CS586 (25 Spring) : Paper Presentation

NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation (RSS 24)

20244076 Zhaoyan Wang

20244050 Xiangchen Liu

KAIST

Review

Efficient Residual Learning with Mixture-of-Experts for Universal Dexterous Grasping (ICLR 2025)

Improvements

- **Residual Policy Learning Framework**
  - Efficiently trainable model
- **Geometry-Agnostic Base Policy**
  - High generalization
- **Mixture-of-Experts (MoE)**
  - High diversity and good performance

Limitation

- **No functional grasping**
- **No experiment on hardware**

Conclusion

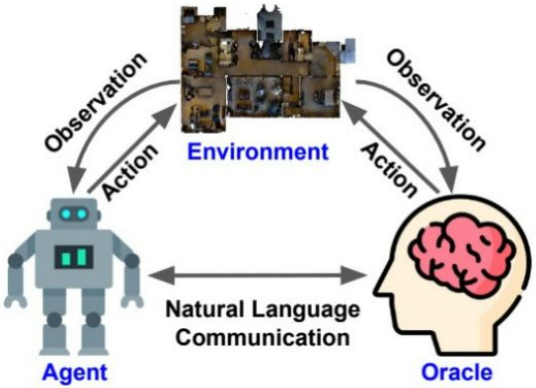
- **Perform better than previous works**
- **zero generalization gap**

Table of contents

Introduction & Motivation

Vision-and-Language Navigation (VLN)

Goal: Humans communicate with each other using natural language to issue tasks and request help.



[Gu et. al. 2022]

Amazon Astro Robot

A robot that can understand human language and navigate intelligently would significantly benefit human society.

[1] Gu, Jing et al. "Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions." .arXiv preprint arXiv:2203.12667 (2022).
[2] https://www.cnet.com/home/smart-home/amazon-astro-review

Introduction & Motivation

Vision-and-Language Navigation (VLN)

Given free-form instruction, the robot is required to follow the instruction to navigate in the unseen environments.

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

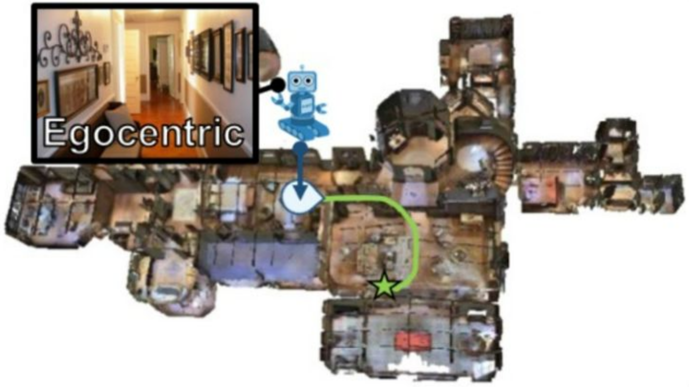
Observation:

- Egocentric color map
- Egocentric depth map
- Location and orientation

Action:

- Low-level actions

(Move forward, Turn left, Turn right, Stop)



VLN-CE

[1] Krantz, Jacob et al. "Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments.", ECCV 2020

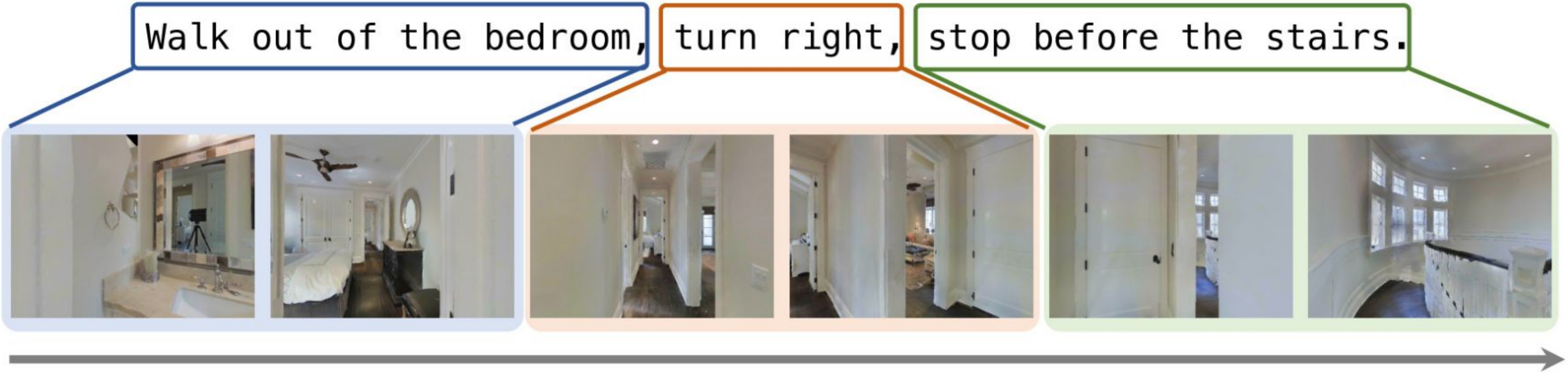
KAIST

Introduction & Motivation

Challenges

High-level understanding:

- Understand long-horzion trajctory with rich visual information.
- Understand free-from text instruction.
- Align the instruction with history trajectory.



