Spatial-Aware Vision-Language Navigation of Mobile Agents

Final Presentation

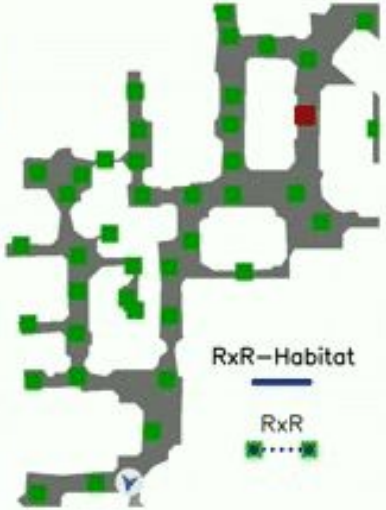
Team 4

Xiangchen Liu 20244050
Zhaoyan Wang 20244074

KAIST

Background: Vision-Language-Navigation (VLN)

"Given the egocentric image observation sequence with corresponding language instruction as input, following the text instruction and reach out to the target area."



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.

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

Literature Survey

ETPNav: Evolving Topological Planning for Vision-Language Navigation in Continuous Environments

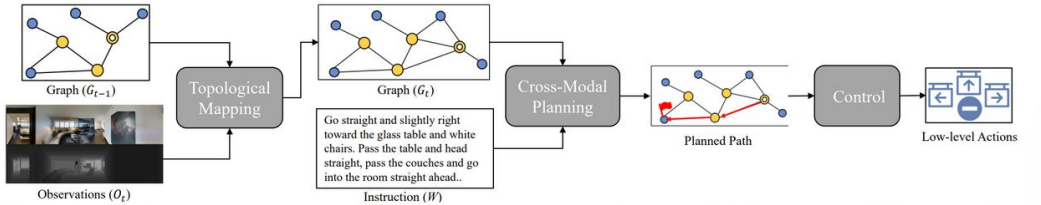


Fig. 1: Overview of the proposed model, ETPNav. It consists of three modules, a topological mapping module that gradually updates the topological map as it receives new observations, a cross-modal planning module that computes a navigational plan based on the instruction and map, and a control module that executes the plan with low-level actions.

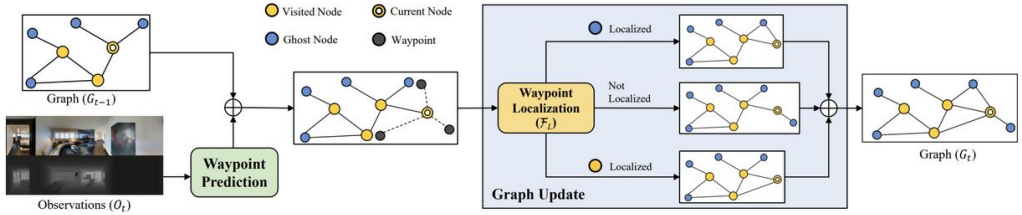


Fig. 2: Illustration of the topological mapping module. It takes the previous graph (G_{t-1}) and the agent observation (O_t) as input. The waypoint prediction submodule first predicts several nearby waypoints. The graph update submodule organizes these waypoints and incorporates them to update the graph using a waypoint localization function ($\mathcal{F}_L$).

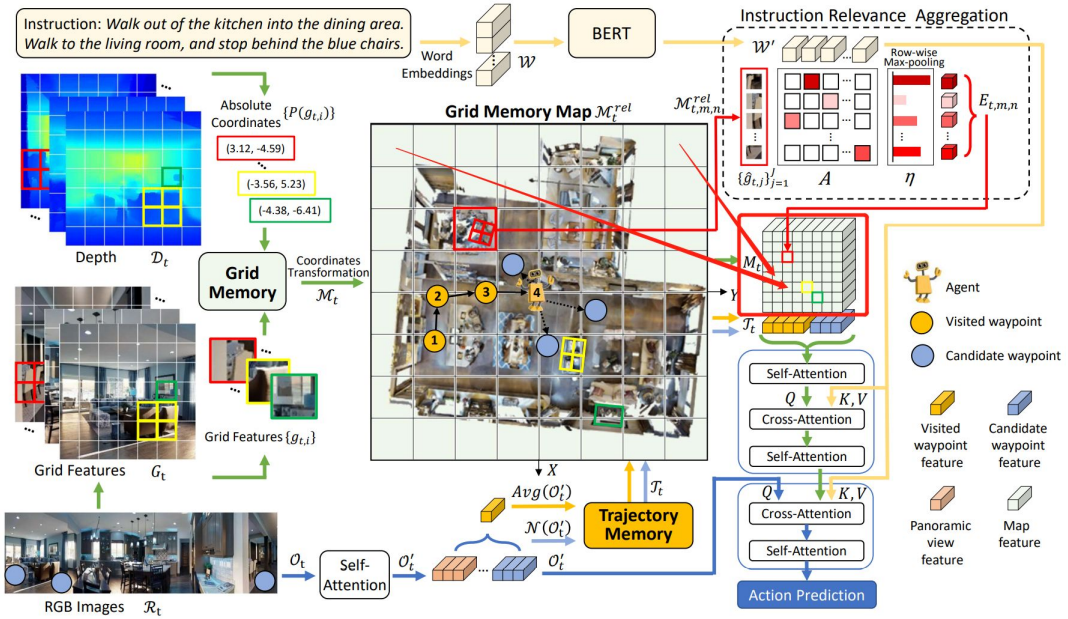
Contribution: The first Topology-graph-based VLN framework, including online/offline training

