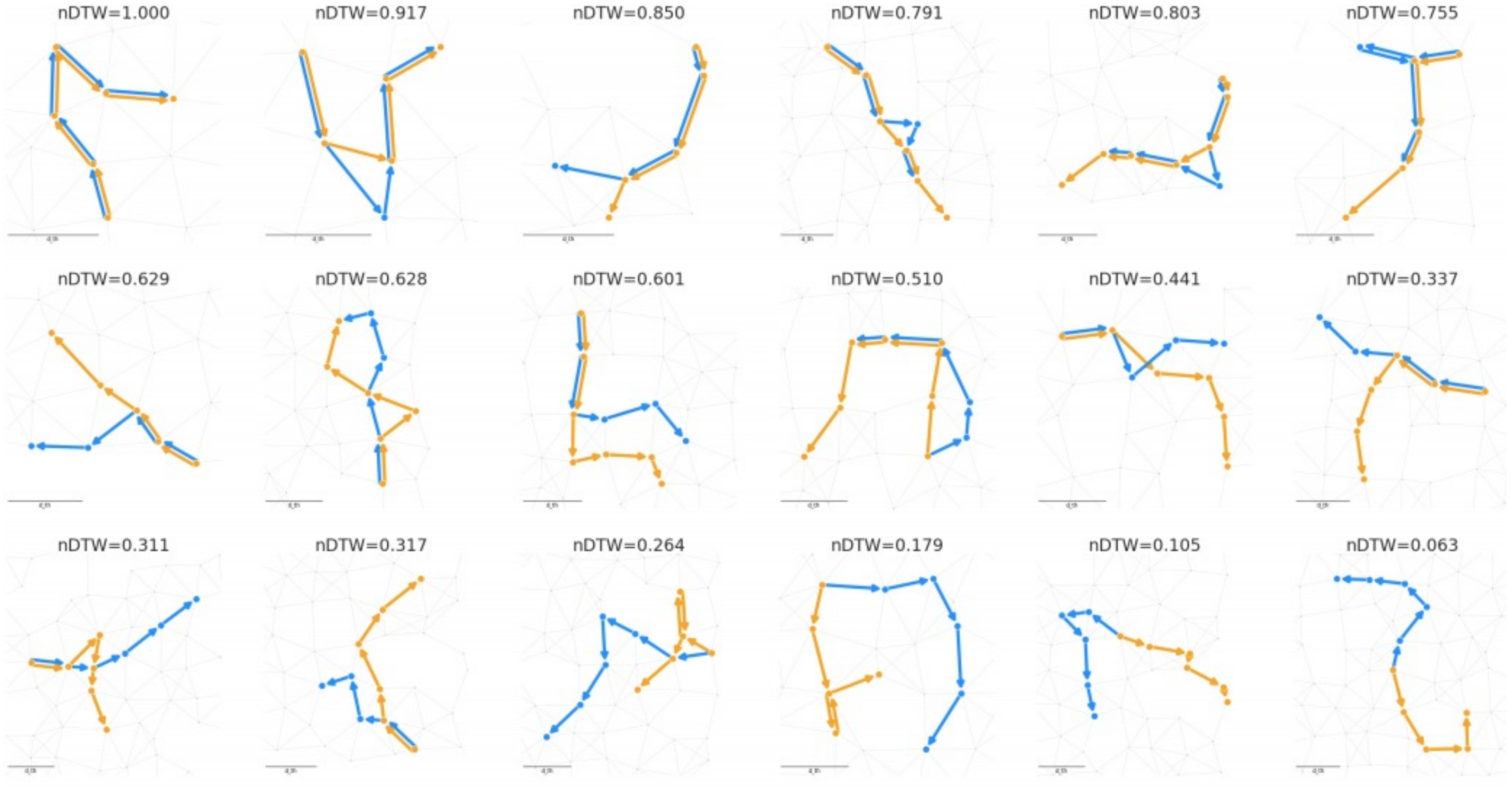


$$\mathrm{nDTW}(R,Q) = \exp\left(-\frac{\mathrm{DTW}(R,Q)}{|R| \cdot d_{th}}\right) = \exp\left(-\frac{\min_{W \in \mathcal{W}} \sum_{(i_k,j_k)\in W} d(r_{i_k}, q_{j_k})}{|R| \cdot d_{th}}\right)$$

General Evaluation for Instruction Conditioned Navigation using Dynamic Time Warping
https://arxiv.org/abs/1907.05446
Ilharco et al, NeurIPS 2019

Slide credit: Stefan Lee

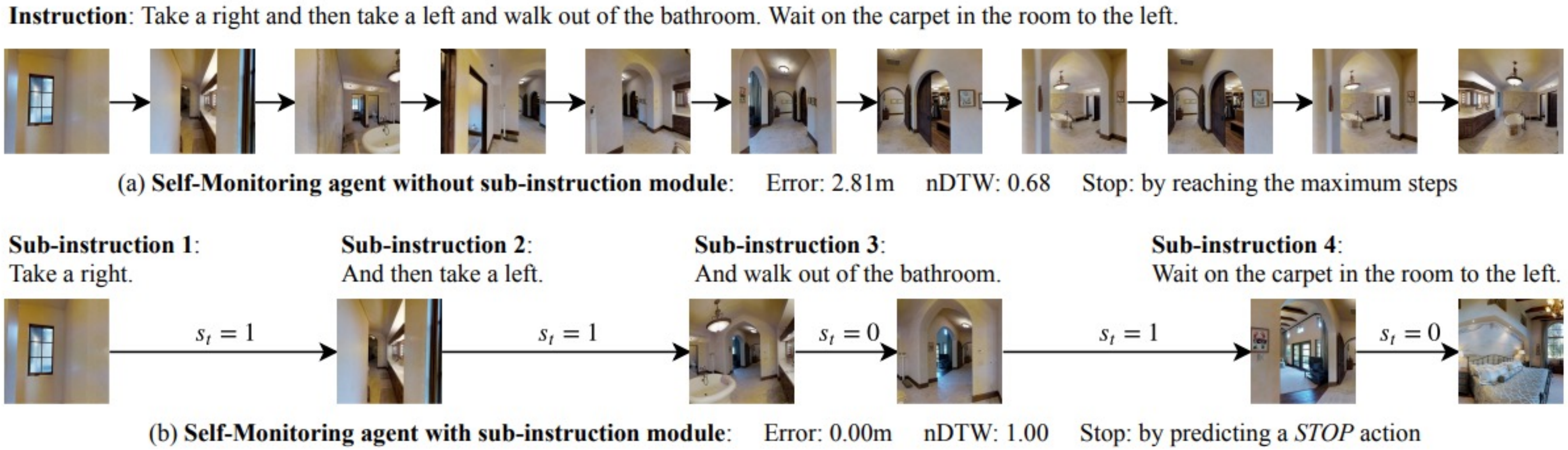
Vision-and-Language Navigation Evaluation



normalized Dynamic Time Warping (nDTW)

Slide credit: Stefan Lee

Sub-instruction aware VLN



Instruction: Take a right and then take a left and walk out of the bathroom. Wait on the carpet in the room to the left.

(a) **Self-Monitoring agent without sub-instruction module**: Error: 2.81m nDTW: 0.68 Stop: by reaching the maximum steps

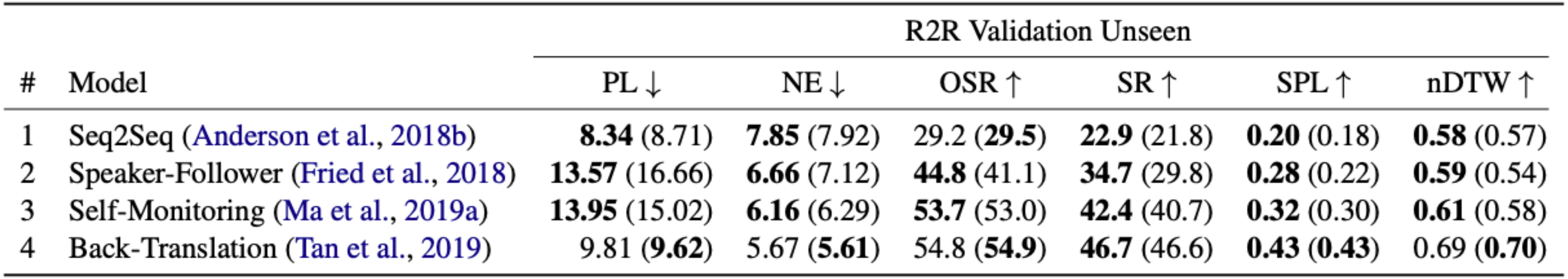
Sub-instruction 1: Take a right.

Sub-instruction 2: And then take a left.

Sub-instruction 3: And walk out of the bathroom.

Sub-instruction 4: Wait on the carpet in the room to the left.