Introduction & Motivation

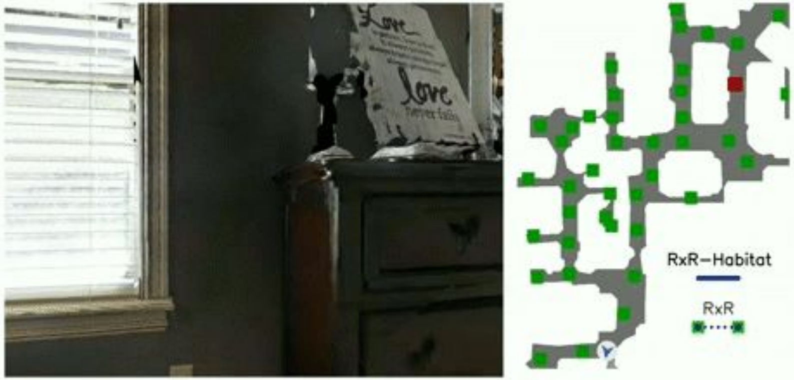
Challenges

High-level understanding:

- Understand long-horzion trajctory with rich visual information.
- Understand free-from text instruction.
- Align the instruction with history trajectory.

Low-level planning:

- Approching landmarks
- Obstacle avoidance

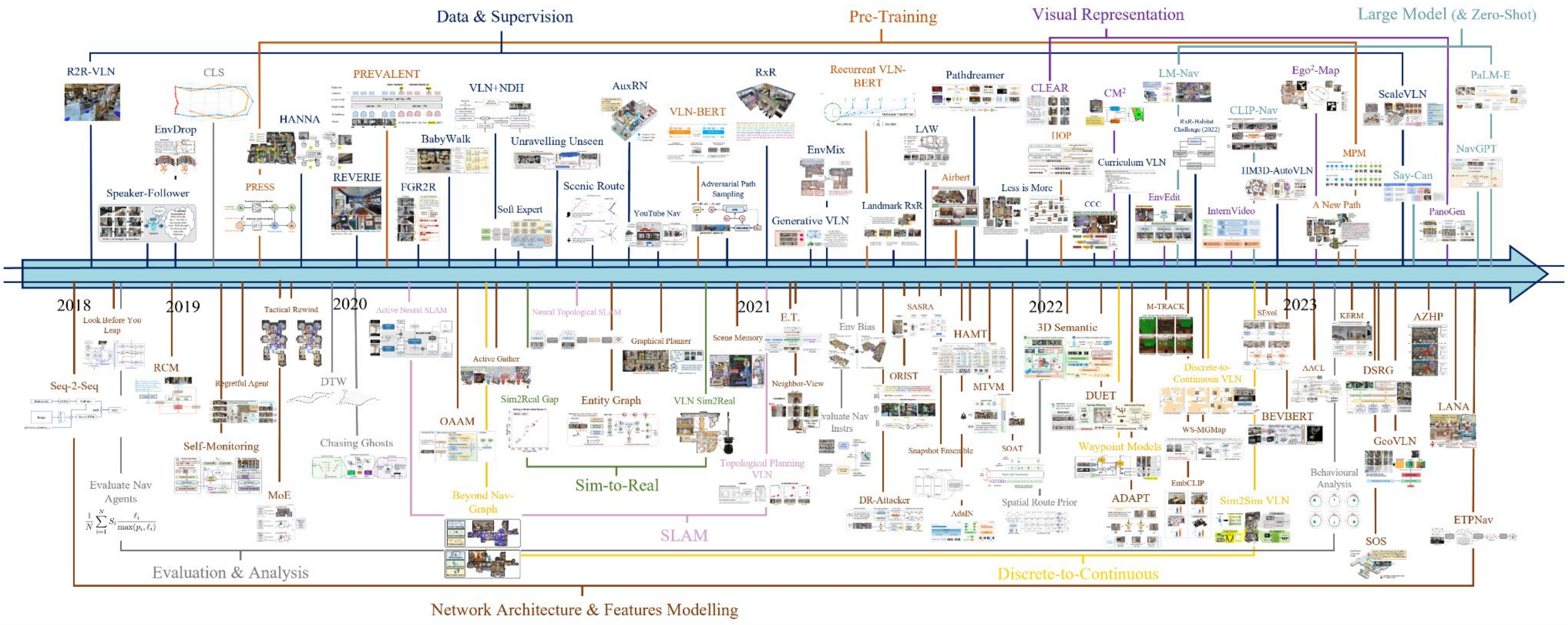


You are in a bedroom. Turn around to the left until you see a door leading out into a hallway, go through it. Hang a right and walk between the island and the couch on your left. When you are between the second and third chairs for the island stop.