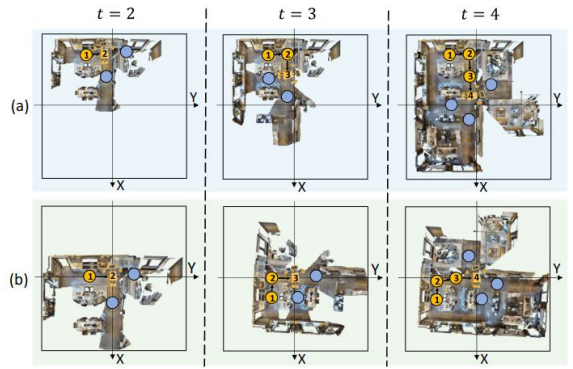
Limitation: Topology graph cannot extract detailed information of surrounding environment and spatial relationships between different rooms/objects

An, et. al. “Evolving Topological Planning for Vision-Language Navigation in Continuous Environment”, TPAMI 2024

Literature Survey

GridMM: Grid Memory Map for Vision-and-Language Navigation





(a): Maps with absolute coordinate
(b): Maps with relative coordinate

Contribution: Ego-centric adaptive resolution Grid Memory map for spatial reasoning and navigation, divide image as many small batches to get detailed feature information.

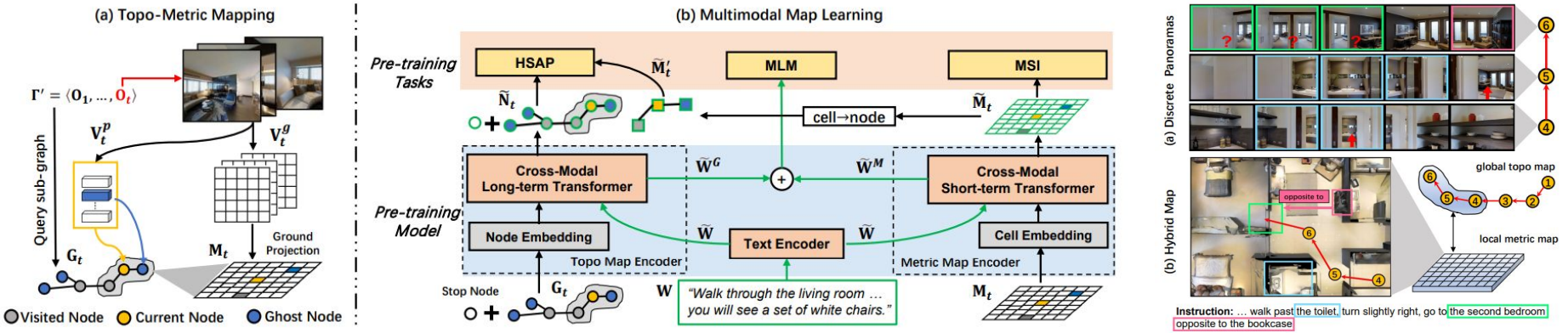
Limitation: It struggles with multi-floor indoor environment and fails to capture the room-level information

KAIST

Wang, et. al. "Grid Memory Map for Vision-and-Language Navigation", ICCV 2023

Literature Survey

BEVBert: Multimodal Map Pre-training for Language-guided Navigation



Contribution: Proposed a spatial-aware multimodal reasoning map for VLN task, use the local map to update the graph and the global map to learn the global-scale spatial relationship.

Limitation: It is not memory-efficient and cannot capture high-level spatial information