(b) **Self-Monitoring agent with sub-instruction module**: Error: 0.00m nDTW: 1.00 Stop: by predicting a *STOP* action



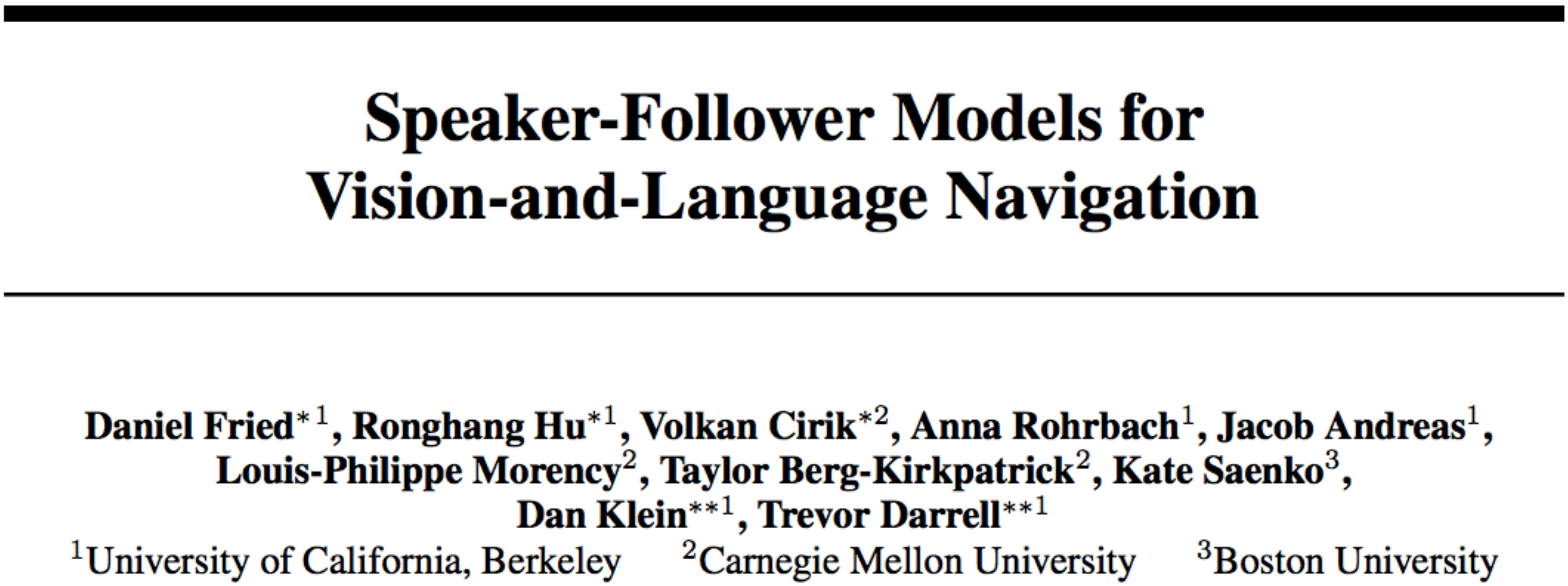
#	Model	R2R Validation Unseen					
		PL ↓	NE ↓	OSR ↑	SR ↑	SPL ↑	nDTW ↑
1	Seq2Seq (Anderson et al., 2018b)	**8.34** (8.71)	**7.85** (7.92)	29.2 (**29.5**)	**22.9** (21.8)	**0.20** (0.18)	**0.58** (0.57)
2	Speaker-Follower (Fried et al., 2018)	**13.57** (16.66)	**6.66** (7.12)	**44.8** (41.1)	**34.7** (29.8)	**0.28** (0.22)	**0.59** (0.54)
3	Self-Monitoring (Ma et al., 2019a)	**13.95** (15.02)	**6.16** (6.29)	**53.7** (53.0)	**42.4** (40.7)	**0.32** (0.30)	**0.61** (0.58)
4	Back-Translation (Tan et al., 2019)	9.81 (**9.62**)	5.67 (**5.61**)	54.8 (**54.9**)	**46.7** (46.6)	**0.43** (**0.43**)	0.69 (**0.70**)

https://arxiv.org/pdf/2004.02707.pdf
Hong et al, EMNLP 2020

With (without)
subinstructions

Vision-and-language Navigation (VLN)

Speaker-Listener Model



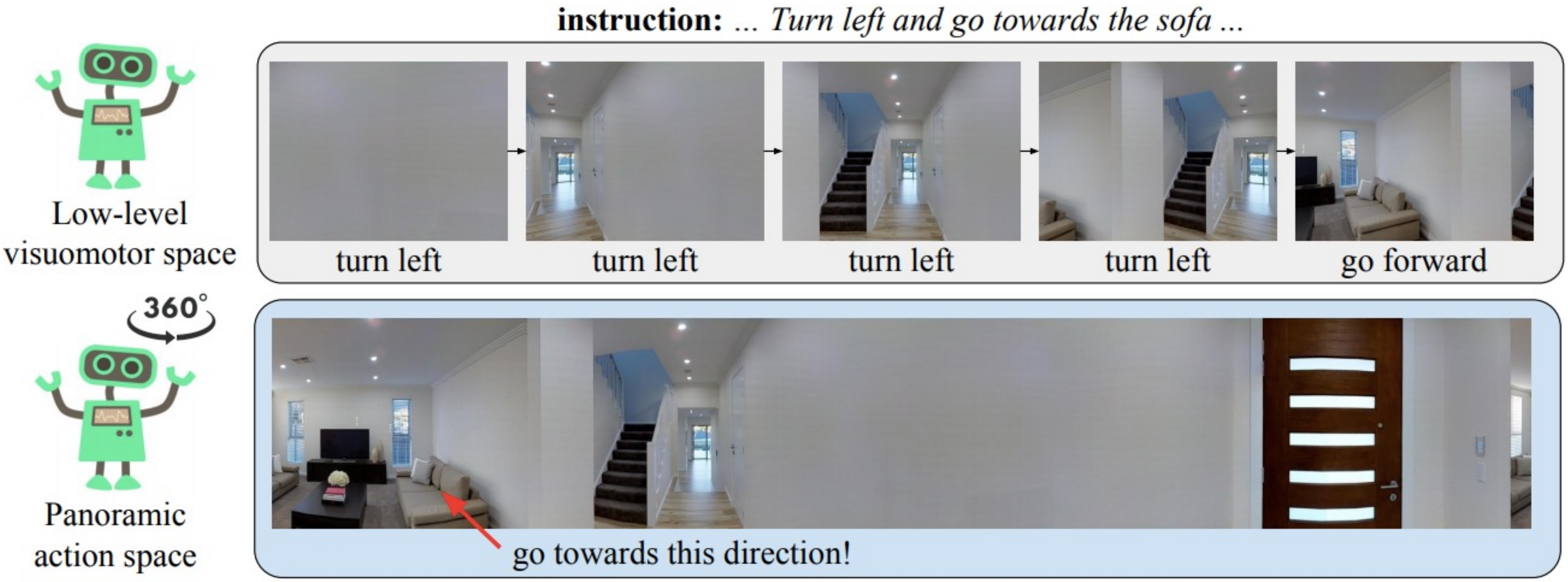
Speaker-Follower Models for Vision-and-Language Navigation

Daniel Fried*[1], Ronghang Hu*[1], Volkan Cirik*[2], Anna Rohrbach[1], Jacob Andreas[1], Louis-Philippe Morency[2], Taylor Berg-Kirkpatrick[2], Kate Saenko[3], Dan Klein[1], Trevor Darrell**[1]**

[1]University of California, Berkeley [2]Carnegie Mellon University [3]Boston University

Slide credit: Stefan Lee

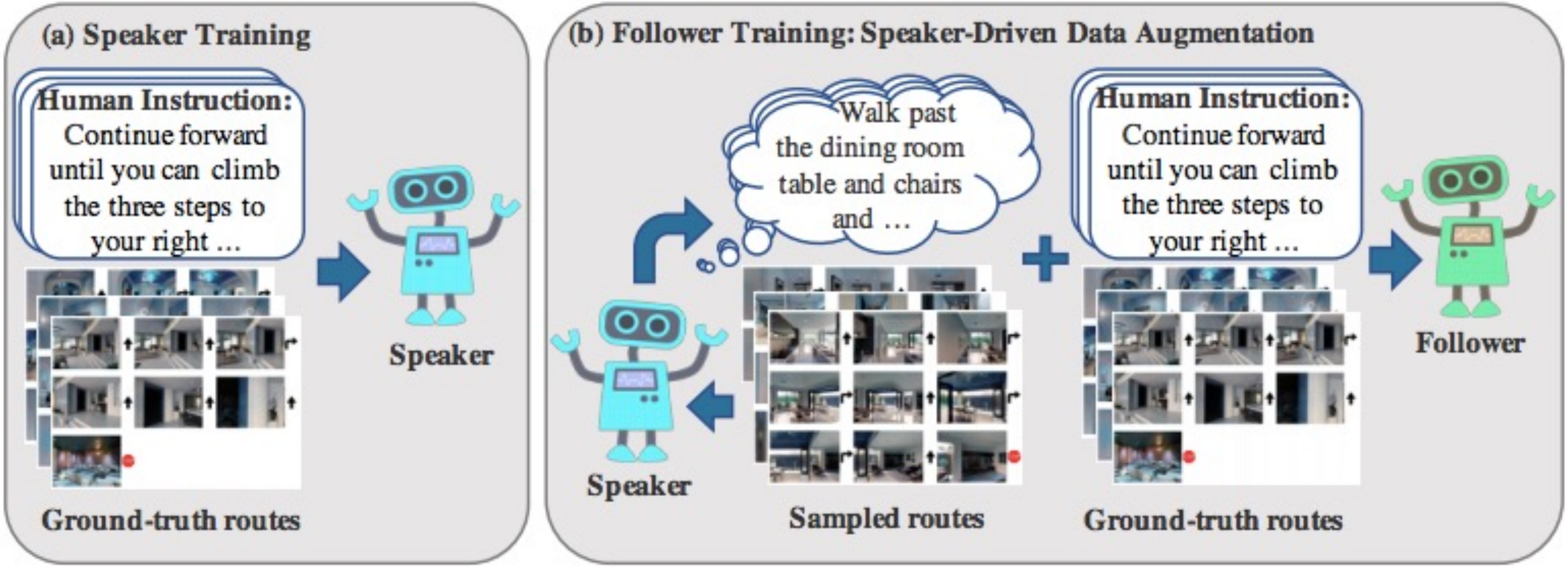
VLN: Speaker-Follower Model


instruction: ... *Turn left and go towards the sofa* ...
Low-level visuomotor space
turn left
turn left
turn left
turn left
go forward
360°
Panoramic action space
go towards this direction!