VLN-CE RxR

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

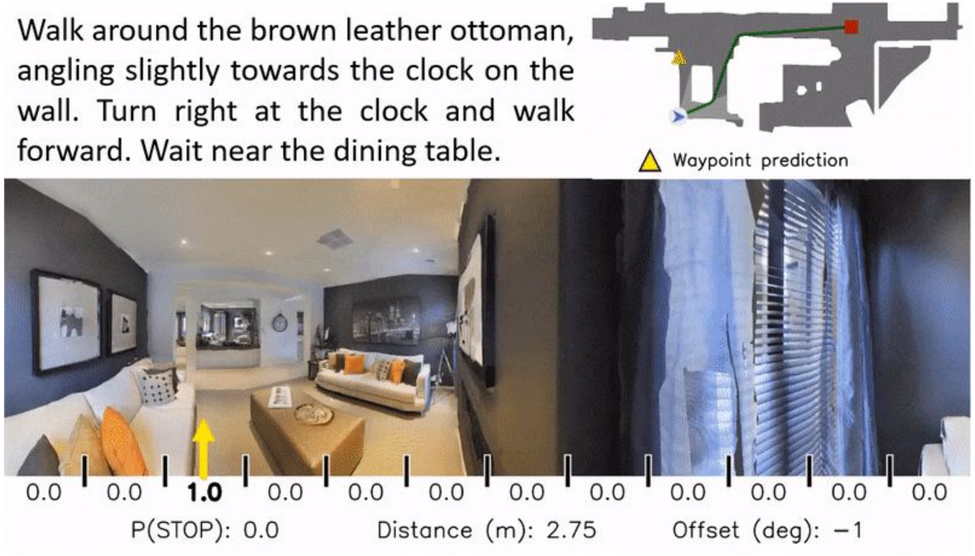
Introduction & Motivation



[1] https://github.com/YicongHong/Thinking-VLN/tree/main

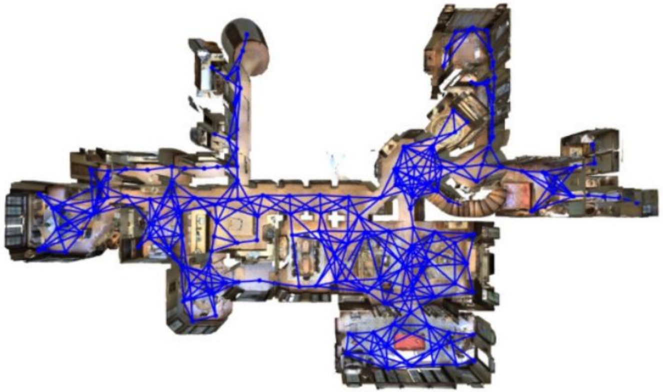
KAIST

Related works



(A) Panoramic images

- Large field of view



(B) Topology node graph

- Known Topology
- Oracle Navigation
- Perfect Location

[1] Krantz, Jacob et al. “Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments.”, ECCV 2020
[2] Krantz, Jacob et al. "Waypoint models for instruction-guided navigation in continuous environments.", ICCV 2021

Related works

Simplified Settings:
1. Panorama Observation
2. Discrete Connectivity Graph
3. Jump-Point motion

Problem:
1. Super Large Sim2Real Gap
2. No discrete graph in physical world
3. Pose Estimation is imperfect.

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

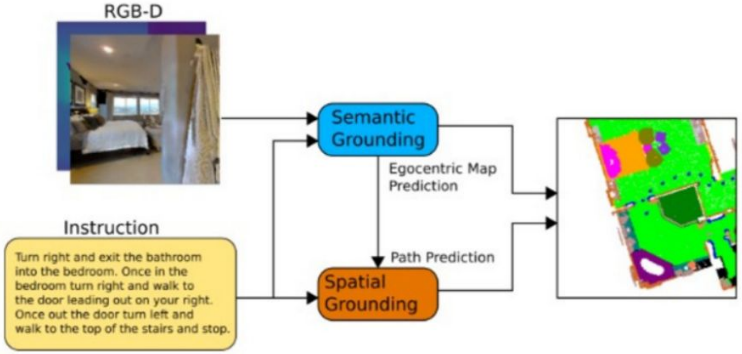
[1] Krantz, Jacob et al. "Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments.", ECCV 2020
[2] Krantz, Jacob et al. "Waypoint models for instruction-guided navigation in continuous environments.", ICCV 2021

Related works

Map-based VLN method

Cross-modal Map Learning for Vision and Language Navigation

Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan,
Eleni Miltsakaki, Dan Roth, Kostas Daniilidis
University of Pennsylvania
{ggeorgak,karls,kwanchoo,sohamdan,elenimi,danroth,kostas}@seas.upenn.edu
Project webpage: https://ggeorgak11.github.io/CM2-project/



Build top-down map → **BERT** → Text-to-map attention → High related path → Local Planner

Noiseless depth image, rotation and translation estimation.

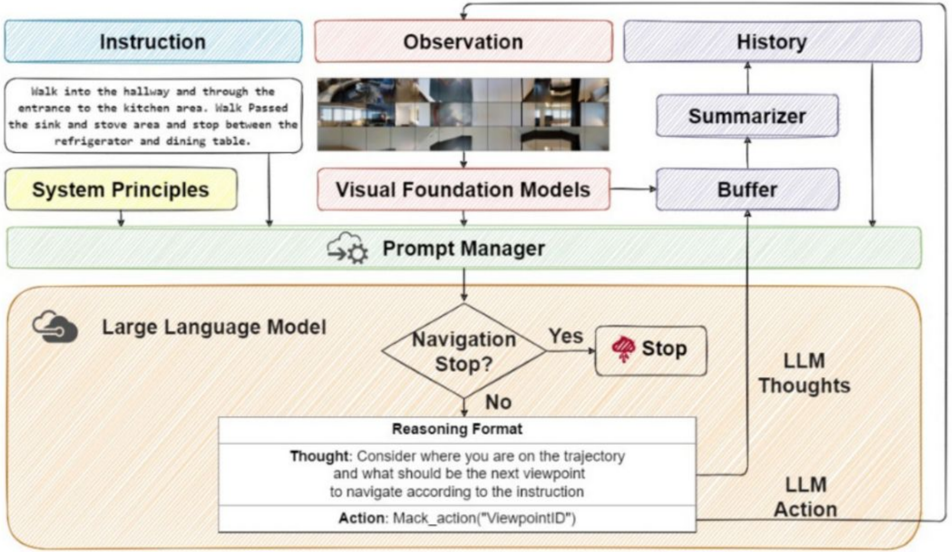
[1] Georgakis, Georgios et al. "Cross-modal map learning for vision and language navigation.", CVPR 2022.