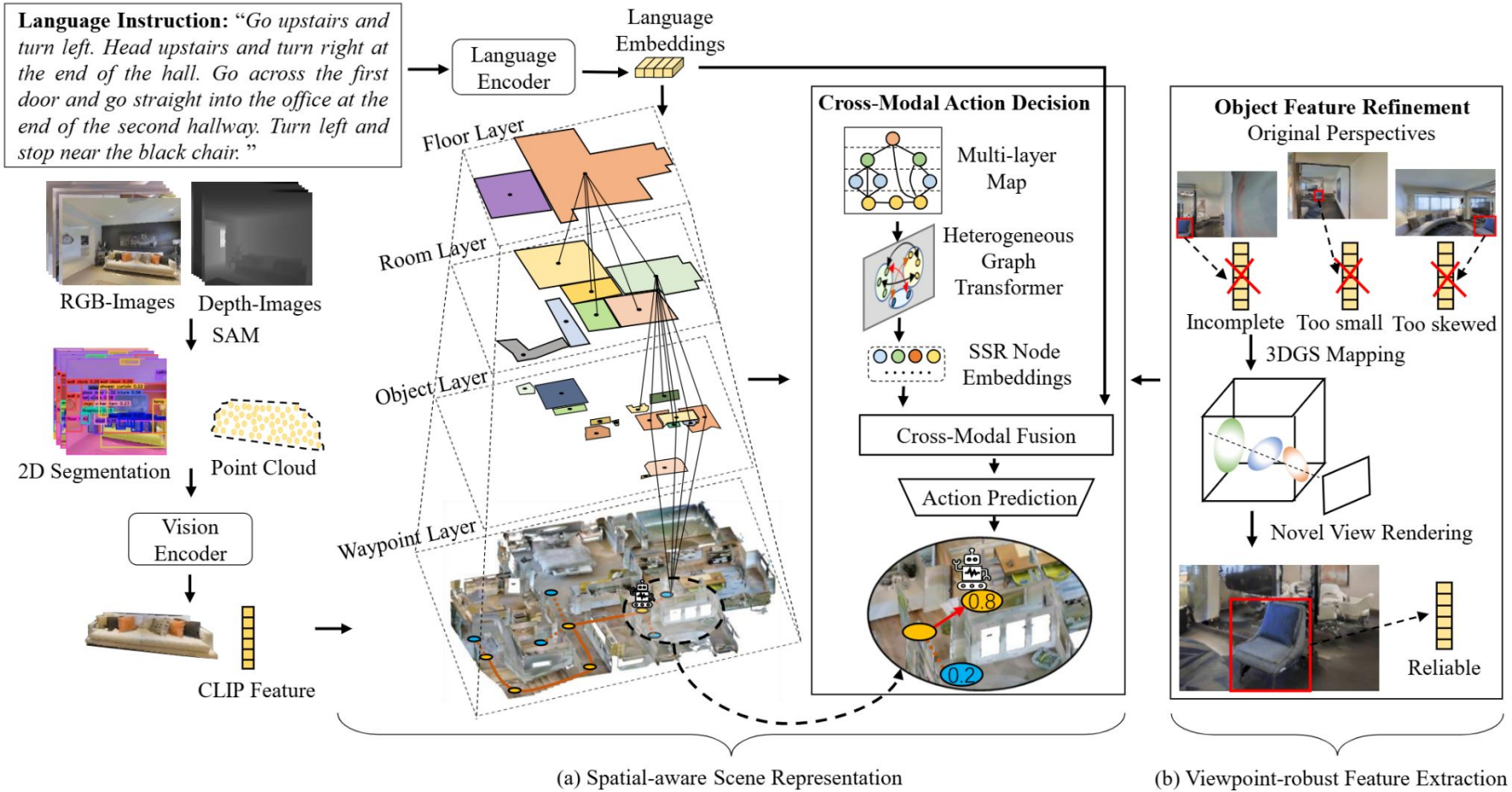
Goals & Limitations of prior works

Our Goals:

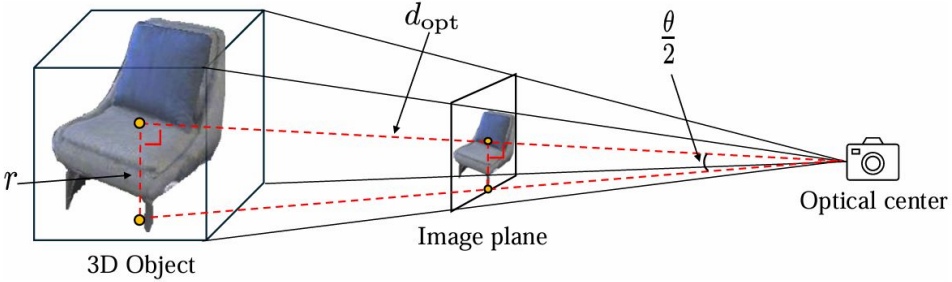
We propose a **Spatial-Aware Vision-Language Navigation (SA-VLN)** designed for **continuous environment**.

1. Design a novel **hierarchical scene representation** that includes waypoint-**object**-**room**-**floor** layers to simultaneously capture different levels of scene semantics and the agent's navigation history, enabling stable long-horizon and cross-floor navigation.
2. Propose to use **heterogeneous graph transformer** to extract semantic and spatial features from the multi-layer map and align them with the instruction embeddings to guide the agent's planning process.