https://arxiv.org/pdf/1806.02724.pdf

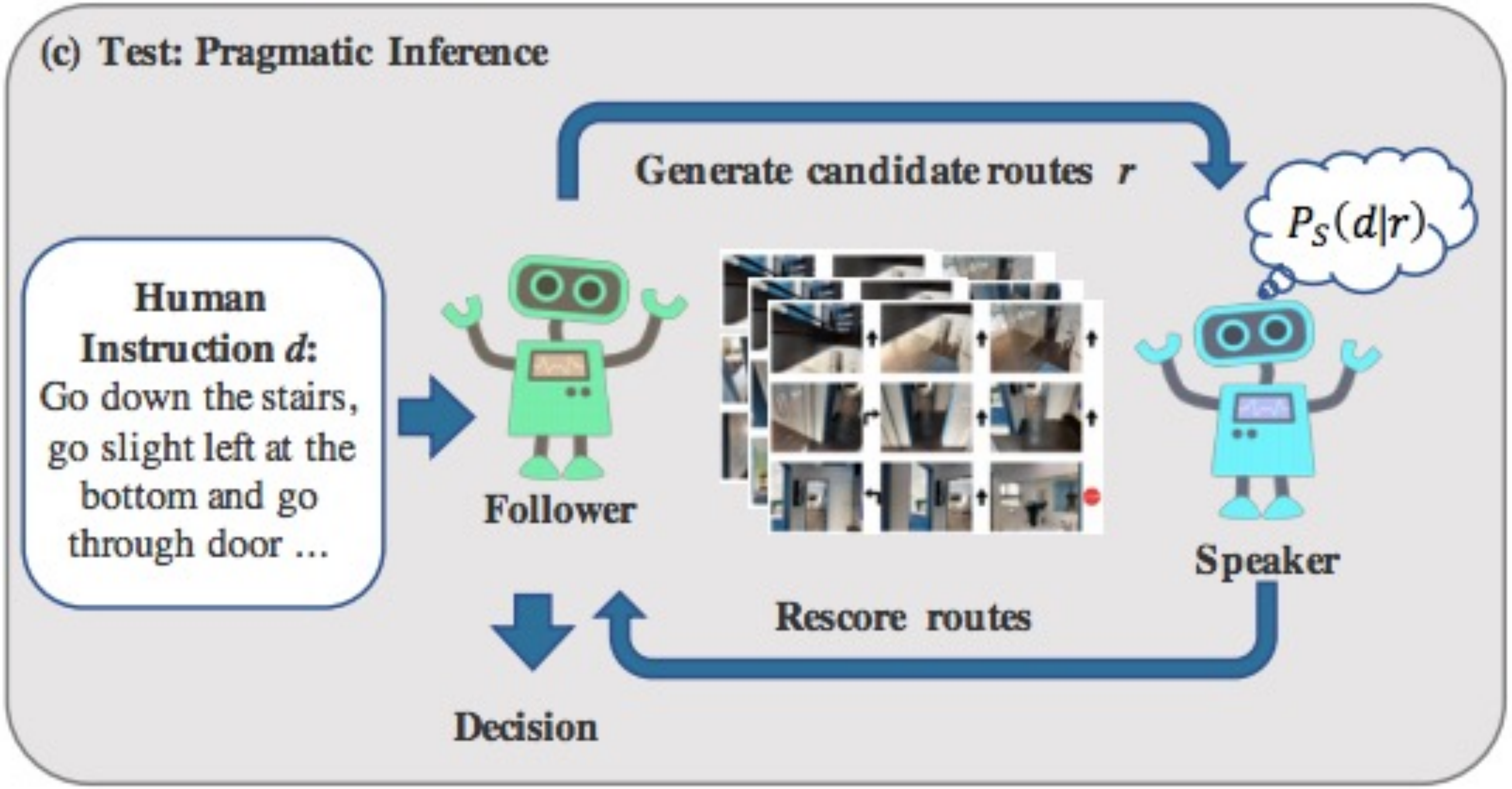
Fried et al, NeurIPS 2018

VLN: Speaker-Follower Model

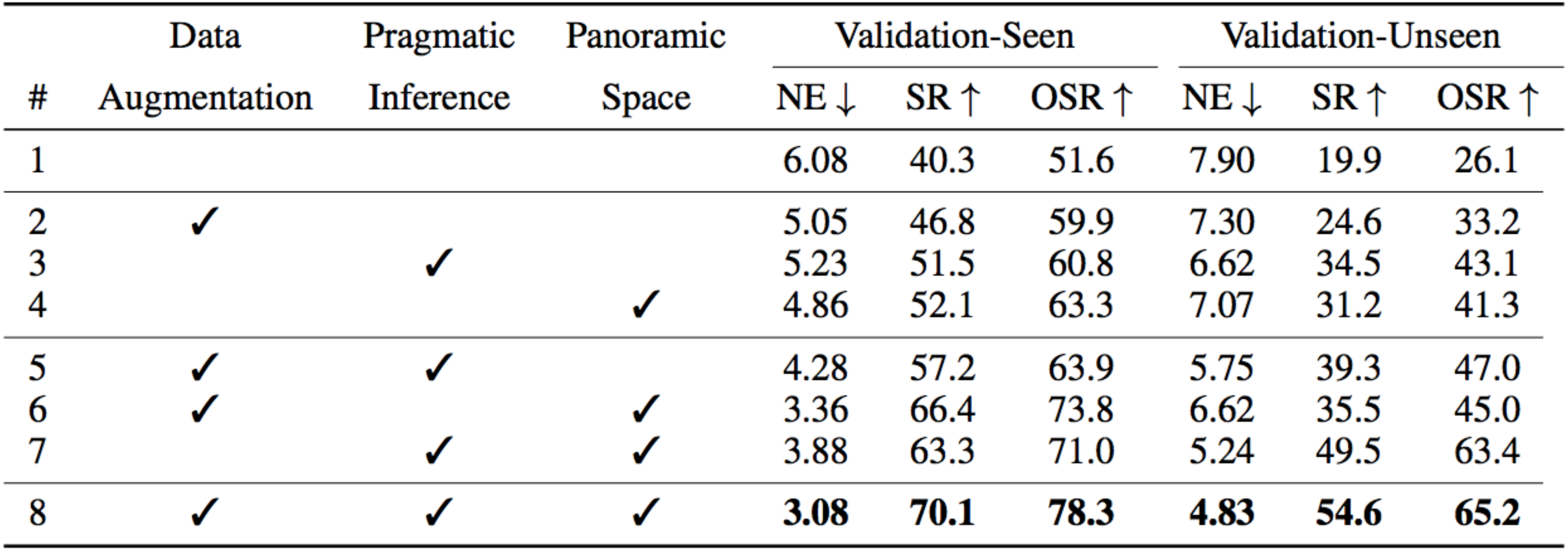

(a) Speaker Training
Human Instruction:
Continue forward until you can climb the three steps to your right ...
Speaker
Ground-truth routes
(b) Follower Training: Speaker-Driven Data Augmentation
Walk past the dining room table and chairs and ...
Speaker
Sampled routes
Human Instruction:
Continue forward until you can climb the three steps to your right ...
Ground-truth routes
Follower

https://arxiv.org/pdf/1806.02724.pdf
Fried et al, NeurIPS 2018

VLN: Speaker-Follower Model


(c) Test: Pragmatic Inference
Generate candidate routes r
$P_S(d|r)$
Human Instruction d: Go down the stairs, go slight left at the bottom and go through door ...
Follower
Speaker
Rescore routes
Decision

VLN: Speaker-Follower Model

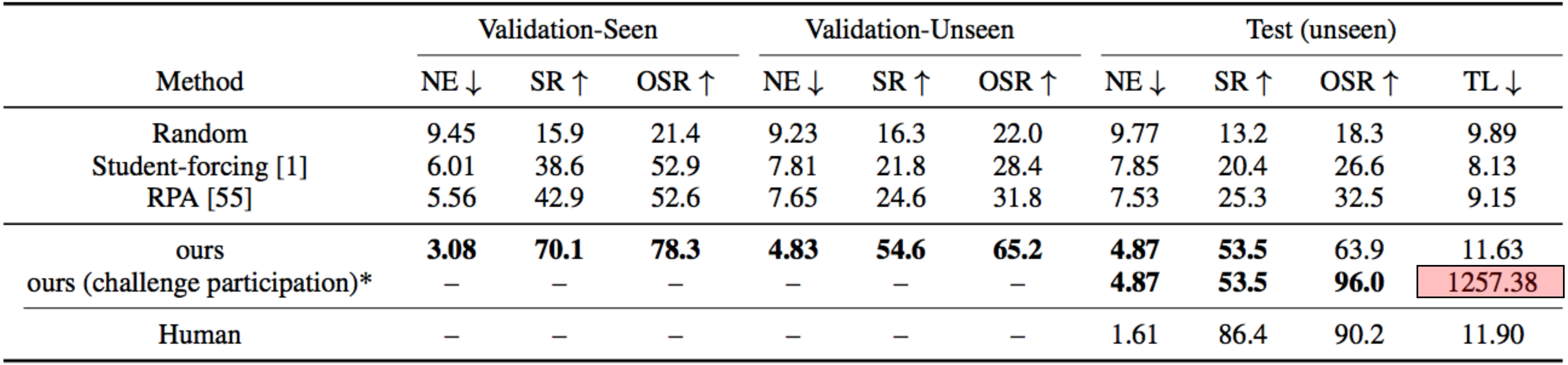


	Data	Pragmatic	Panoramic	Validation-Seen			Validation-Unseen		
#	Augmentation	Inference	Space	NE ↓	SR ↑	OSR ↑	NE ↓	SR ↑	OSR ↑
1				6.08	40.3	51.6	7.90	19.9	26.1
2	✓			5.05	46.8	59.9	7.30	24.6	33.2
3		✓		5.23	51.5	60.8	6.62	34.5	43.1
4			✓	4.86	52.1	63.3	7.07	31.2	41.3
5	✓	✓		4.28	57.2	63.9	5.75	39.3	47.0
6	✓		✓	3.36	66.4	73.8	6.62	35.5	45.0
7		✓	✓	3.88	63.3	71.0	5.24	49.5	63.4
8	✓	✓	✓	**3.08**	**70.1**	**78.3**	**4.83**	**54.6**	**65.2**