Related works

LLM-based VLN method

NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models

Gengze Zhou[1] Yicong Hong[2] Qi Wu[1]
[1]The University of Adelaide [2]The Australian National University
{gengze.zhou, qi.wu01}@adelaide.edu.au yicong.hong@anu.edu.au
https://github.com/GengzeZhou/NavGPT



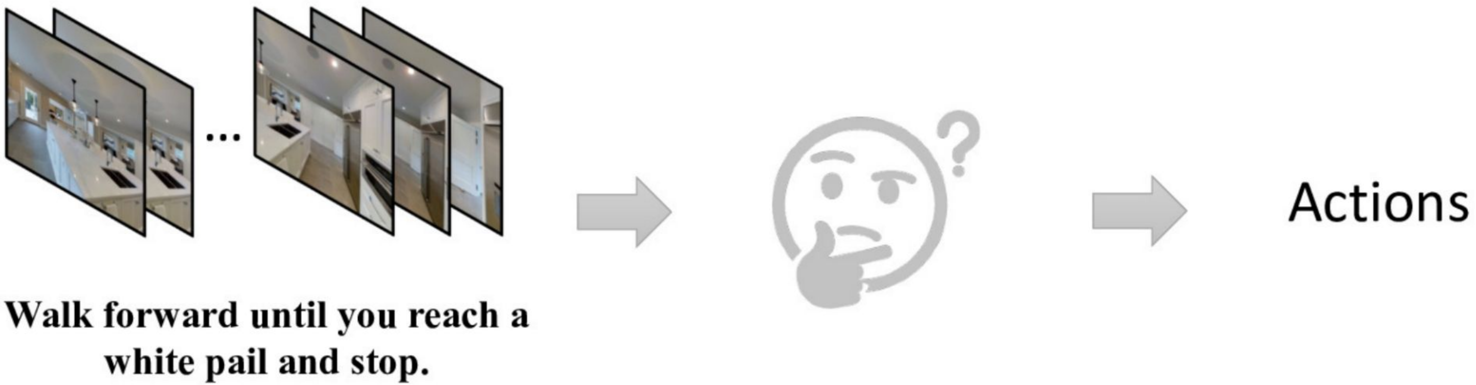
Foundation models + observation images → Text-based navigation history → **ChatGPT** → Node selection

[1] Zhou, Gengze, et al. "Navgpt: Explicit reasoning in vision-and-language navigation with large language models." AAAI 2023.

Related works

Is there a simple way to achieve VLN?

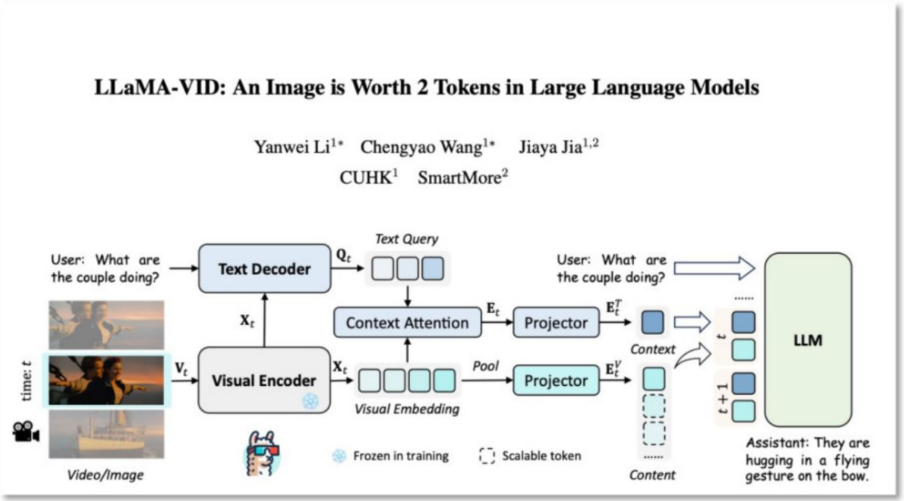
We want a straightforward solution for VLN.



Related works

Video-Based Vision-Language Model (VLM)

- Strong performance in understanding video and text.
- Generalizability to novel videos and texts.



LLaMA-VID

VID-LLaVA

[1] Li, Yanwei et al. "LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models." ArXiv abs/2311.17043
[2] Lin, Bin et al. "Video-LLaVA: Learning United Visual Representation by Alignment Before Projection." ArXiv abs/2311.10122

Problem Formulation

We are focusing on a straightforward and challenging solution: using only RGB video as input and directly output low-level actions from a video-based VLM.