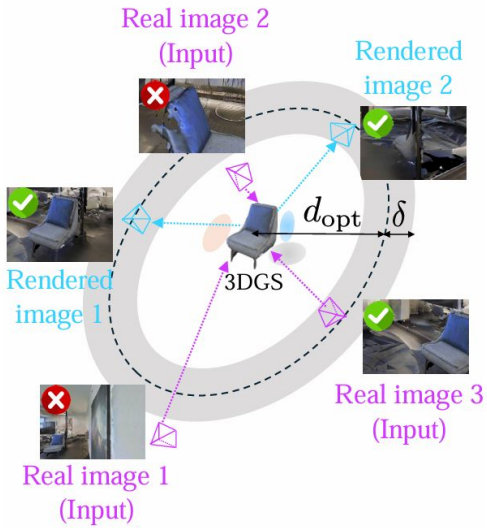
System Overview



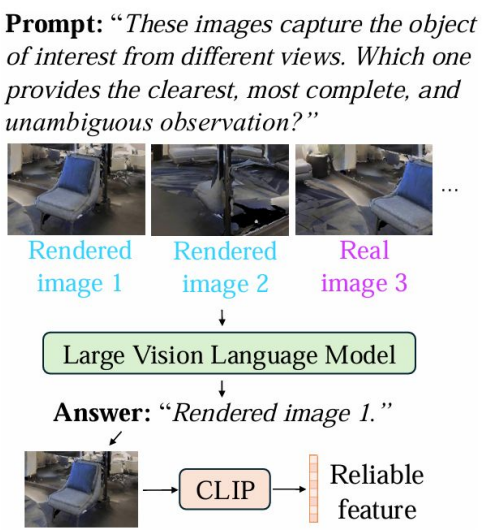
Methodology



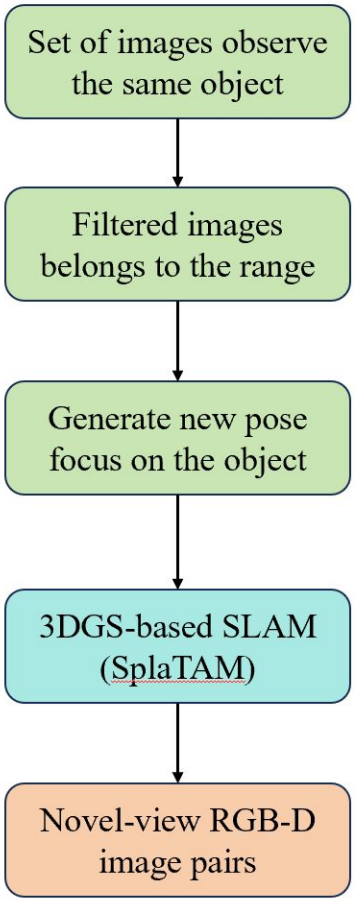
(a) Definition of the optimal observation distance d_{opt}



(b) Filtering and augmentation of views



(c) Optimal view selection



Our Contribution

1. Multi-layer Map Design: Introduces a novel map with waypoint layer to encode scene semantics and agent navigation history, enabling long-horizon and cross-floor navigation.
2. Heterogeneous Graph Transformer: Extracts semantic and spatial features from the map and aligns them with language instructions to guide effective planning.
3. **Viewpoint-Robust Feature Extraction: Filters unreliable views, synthesizes novel perspectives, and selects the optimal views based on vision-language model to improve map robustness and quality.**
4. **Superior Performance: Outperforms all state-of-the-art VLN models in continuous VLN environments, validating our proposed multi-layer scene representation and feature extraction methods.**

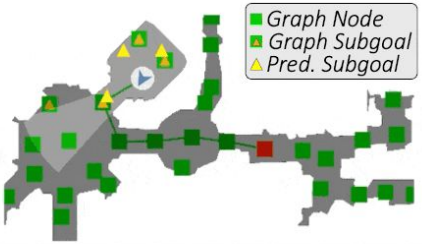
Experiment settings and datasets

Experiment settings: Vision-Language Navigation in Continuous environment (VLN-CE)

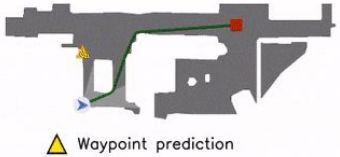
Datasets:

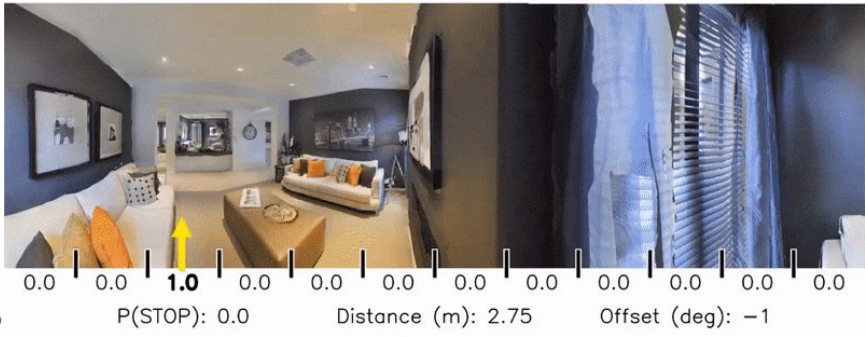
1. Room-to-Room in Continuous Environment (**R2R-CE** [1])
2. Room-across-Room in Continuous Environment (**RxR-CE** [2])

Walk through the arched entry way that leads into the tiled room with a mosaic floor. Walk through the arched entry way to the left of the stairs. Continue down the tiled hall through the open door ahead into the bedroom with red bedspread.