https://arxiv.org/pdf/1806.02724.pdf

Fried et al, NeurIPS 2018

VLN: Speaker-Follower Model

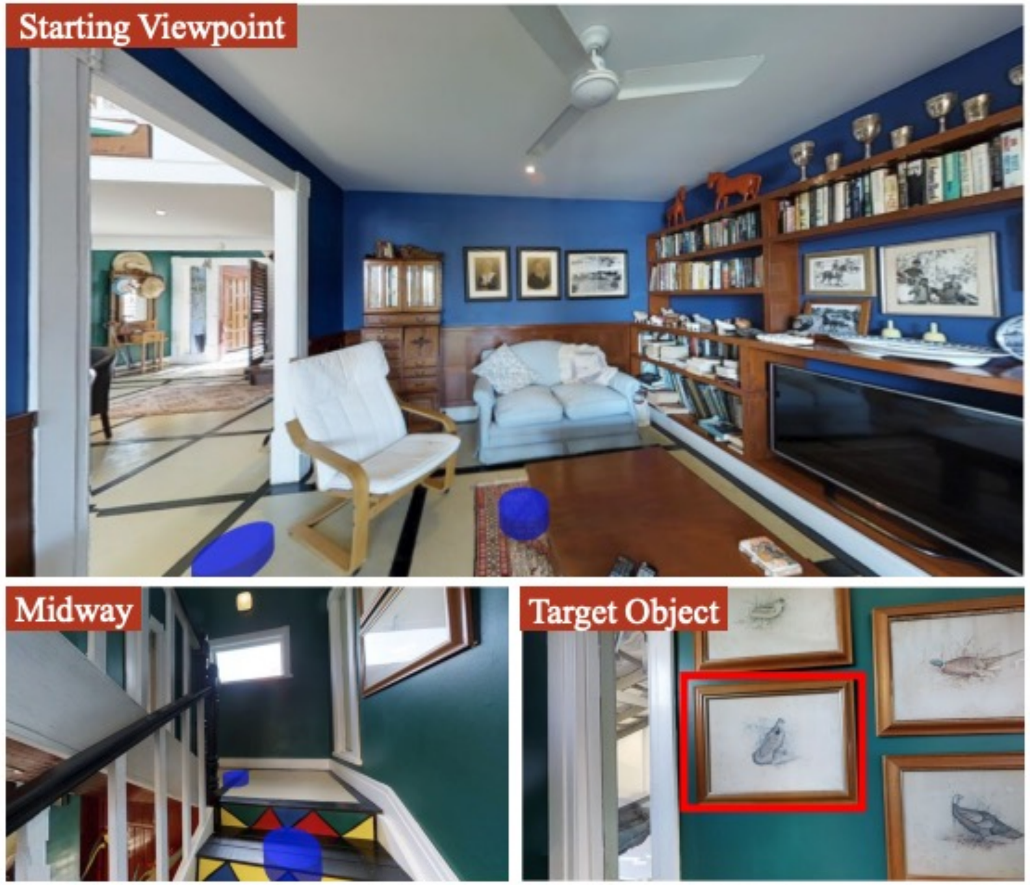


	Validation-Seen			Validation-Unseen			Test (unseen)			
Method	NE ↓	SR ↑	OSR ↑	NE ↓	SR ↑	OSR ↑	NE ↓	SR ↑	OSR ↑	TL ↓
Random	9.45	15.9	21.4	9.23	16.3	22.0	9.77	13.2	18.3	9.89
Student-forcing [1]	6.01	38.6	52.9	7.81	21.8	28.4	7.85	20.4	26.6	8.13
RPA [55]	5.56	42.9	52.6	7.65	24.6	31.8	7.53	25.3	32.5	9.15
ours	**3.08**	**70.1**	**78.3**	**4.83**	**54.6**	**65.2**	**4.87**	**53.5**	63.9	11.63
ours (challenge participation)*	–	–	–	–	–	–	**4.87**	**53.5**	**96.0**	1257.38
Human	–	–	–	–	–	–	1.61	86.4	90.2	11.90

https://arxiv.org/pdf/1806.02724.pdf

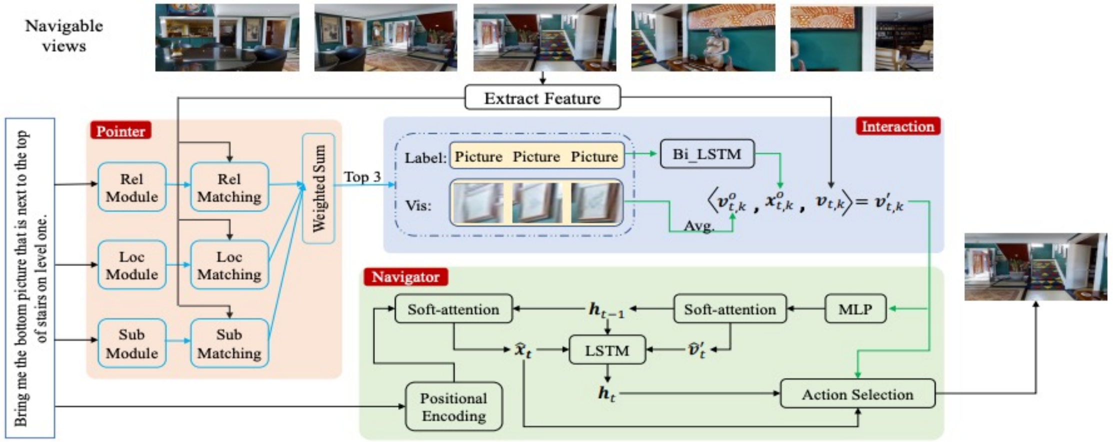
Fried et al, NeurIPS 2018

REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments



Instruction: Bring me the bottom picture that is next to the top of stairs on level one.

- Navigate + Localize
    - Follow instructions to a specified location
    - Identify the object that is being referred to
- Combines VLN + Referring Expressions