Walk forward until you reach a white pail and stop.

Vision Language Model

Actions

- Move forward
- Turn left
- Turn right
- Stop

➢ An intuitive way to drive the agent.

➢ Eliminates the need for location, orientation, and depth information.

Problem Formulation

From Video-based Question & Answer (VQA) to Navigation

Walk forward until you reach a white pail and stop.

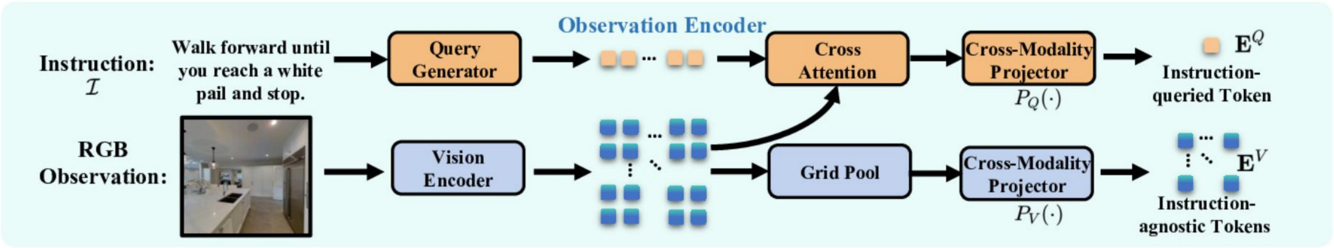
History RGB sequence | Current Frame | Instruction

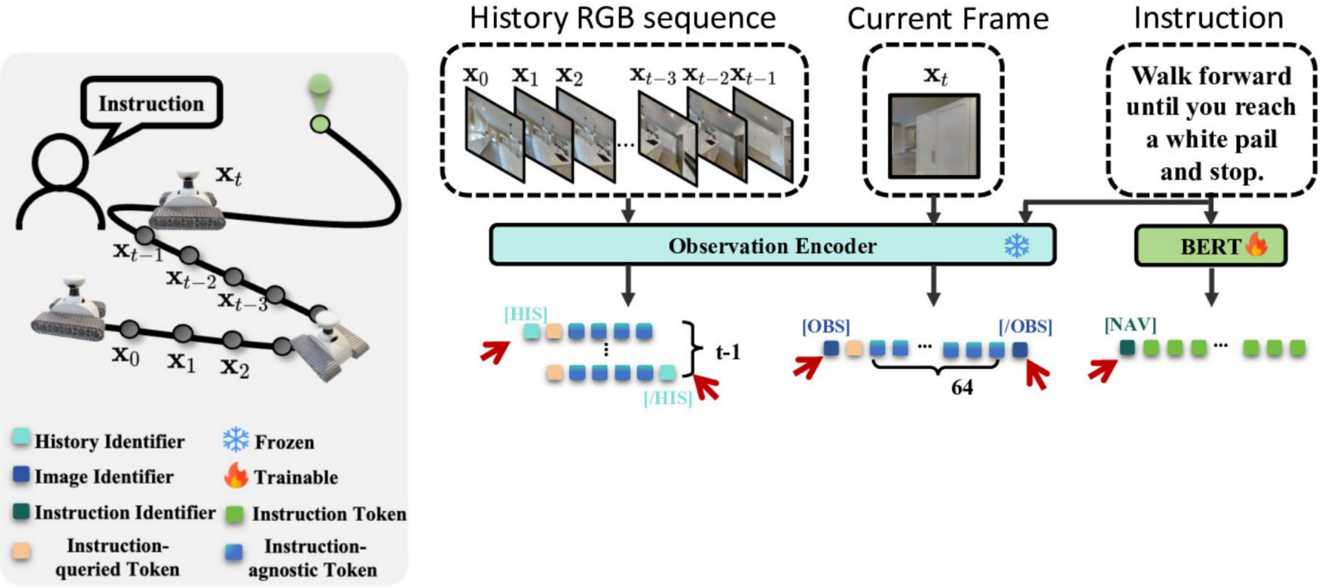
- The modality of VLN is different from the common modalities of large models.