Walk around the brown leather ottoman, angling slightly towards the clock on the wall. Turn right at the clock and walk forward. Wait near the dining table.





Krantz, et. al. “Beyond the nav-graph: Vision-and-language navigation in continuous environments”, ECCV 2020
Ku, et. al. “Room-Across-Room: Multilingual vision-and-language navigation with dense spatio-temporal grounding”, EMNLP 2020

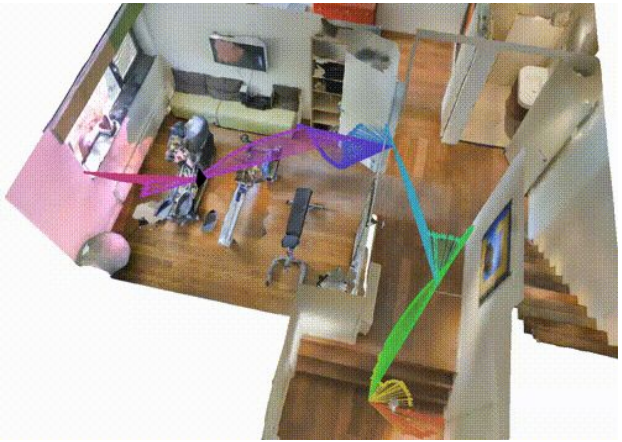
Experiment settings and datasets

Evaluation Metrics:

1. Navigation Error (NE): geometric distance in meters between the final and target location
2. Success Rate (SR): the ratio of paths with NE less than 3 meters
3. Oracle SR (OSR): SR given oracle stop policy
4. SR penalized by Path Length (SPL):

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

where $S_i = 1$ if the agent stops within the 3 m success radius (else 0), ℓ_i is the geodesic shortest-path length to the goal, and p_i is the agent's actual path length.



Preliminary results

Preliminary results on R2R-CE dataset: We outperform the single-layer scene representation-based methods on all sets in terms of NE, OSR, SR and SPL metrics. Especially, compared with our baseline methods, we surpasses in average 4% on val unseen split, 5% on test unseen split and 8% on val seen split, which validated the efficiency of our proposed method. This is the initial version of our SA-VLN and we are still in further refinement.

Methods	Val Seen				Val Unseen				Test Unseen			
	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑
GridMM [20]	4.21	69	60	53	4.44	58	50	44	5.64	56	46	39
ETPNav [17]	3.95	72	66	59	4.71	65	57	49	5.12	63	55	48
BEVBert [21]	3.45	78	71	61	4.57	67	59	50	4.70	67	59	50
SA-VLN (our)	**3.31**	**80**	**74**	**63**	**4.46**	**69**	**61**	**51**	**4.64**	**68**	**60**	**51**

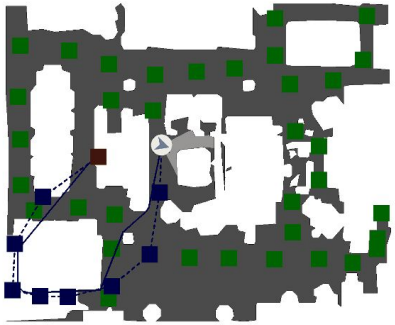
TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON R2R-CE [5] DATASET.

Methods	VS				VU				TU			
	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑
Reborn [43]	4.34	67	59	51	5.40	56	47	41	5.55	55	48	45
CM^2 [44]	6.10	50	42	34	7.02	41	34	27	7.70	39	31	24
WS-MGMap [45]	5.65	52	47	43	6.28	48	39	34	7.70	39	31	24
Sim-2-Sim [46]	4.67	61	52	44	6.07	52	43	36	6.17	52	44	37
CWP-RecBERT [26]	5.02	59	50	42	5.74	53	44	39	5.89	51	44	36
Ego^2-Map [47]	4.13	68	61	49	4.94	60	52	42	5.54	56	47	41
GridMM [16]	4.21	69	60	53	4.44	58	50	44	5.64	56	46	39
DREAMWALKER [48]	4.09	66	59	48	5.53	59	49	44	5.48	57	45	36
ScaleVLN [49]	3.82	74	68	59	4.80	64	55	51	5.11	63	55	50
ETPNav [13]	3.95	72	66	59	4.71	65	57	49	5.12	63	55	48
BEVBert [17]	3.45	78	71	61	4.57	67	59	50	4.70	67	59	50
Energy [33]	3.90	73	68	59	4.69	65	58	50	5.08	64	56	48
HNR-VLN [32]	3.67	76	69	61	4.42	67	61	51	4.81	67	58	50
SV-VLN (our)	**3.22**	**81**	**75**	**65**	**4.30**	**70**	**62**	**53**	**4.62**	**69**	**61**	**51**

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON RXR-CE [2] DATASET.