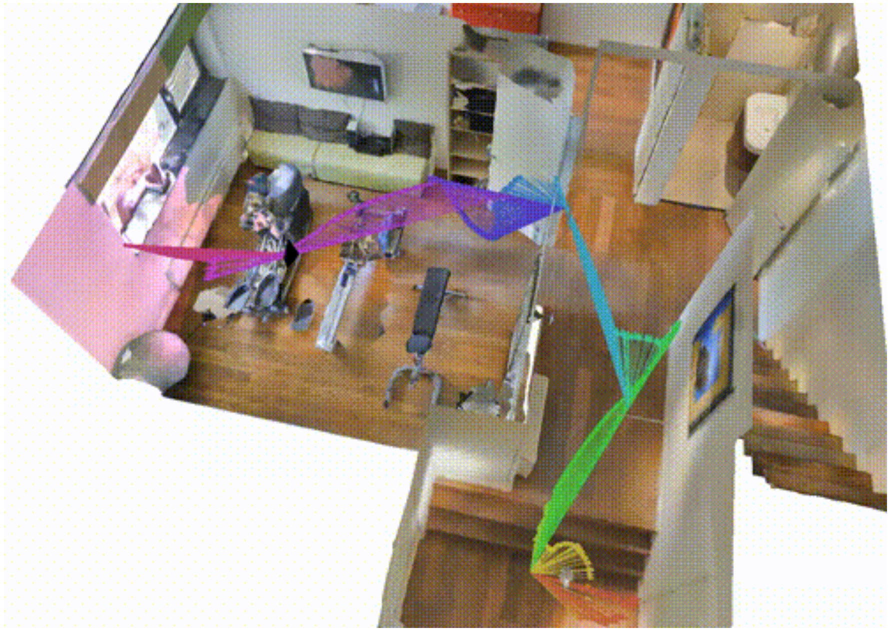


https://arxiv.org/pdf/1904.10151.pdf

Qi et al, EMNLP 2020

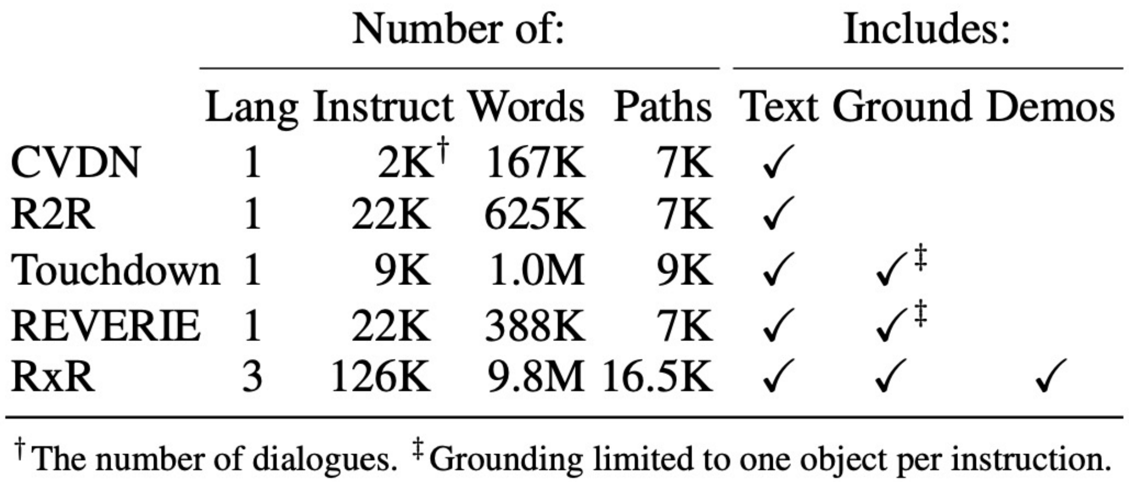
https://github.com/YuankaiQi/REVERIE

Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding



Now you are standing in-front of a closed door,
turn to your left, you can see two wooden steps,
climb the steps and walk forward by
crossing a wall painting which is to your right side,
you can see open door enter into it.
This is a gym room, move forward,
walk till the end of the room,
you can see a grey colored ball to the corner of the room,
stand there, that's your end point.

- Instructions spatially/temporally aligned to poses
- Larger, multilingual (English, Hindi, Telugu)

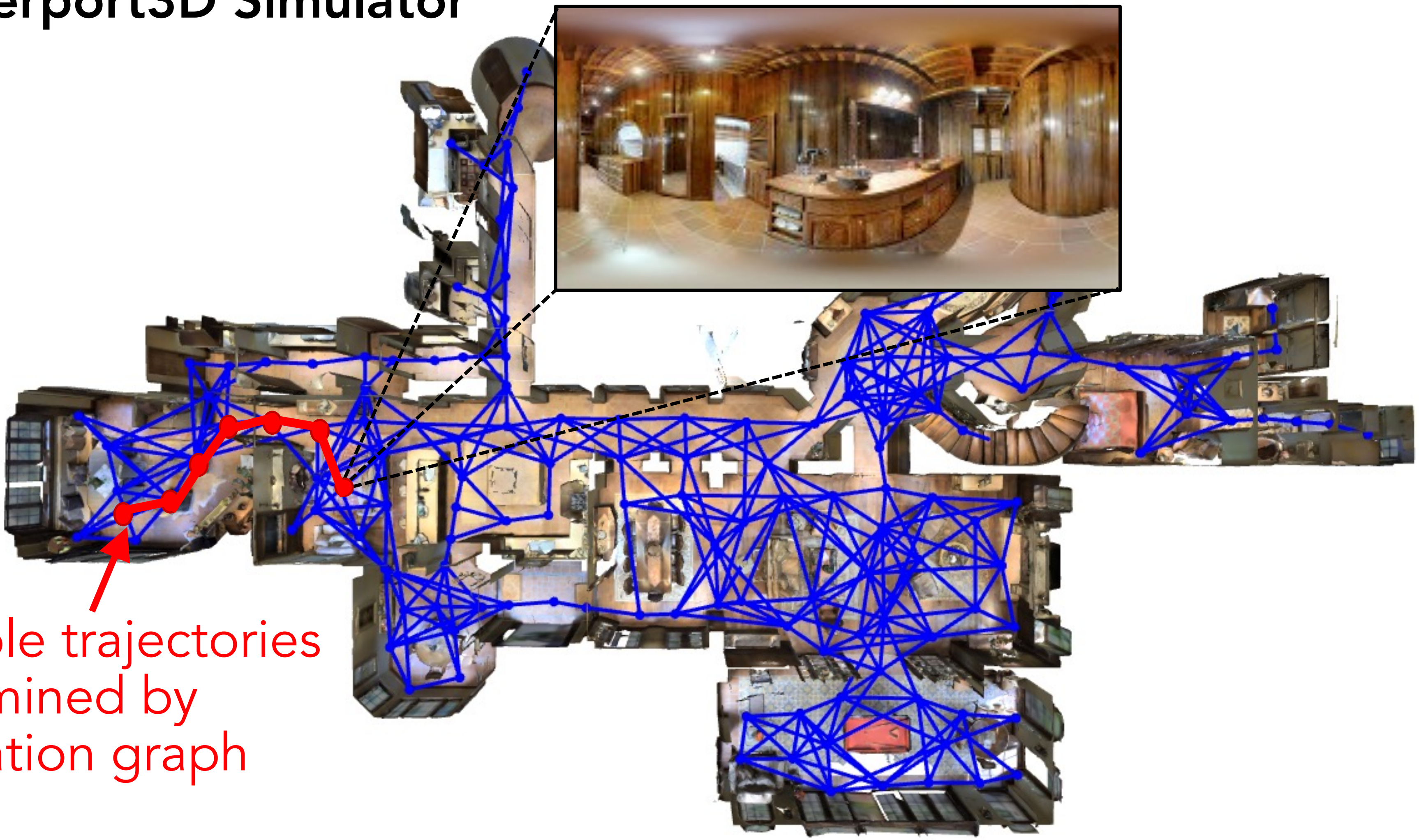


	Number of:				Includes:		
	Lang	Instruct	Words	Paths	Text	Ground	Demos
CVDN	1	2K†	167K	7K	✓		
R2R	1	22K	625K	7K	✓		
Touchdown	1	9K	1.0M	9K	✓	✓‡	
REVERIE	1	22K	388K	7K	✓	✓‡	
RxR	3	126K	9.8M	16.5K	✓	✓	✓

† The number of dialogues. ‡ Grounding limited to one object per instruction.

https://arxiv.org/pdf/2010.07954.pdf
Ku et al, EMNLP 2020
https://ai.google.com/research/rxr/

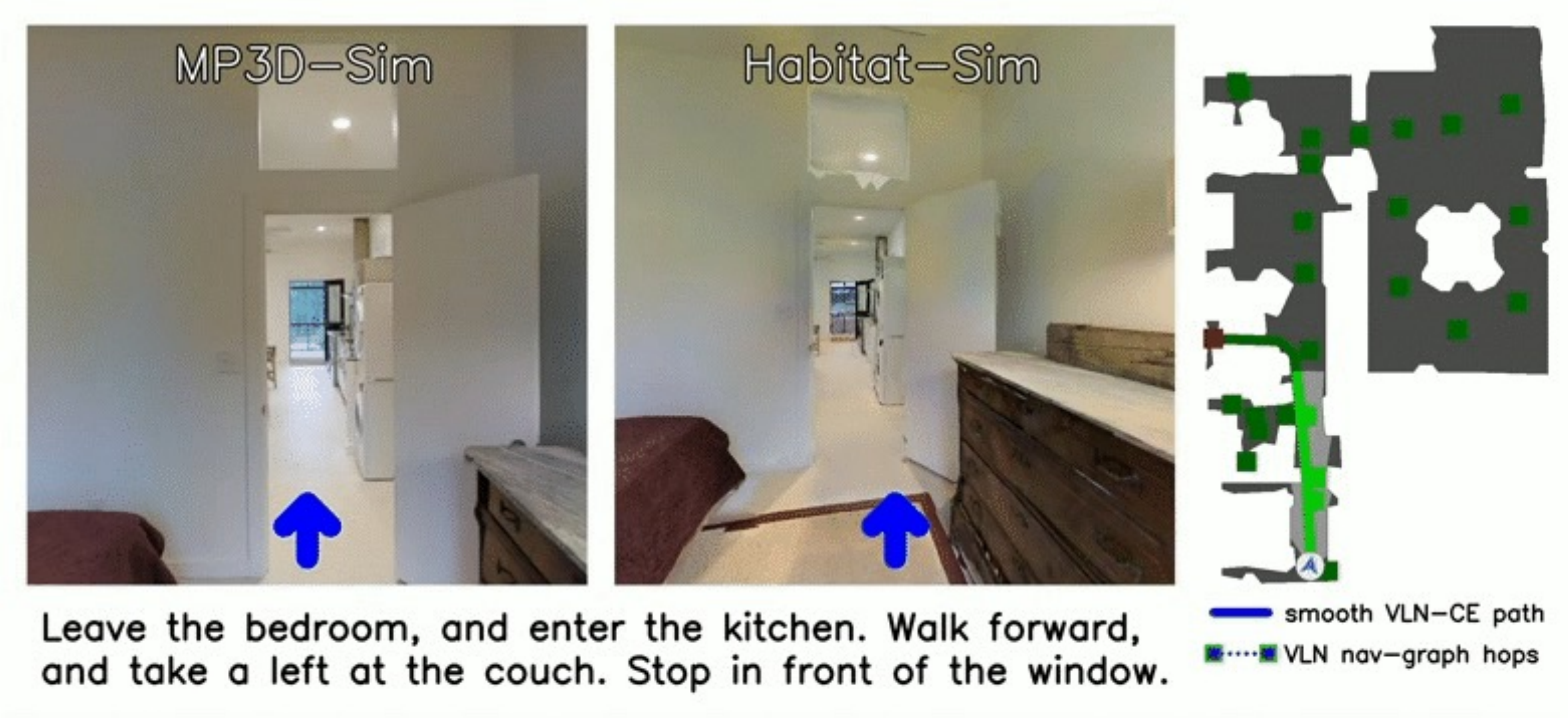
Vision-and-Language Navigation (VLN)