  <u>Design a new pipeline of video-based VLM for VLN</u>

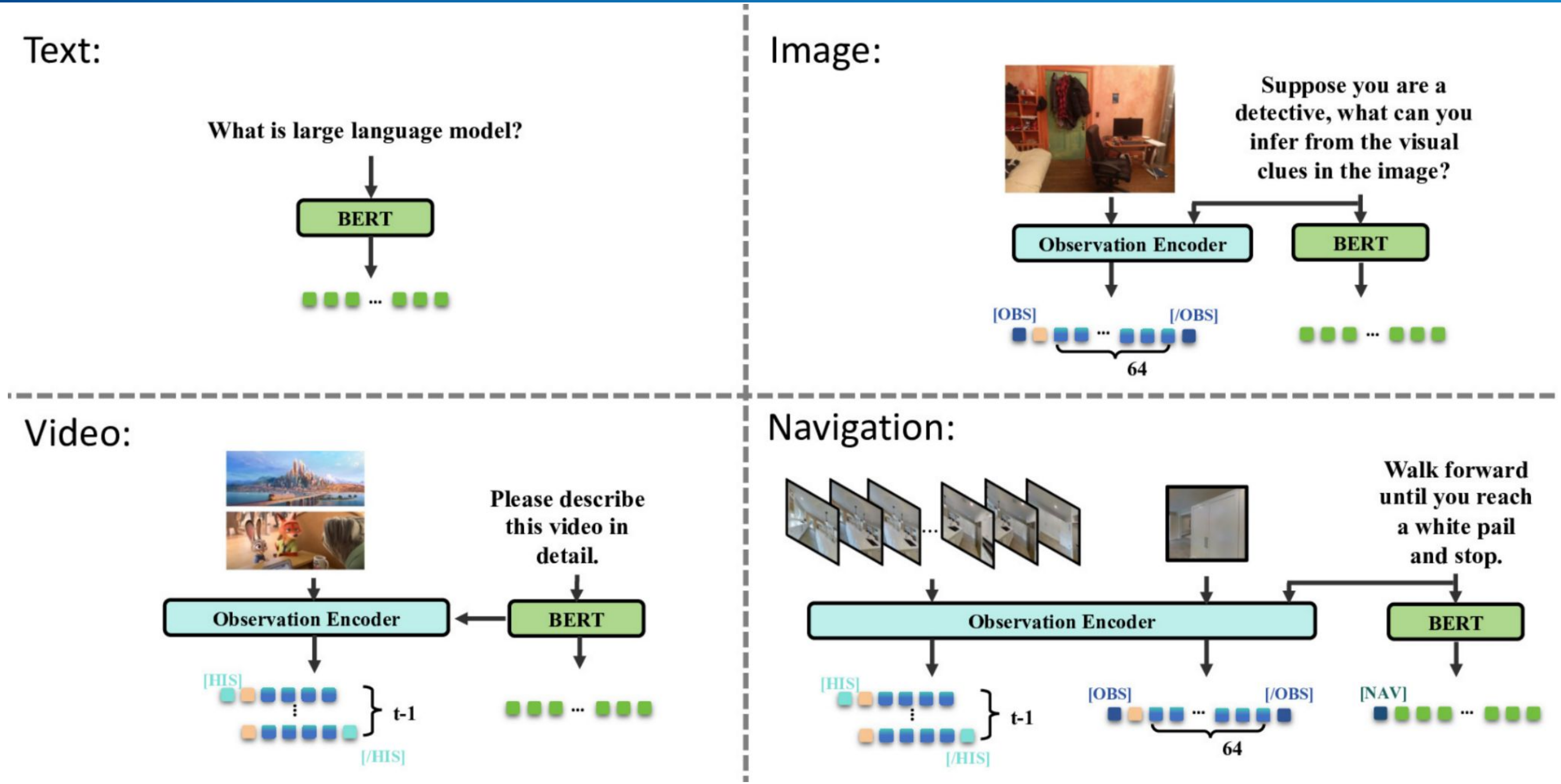
- There is a lack of large amounts of high-quality real data for VLN task.