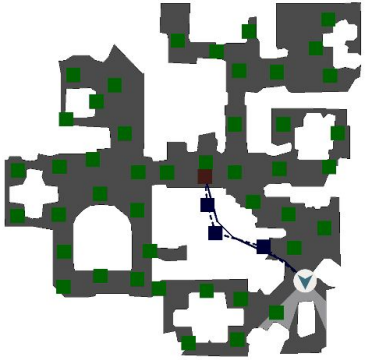
Methods	VS					VU				
	NE↓	SR↑	SPL↑	NDTW↑	SDTW↑	NE↓	SR↑	SPL↑	NDTW↑	SDTW↑
CWP-CMA [26]	8.62	32.64	26.01	51.14	26.72	8.76	26.59	22.16	47.05	23.65
CWP-RecBERT [26]	8.61	33.41	26.57	50.01	27.86	8.98	27.08	22.65	46.71	24.05
Reborn [43]	5.69	52.43	45.46	66.27	44.47	5.98	48.60	42.05	63.35	41.82
ETPNav [13]	5.31	60.32	49.63	65.42	49.86	5.92	53.93	43.91	61.06	44.25
Energy [33]	5.10	62.01	51.18	67.22	51.90	5.51	55.27	45.11	62.97	45.83
HNR-VLN [32]	4.85	63.72	53.17	68.81	52.78	5.51	56.39	46.73	63.56	47.24
SV-VLN (our)	**4.34**	**67.43**	**54.81**	**69.59**	**55.22**	**5.28**	**58.01**	**47.30**	**63.95**	**47.72**

Experiment: Cross-Floor Navigation

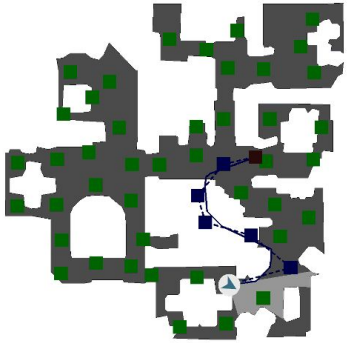


We're facing some armchairs and a couch. Let's turn to our right and go straight. And go down the stairs. On this floor we should be able to see a white open door. Go through the open door. We are in a bathroom and we can stop here.
The next action is to turn right 90 degrees.

Walk down the stairs and go to the room on the right and wait in the doorway of the bathroom.
The next action is to turn left 30 degrees.

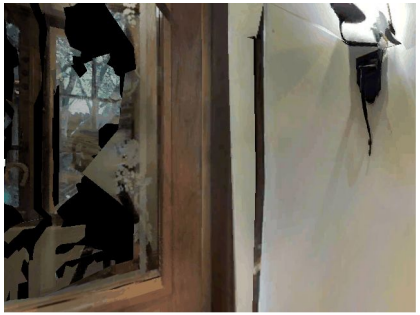


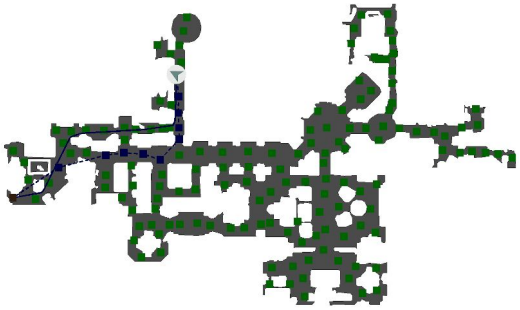
Go up the stairs and stop at the top in front of a mirror.
The next action is to turn left 90 degrees.



Exit dining room, walk towards the front door, go up stair case, turn right, stop in the doorway to bedroom.
The next action is to turn right 90 degrees.

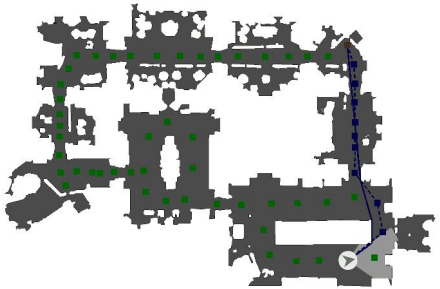
Experiment: Room-across-Room





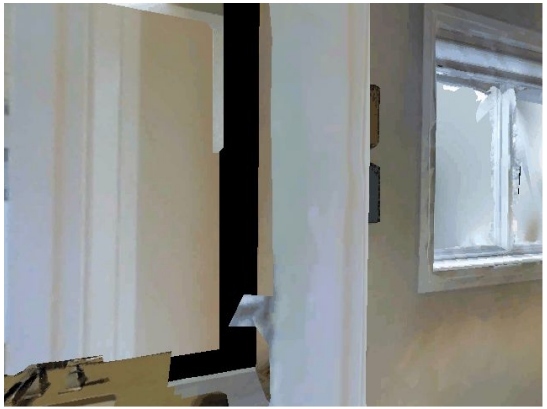
You will start by facing a door. Continue to walk down this hallway towards the arched opening at the end. You will see a kitchen island. Turn right and walk towards the sink. Turn right once again. Walk down the right side of this room towards the fireplace. On the right side of the fireplace you will see a double door opening. Walk inside this room. You will see a dining table just ahead of you. Jump over the chair and coffee table towards the left side of the dining table. Once you're standing in front of the flower decoration and the dining table is on your right then you're done.