Feasible trajectories determined by navigation graph

Slide credit: Peter Anderson

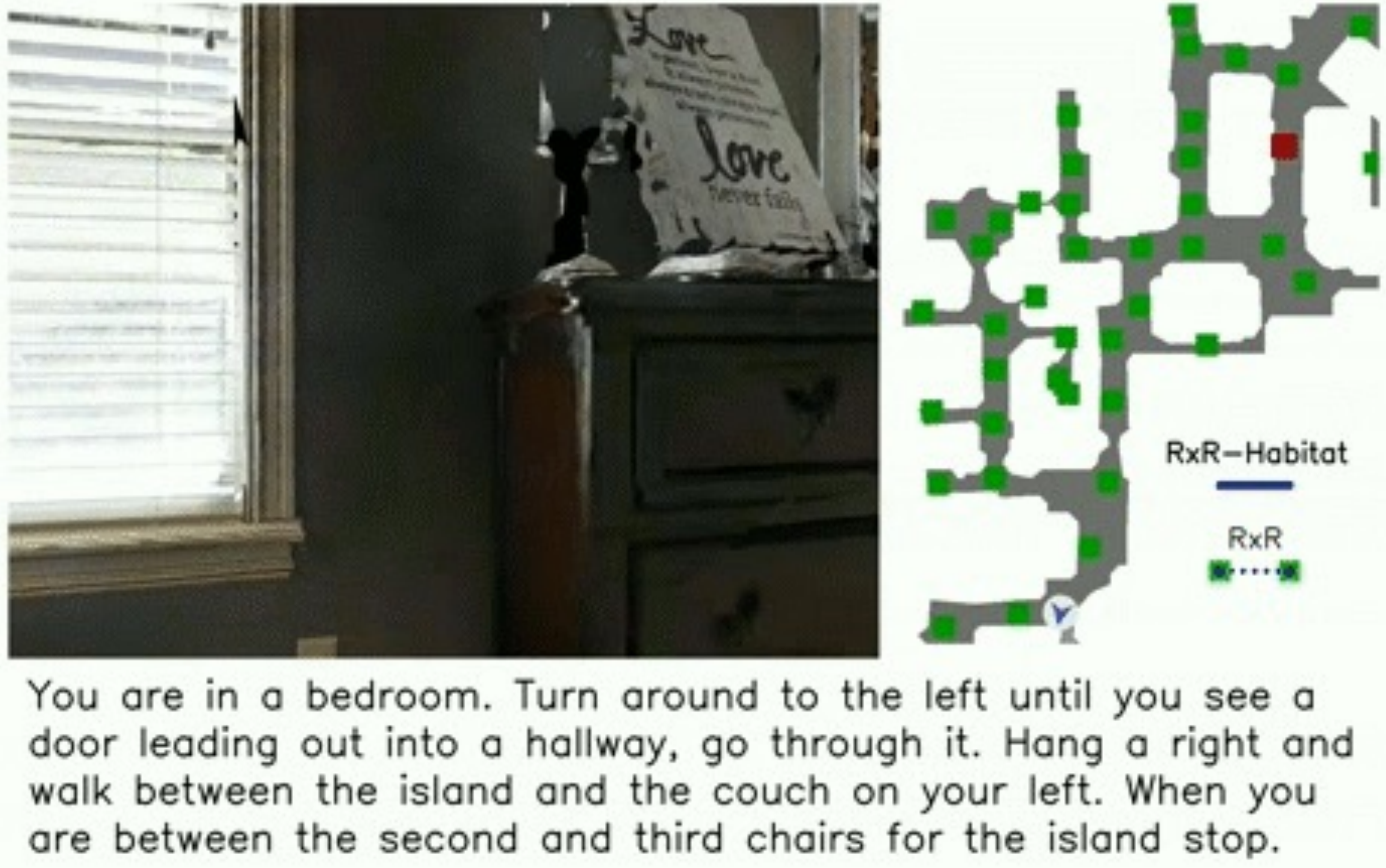
VLN with Continuous Environment



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 in Continuous Environments
https://arxiv.org/pdf/2010.07954.pdf
Krantz et al, ECCV 2020
https://jacobkrantz.github.io/vlnce/

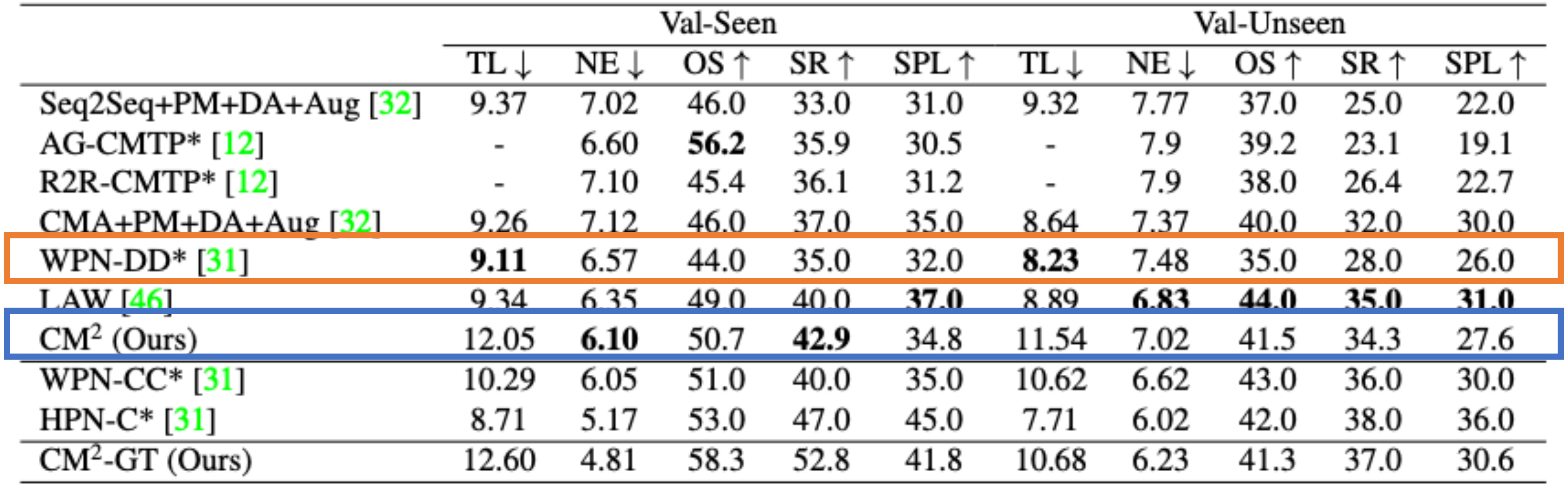
VLN with Continuous Environment



You are in a bedroom. Turn around to the left until you see a door leading out into a hallway, go through it. Hang a right and walk between the island and the couch on your left. When you are between the second and third chairs for the island stop.

https://ai.google.com/research/rxr/habitat

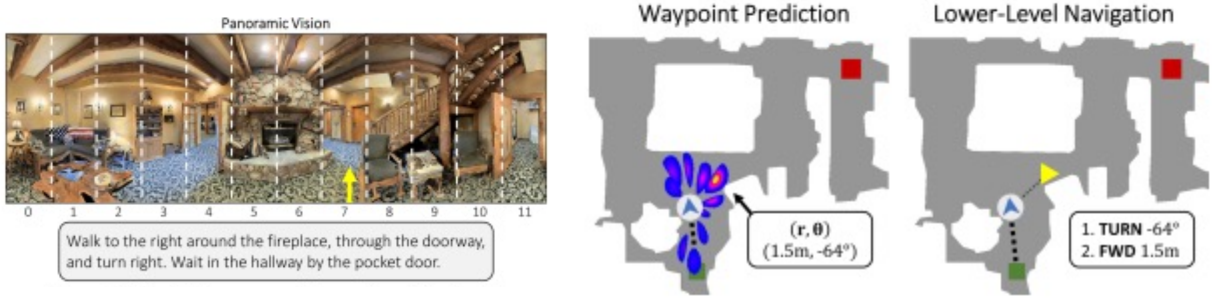
VLE-CE methods and results



	Val-Seen					Val-Unseen				
	TL ↓	NE ↓	OS ↑	SR ↑	SPL ↑	TL ↓	NE ↓	OS ↑	SR ↑	SPL ↑
Seq2Seq+PM+DA+Aug [32]	9.37	7.02	46.0	33.0	31.0	9.32	7.77	37.0	25.0	22.0
AG-CMTP* [12]	-	6.60	**56.2**	35.9	30.5	-	7.9	39.2	23.1	19.1
R2R-CMTP* [12]	-	7.10	45.4	36.1	31.2	-	7.9	38.0	26.4	22.7
CMA+PM+DA+Aug [32]	9.26	7.12	46.0	37.0	35.0	8.64	7.37	40.0	32.0	30.0
WPN-DD* [31]	**9.11**	6.57	44.0	35.0	32.0	**8.23**	7.48	35.0	28.0	26.0
LAW [46]	9.34	6.35	49.0	40.0	**37.0**	8.89	**6.83**	**44.0**	**35.0**	**31.0**
CM^2 (Ours)	12.05	**6.10**	50.7	**42.9**	34.8	11.54	7.02	41.5	34.3	27.6
WPN-CC* [31]	10.29	6.05	51.0	40.0	35.0	10.62	6.62	43.0	36.0	30.0
HPN-C* [31]	8.71	5.17	53.0	47.0	45.0	7.71	6.02	42.0	38.0	36.0
CM^2-GT (Ours)	12.60	4.81	58.3	52.8	41.8	10.68	6.23	41.3	37.0	30.6

* uses panorama, WPN-CC and HPN-C used enhanced action space

With waypoints

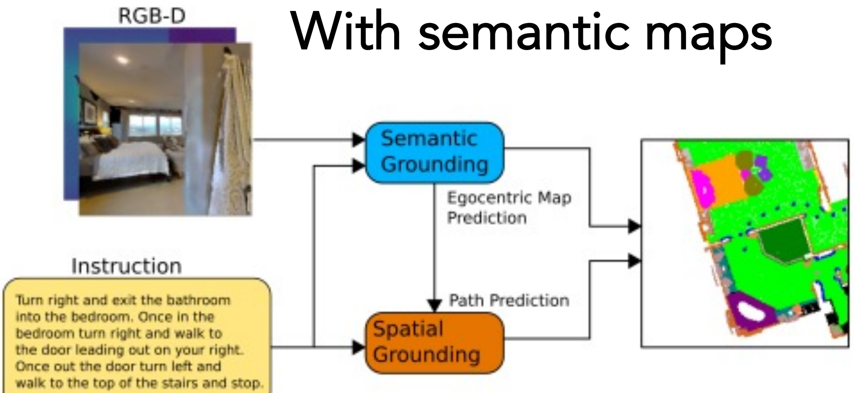


Walk to the right around the fireplace, through the doorway, and turn right. Wait in the hallway by the pocket door.