Methodology: THE PROPOSED NAVID AGENT

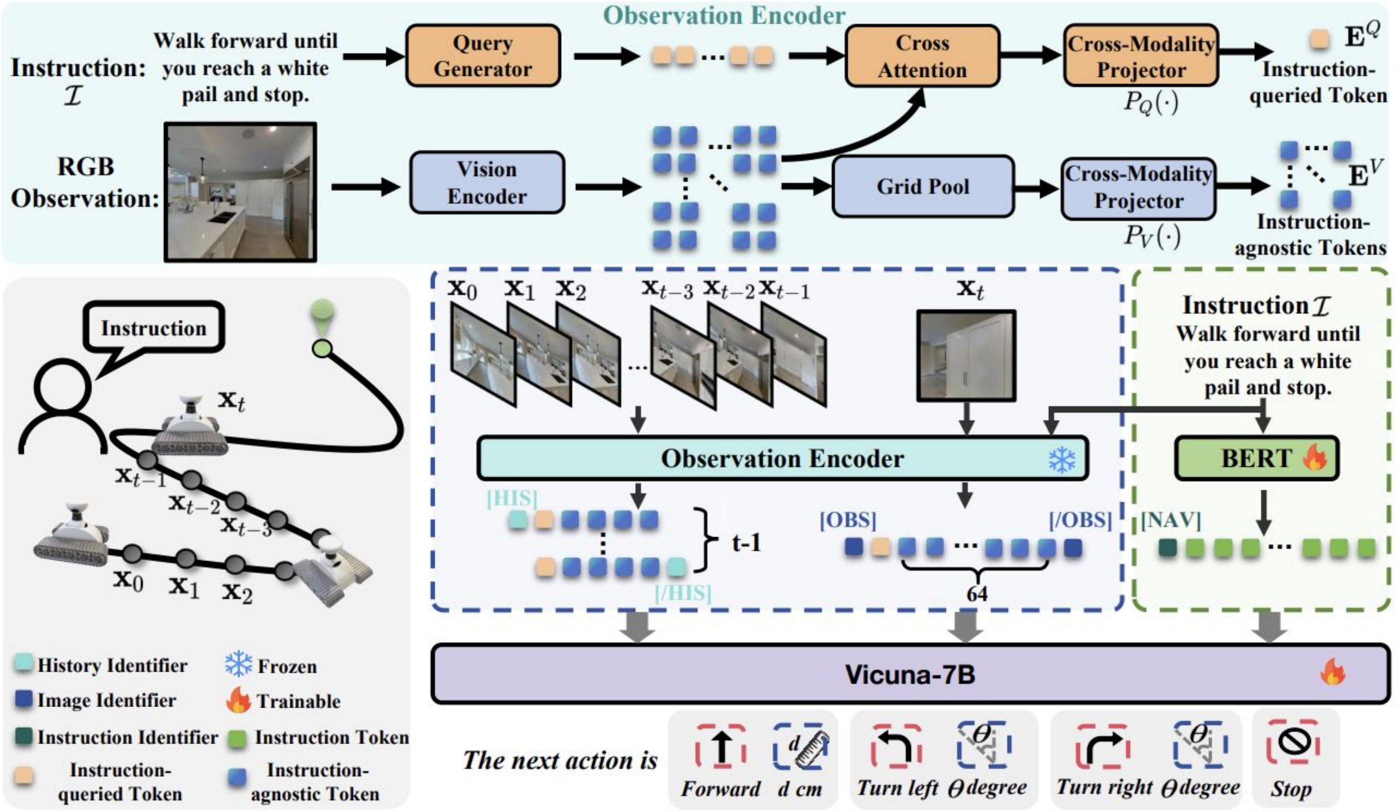




Methodology: THE PROPOSED NAVID AGENT



Methodology: THE PROPOSED NAVID AGENT



Methodology: THE PROPOSED NAVID AGENT

From video QA to navigation

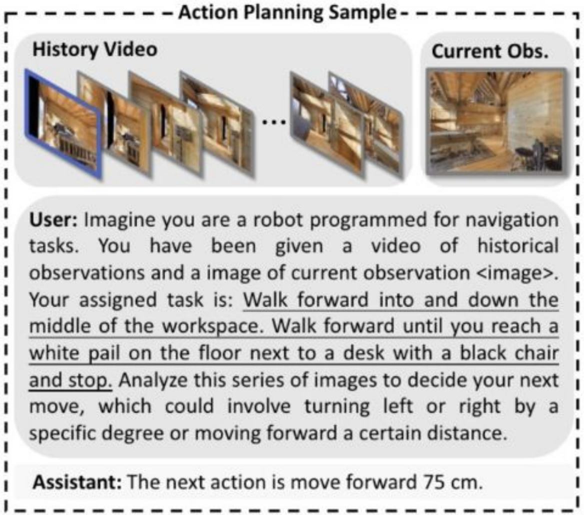
The modality of VLN is different from the common modalities of large models.

There is a lack of large amounts of high-quality real data for VLN task.

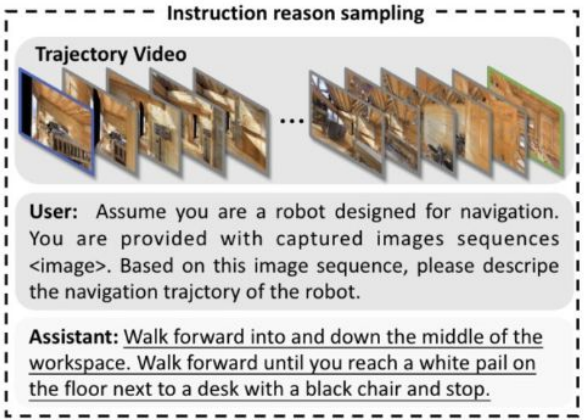
<u>Collect simulator data for training NaVid</u>

Data collection

We collect the navigation data based on R2R dataset training-split on VLN-CE simulator: 10819 episodes, 61 scenes (MP3D).



Sample video segment + action
(Action Planning Sample)



Video + instruction
(Instruction Reasoning Sample)

Experiments

We collect the navigation data based on R2R dataset training-split on VLN-CE simulator:
10819 episodes, 61 scenes (MP3D).