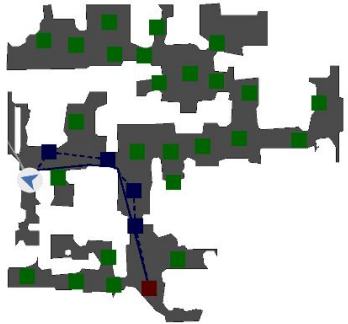
The next action is to turn right 90 degrees.



You are facing towards a pink wall and a chair, slightly turn to your left and move forward, again slightly turn to your left move forward, now slightly turn to your left there is an open doors exit the room through the open doors, now slightly turn to your right and move forward until you reach another open doors, in front of you exit the room through the open doors move forward, move forward until the white clay pot it is your end point.

The next action is to move forward 75 cm.





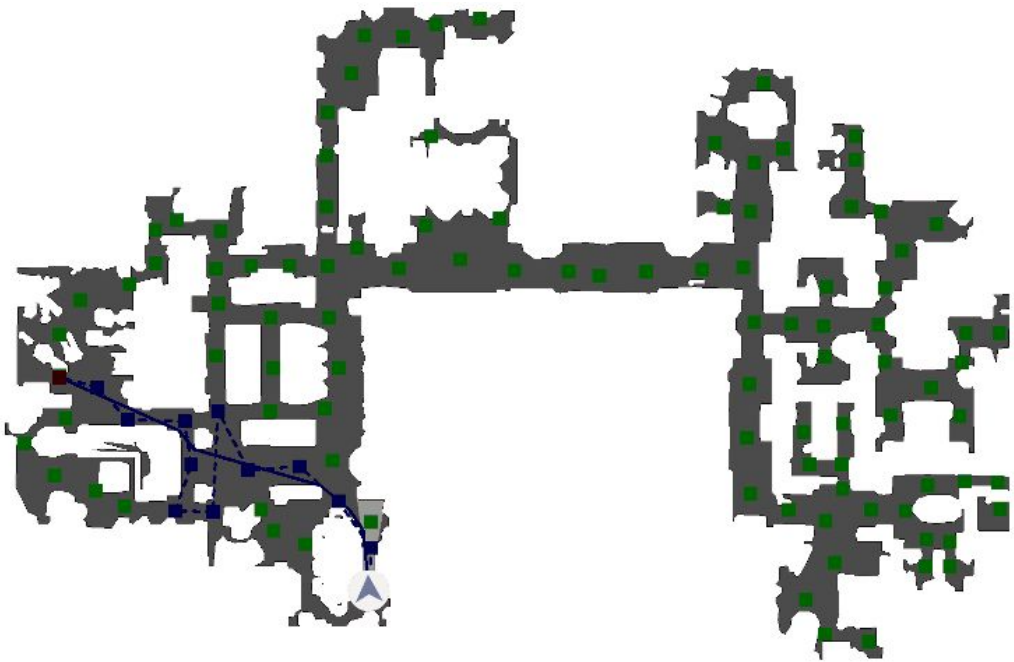
For your start point you will begin in a washroom exit the washroom and enter into the bedroom straight ahead, once you have done so turn to the right and exit the bedroom moving forward into the hallway. Once you are in the hallway turn to the right and enter into the bedroom through the open doorway straight ahead. In this bedroom you will see a shaggy area rug and a bed to your left and a single accent chair to your right. Move forward crossing the shaggy carpet move to the end corner of that shaggy carpet so that you right at that very end corner. You should be in–between the end of the bed on your left and the zebra printed accent chair to your right that has a little cute dog. I'm not sure if that's real, a real dog or a toy dog whatever it is once you are there you have reached your end point. And your done.

The next action is to turn right 60 degrees.

Right now you're facing towards a chair. Now turn left, you can see two pillars in front of you, move forward and stand in between them and move a bit forward. Now turn left and exit the room through the open door which is in front of you. There is an open door in front of you, move towards the door and enter in to the room, there is a table in front of you. Now turn left and move towards the second door which is in front of you. You can see a cross statue in front of you, move towards the cross statue and stand in front of it and it is your end point.

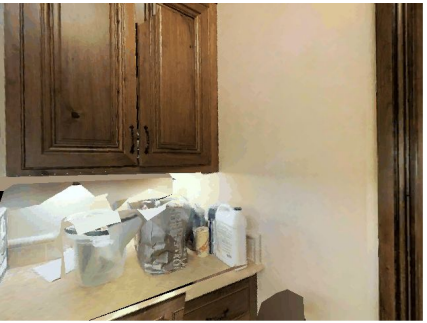
The next action is to turn left 90 degrees.