Krantz et al, ICCV 2021

https://arxiv.org/pdf/2110.02207.pdf

With semantic maps



Turn right and exit the bathroom into the bedroom. Once in the bedroom turn right and walk to the door leading out on your right. Once out the door turn left and walk to the top of the stairs and stop.

Georgakis et al, 2022

https://arxiv.org/pdf/2203.05137.pdf

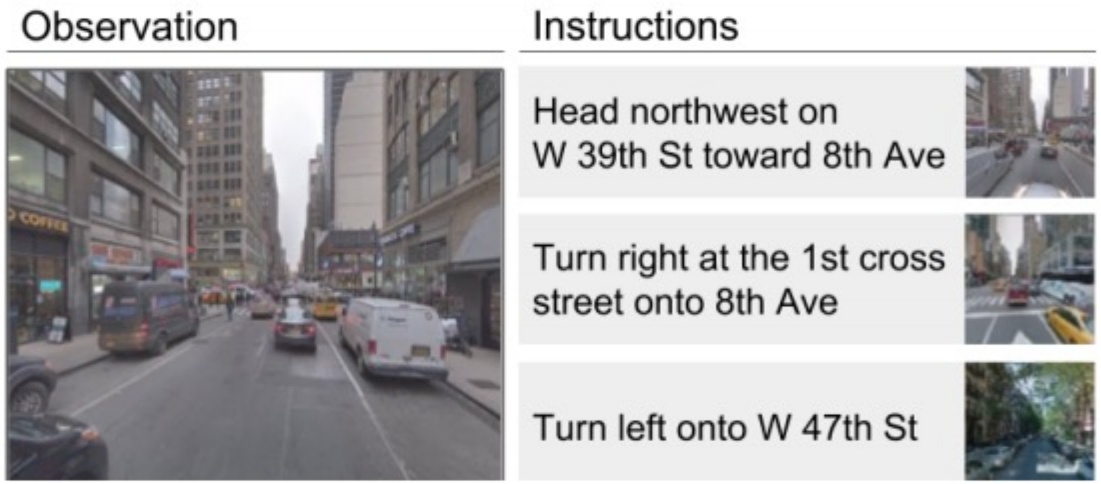
Vision-and-language Navigation (VLN)

Other Environments

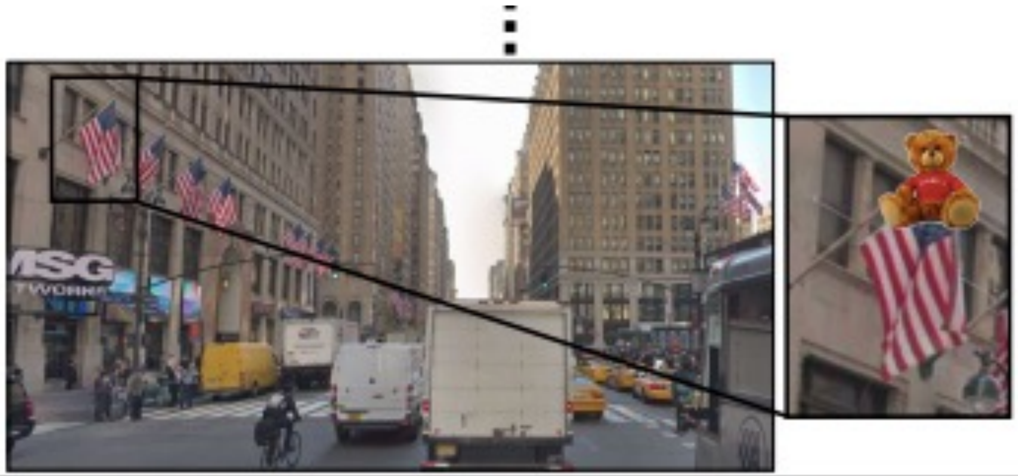
Instruction-guided Visual Navigation: StreetLearn

StreetLearn

- **Google Street View + Google Maps directions**
- The StreetLearn Environment and Dataset arxiv.org/abs/1903.01292
- Learning To Follow Directions in Street View arxiv.org/abs/1903.00401
- **Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments** arxiv.org/abs/1811.12354



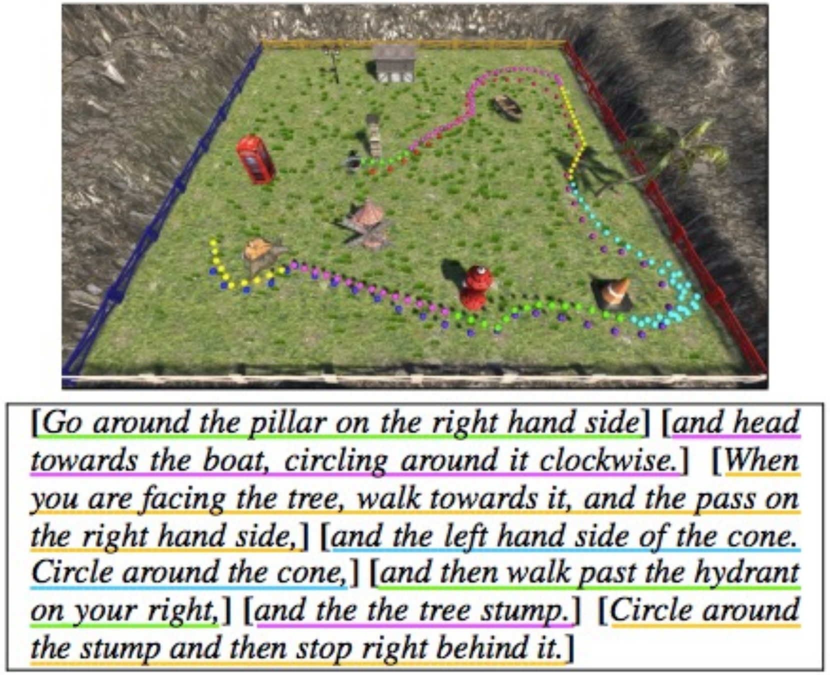
Observation	Instructions
	Head northwest on W 39th St toward 8th Ave
	Turn right at the 1st cross street onto 8th Ave
	Turn left onto W 47th St



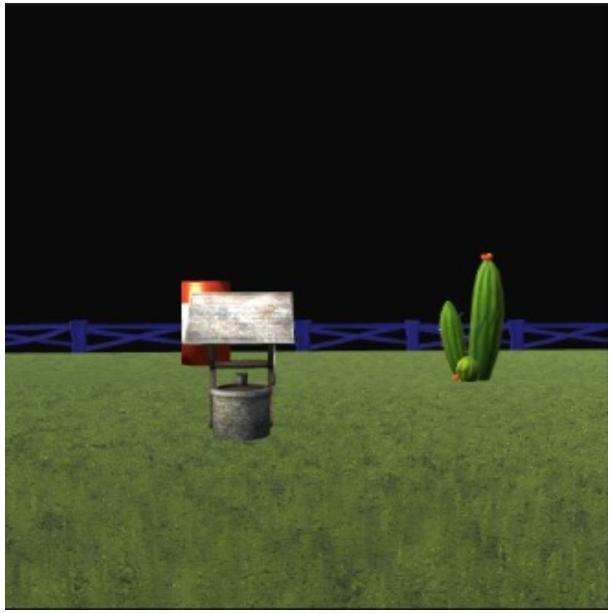
Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

Slide credit: Stefan Lee

Instruction-guided Visual Navigation



[Go around the pillar on the right hand side] [and head towards the boat, circling around it clockwise.] [When you are facing the tree, walk towards it, and the pass on the right hand side,] [and the left hand side of the cone. Circle around the cone,] [and then walk past the hydrant on your right,] [and the the tree stump.] [Circle around the stump and then stop right behind it.]



