Memory-Maze: Benchmark and Visual Language Navigation Model for Guiding Blind People

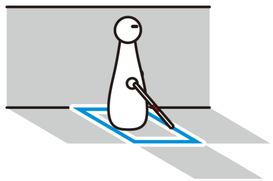
Masaki Kuribayashi*[1], Kohei Uehara *[2], Allan Wang [2,3], Daisuke Sato [3], Simon Chu [3], Shigeo Morishima [1] (* - Equal contribution)
[1] Waseda University, [2] Miraikan, [3] Carnegie Mellon University

VLN Gap for Blind Navigation

Environment Difference
Need to use robots in public environments.

Static House in R2R dataset[2]

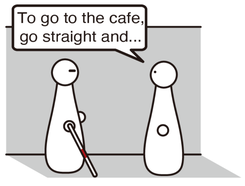
Maze-like Environment
with Intersections
e.g., shopping malls



Instruction Difference
Actual instructions are obtained from human memory.

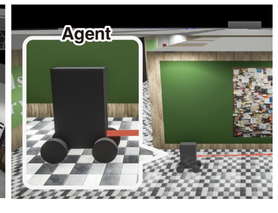
Observation-Based Instruction
from Past Research[2]

Memory-Based Instruction
from actual scenario[1]



Memory-Maze

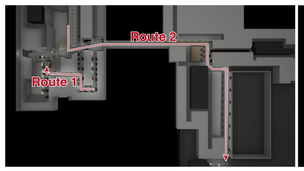
Virtual environments with CARLA [3]
- ✅ Realistic maze-like structure with by turns
- ✅ Can place static & dynamic obstacles
- ✅ Can obtain sensor data (e.g., LiDAR sensor)
- ✅ Can test with actual instruction collected

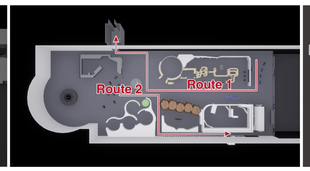


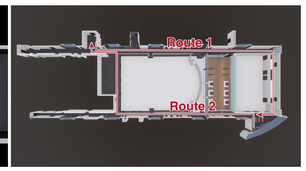
Instruction Collection

Setup: Participants and Routes

Online - 78 participants, 312 inst. Watch video and type in.
Online - 83 participants, 252 inst. Described solely from memory.







Memory-Based Instructions were longer, had diverse wordings and contained more errors.

	Route	Avg. Word #	Failure Rate	Alternative %
Online	University R1	60.8	0.00%	–
	University R2	89.8	6.06%	–
	Museum 5F R1	85.2	17.39%	–
	Museum 5F R2	65.6	6.82%	–
	Museum 7F R1	97.5	15.91%	–
	Museum 7F R2	83.2	2.18%	–
Onsite	University R1	88.4	25.0%	10.0%
	University R2	139.2	37.5%	15.0%
	Museum 5F R1	82.7	42.86%	0.0%
	Museum 5F R2	74.5	25.0%	61.35%
	Museum 7F R1	78.9	34.09%	6.75%
	Museum 7F R2	89.3	38.10%	87.9%

Example Memory-Based Instruction:
"So go straight for **20 meters and turn left. And then turn left.** There will be a hall connecting **Scott Hall and this is... Wait, this is... Newell Simon Hall.**[…] **If you turn right, there is a small... You can cross it**, but it's **probably** like one to two meters opening there. […] And then cross that hall. And then **I think** you can just go straight downstairs, **probably like 10 steps**, and then you will be there, the cafe."

Example Observation-Based Instruction:
"Go a little way down this road, then continue straight after turning left. Along the way, you will pass through a path lined with glass on both sides. After that, turn right at the dead end and follow the road, then turn right again before the stairs. Continue straight to *[…]*."

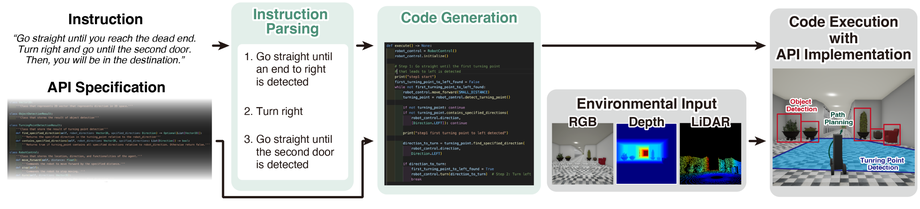
Observation-Based Instruction Word Cloud (University)

Memory-Based Instruction Word Cloud (University)

Model & Experiment

Model Summary
1. Parse the input inst. to step-by-step inst. with LLM (GPT-4).
2. Generate a navigation code for the agent by referring to API specification for robot control, using the same LLM.
3. Execute the generated code.

Python API for Robot Control
Examples:
```
move_forward(distance)
turn(direction)
detect_from_RGB_image(object)
detect_turning_point(object)
```



Our method outperformed state-of-the-art model and the results indicate the challenge of the benchmark.

Condition	Online Study Data SR↑	OSR↑	SPD↓	CLS↑	Onsite Study Data SR↑	OSR↑	SPD↓	CLS↑
NavGPT	0.01	0.03	67.37	0.04	0.00	0.02	64.19	0.05
NaVid	0.00	0.00	70.83	0.02	0.00	0.00	71.32	0.02
Proposed	0.09	0.15	45.45	0.34	0.06	0.09	51.75	0.32
Proposed (+Parser)	**0.11**	**0.22**	**37.29**	**0.42**	**0.08**	**0.12**	**40.39**	**0.44**

This work was supported by JSPS KAKENHI No. 23KJ2048 and 21H05054.
[1] Engel *et al.* 2020. ASSTES [2] Anderson *et al.* 2018. CVPR [3] Dosovitskiy *et al.* 2017. PMLR