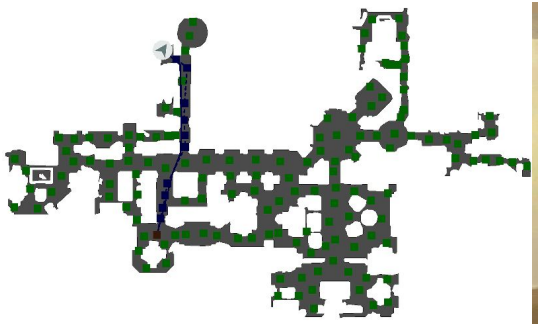
More visualization result



We start off looking at a pathway between a dining room table and a cabinet. Take a step down that pathway. Take one more step into the archway that's slightly to the left. You will now be looking into a hallway with that is L shaped. Take the hallway to the left. Take one more step down the hallway. And turn to your right you will now see the kitchen are. If you look towards the stove take a step towards that stove and oven and then turn back around. Take a step across the hallway where you just came from and into the small table area in the other room. So the room opposite to the kitchen room that you are in. Hop across the hall. And turn to your right. You should now take one step forward and you will now be next to some curtains and a painting of a native american woman holding her daughter. Turning around so that painting is behind you. Walk down that pathway that's goes next to the table that's next to the L-shaped couch. Walk down that pathway and you will see a nother painting of some farm land. Turn to the left and walk down that pathway. Now on your right you will see a workout area. Take a step to the front of the treadmill. Now take a step right next to the rowing machine and you don't know what that is then maybe the brown cabinet on your left and you are done.

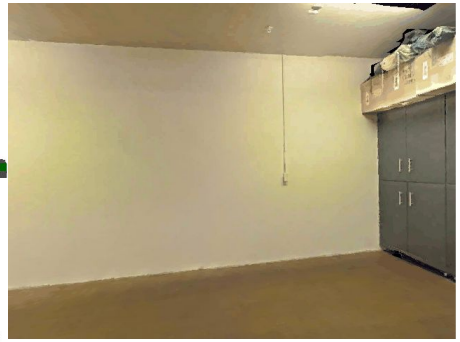
More visualization result

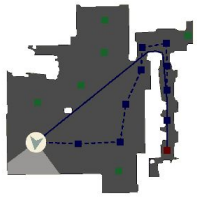




You are facing some cabinets, turn right exit the room in the brown floor hall way, turn right and proceed straight ahead towards the large arched hall way, once you are near the white and brown counter in the kitchen turn right towards the counter and the brown and the sink counter on right, proceed straight ahead through the arch way moving towards the round table with chairs, once you are at the back of the front chair, you are done.

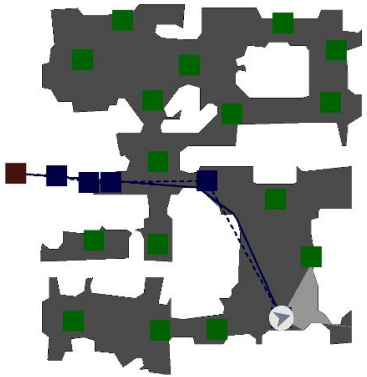
The next action is to turn right 90 degrees.





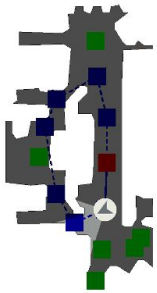
Turn around. You are standing in a garage, to the left there is a black vehicle. Look forward, straight ahead walk towards the storage unit. Once you have reached the storage unit, turn left. In front of you there is a doorway, walk towards the door way passing the storage unit. Once you have made your way through the door way you will see two doors and a white stair case on your right hand side. To the right there's a doorway, go through that door way. Once you have reach through the door way you will see two sets of wine racks on your right and left hand side. Go straight ahead. There is a second set of wine racks on your left hand side and a small seating area on your right hand side, go between the wine rack and the seating area straight ahead. In front of you is a barrel and seating area, turn around you have reached your final destination.

The next action is to turn left 90 degrees.



You are facing towards the round table, turn left and move towards the open door. Turn slight left and move towards the steps. Now get down of the steps. You are facing towards an open door, now enter into the room. You are facing towards the window, which is your final destination.

The next action is to turn left 90 degrees.



Right now you are standing in a wooden room beside the door which has a bed. Move forward towards the bed, turn right and exit the room. You will enter into a washroom with a commode, walk towards the commode and exit the bathroom as well. You will enter into a room with another bed and also a chair in it. Walk towards the chair and turn right. You will find a brown carpet, walk towards the carpet, take a few steps forward and then you will reach the endpoint beside the cupboard.

The next action is to move forward 75 cm.

Roles of each member

- Paper reading (XL, ZY)
- Idea development & Brainstorming (XL, ZY)
- System development (XL, ZY)
- HGT development (ZY)
- Hierarchical Representation (XL)
- Experiment (XL, ZY)

XL: Xiangchen Liu
ZY: Zhaoyan Wang