LANI

- **Simulated quadcopter in an open environment with landmark objects**
- Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction https://arxiv.org/abs/1809.00786
- Mapping Navigation Instructions to Continuous Control Actions with Position-Visitation Prediction arxiv.org/abs/1811.04179

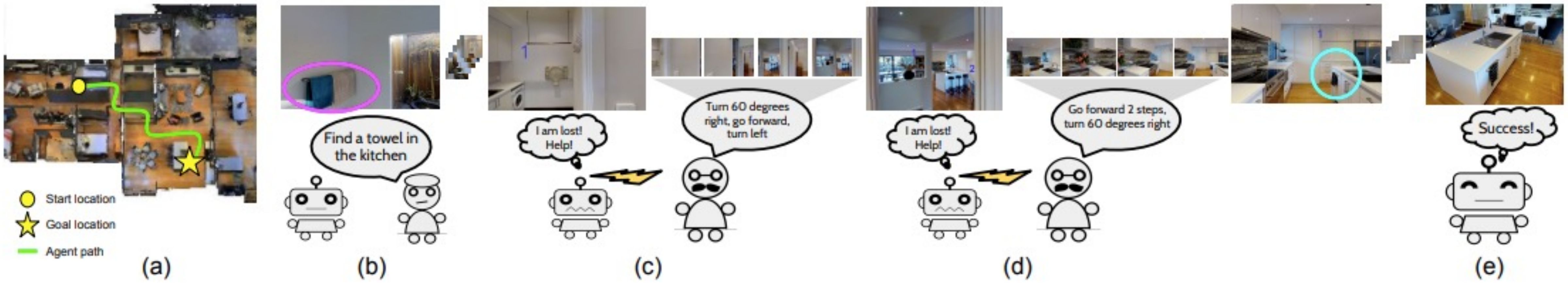
Slide credit: Stefan Lee

Task

Dialog-guided
Visual Navigation

Instruction-guided Visual Navigation

Agent can ask for directions or for help during the navigation.

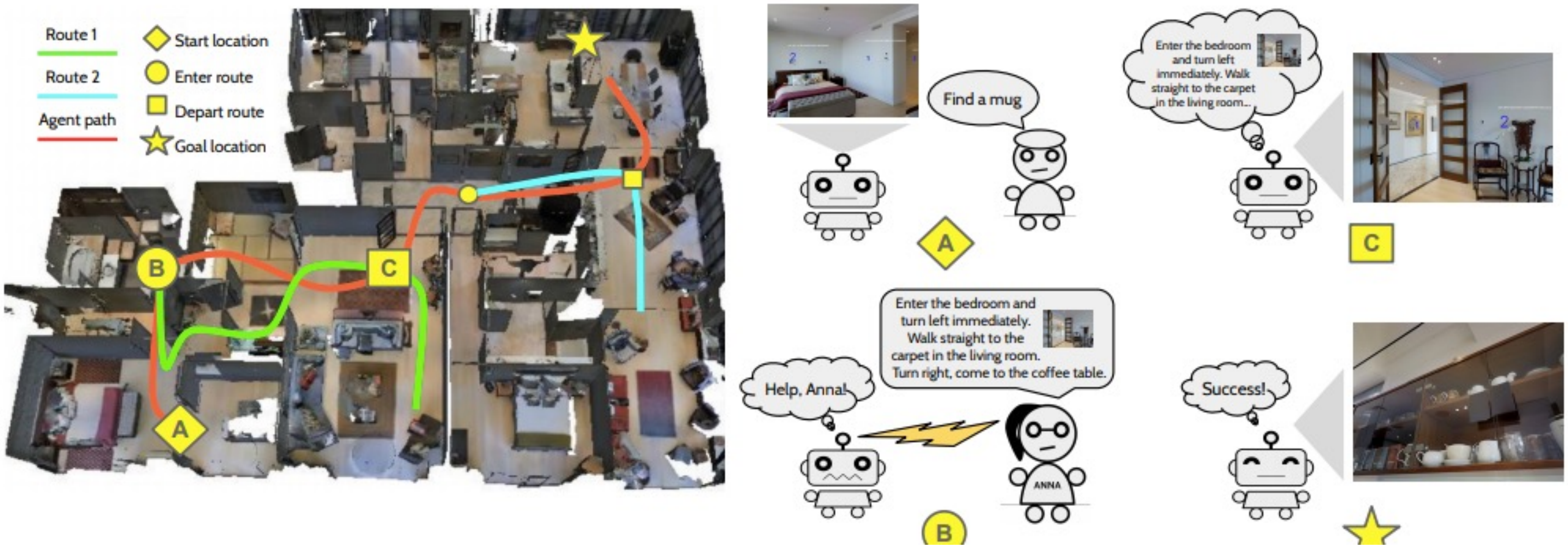

Start location
Goal location
Agent path
(a)
Find a towel in the kitchen
(b)
I am lost! Help!
Turn 60 degrees right, go forward, turn left
(c)
I am lost! Help!
Go forward 2 steps, turn 60 degrees right
(d)
Success!
(e)

- Vision-based Navigation with Language-based Assistance via Imitation Learning with Indirect Intervention arxiv.org/abs/1812.04155

Slide credit: Stefan Lee

Instruction-guided Visual Navigation

Agent can ask for directions or for help during the navigation.

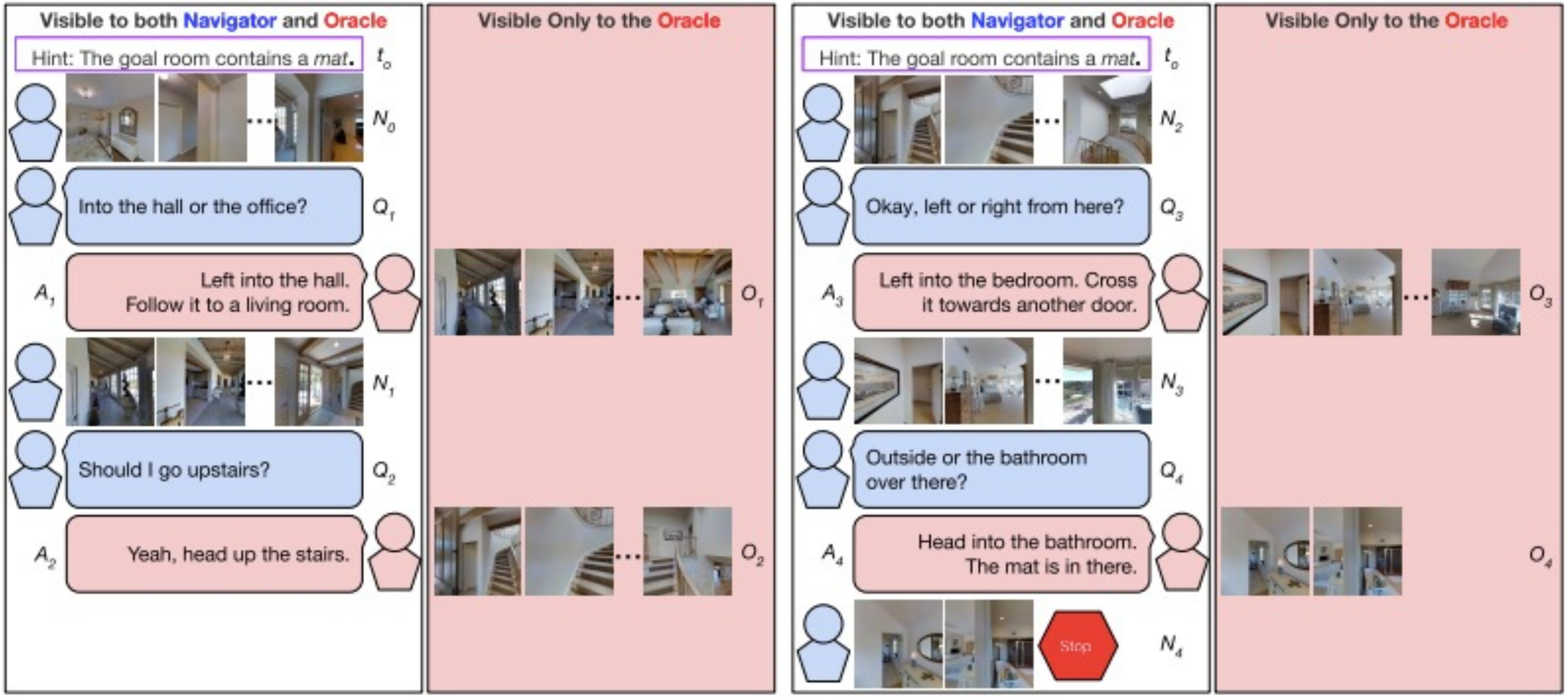

Route 1
Route 2
Agent path
Start location
Enter route
Depart route
Goal location
A
B
C
Find a mug
Enter the bedroom and turn left immediately. Walk straight to the carpet in the living room...
Help, Anna!
Enter the bedroom and turn left immediately. Walk straight to the carpet in the living room. Turn right, come to the coffee table.
ANNA
Success!

- Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning arxiv.org/abs/1909.01871

Slide credit: Stefan Lee

Instruction-guided Visual Navigation

Agent can ask for directions or for help during the navigation.


Visible to both Navigator and Oracle
Hint: The goal room contains a *mat*.
t_0
N_0
Into the hall or the office?
Q_1
A_1
Left into the hall. Follow it to a living room.
N_1
Should I go upstairs?
Q_2
A_2
Yeah, head up the stairs.
Visible Only to the Oracle
O_1
O_2
Visible to both Navigator and Oracle
Hint: The goal room contains a *mat*.
t_0
N_2
Okay, left or right from here?
Q_3
A_3
Left into the bedroom. Cross it towards another door.
N_3
Outside or the bathroom over there?
Q_4
A_4
Head into the bathroom. The mat is in there.
Stop
N_4
Visible Only to the Oracle
O_3
O_4

- Vision-and-Dialog Navigation arxiv.org/abs/1907.04957

Slide credit: Stefan Lee

Next time

- Paper presentations (3/21)
  - Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding (Yanshu)
  - REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments (Shichong)

- Wednesday (3/23): Instruction following – Rearrangement