320k Action Planning Samples
10k Instruction Reasoning Sample

40k Text QA
625k Image QA
98k Video QA

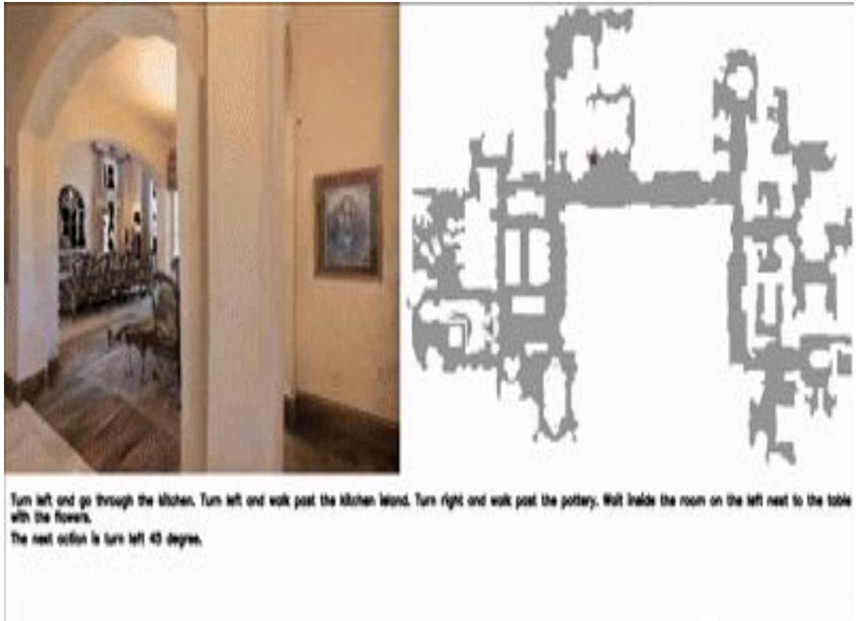
DAGGER

180k Action Planning Samples

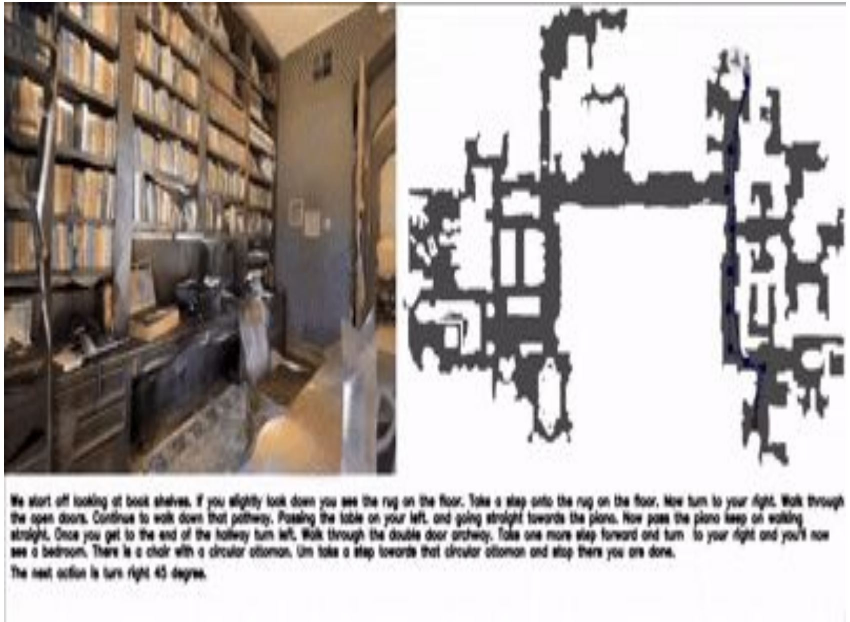
(Total 1.2M training samples)

Experiments

R2R train -> R2R val-unseen



R2R train -> RxR val-unseen



KAIST

Experiments

R2R train -> R2R val-unseen (cross split)

	Observation				VLN-CE R2R Val-Unseen				
	Pan.	S.RGB	Depth	Odo.	TL	NE↓	OS↑	SR↑	SPL↑
AG-CMTP [15]	✓		✓	✓	–	7.90	39.2	23.1	19.1
R2R-CMTP [15]	✓		✓	✓	–	7.90	38.0	26.4	22.7
LAW [73]		✓	✓	✓	8.89	6.83	44.0	35.0	31.0
CM2 [29]		✓	✓	✓	11.54	7.02	41.5	34.3	27.6
WS-MGMap [16]		✓	✓	✓	10.00	6.28	47.6	**38.9**	34.3
Seq2Seq [43]		✓	✓		9.30	7.77	37.0	25.0	22.0
CMA [43]		✓	✓		8.64	7.37	40.0	32.0	30.0
RGB-Seq2Seq		✓			4.86	10.1	8.10	0.00	0.00
RGB-CMA		✓			6.28	9.55	10.8	5.00	4.43
Ours		✓			7.63	**5.47**	**49.1**	37.4	**35.9**

↑ SR (success rate)
↑ OS (oracle success rate)
↑ SPL (success weighted by path length)
↓ NE (Navigation error)

SOTA level performance with only RGB video inputs

Experiments

R2R train -> R2R val-unseen (cross split)

	Observation				VLN-CE R2R Val-Unseen				
	Pan.	S.RGB	Depth	Odo.	TL	NE↓	OS↑	SR↑	SPL↑
AG-CMTP [15]	✓		✓	✓	–	7.90	39.2	23.1	19.1
R2R-CMTP [15]	✓		✓	✓	–	7.90	38.0	26.4	22.7
LAW [73]		✓	✓	✓	8.89	6.83	44.0	35.0	31.0
CM2 [29]		✓	✓	✓	11.54	7.02	41.5	34.3	27.6
WS-MGMap [16]		✓	✓	✓	10.00	6.28	47.6	**38.9**	34.3
Seq2Seq [43]		✓	✓		9.30	7.77	37.0	25.0	22.0
CMA [43]		✓	✓		8.64	7.37	40.0	32.0	30.0
RGB-Seq2Seq		✓			4.86	10.1	8.10	0.00	0.00
RGB-CMA		✓			6.28	9.55	10.8	5.00	4.43
Ours		✓			7.63	**5.47**	**49.1**	37.4	**35.9**

↑SR (success rate)
↑OS (oracle success rate)
↑SPL (success weighted by path length)
↓NE (Navigation error)

Under the same setting, our method demonstrates signifcant improvements on all metrics. (648% improvment on SR and 710% imprivment on SPL.)

KAIST

Experiments

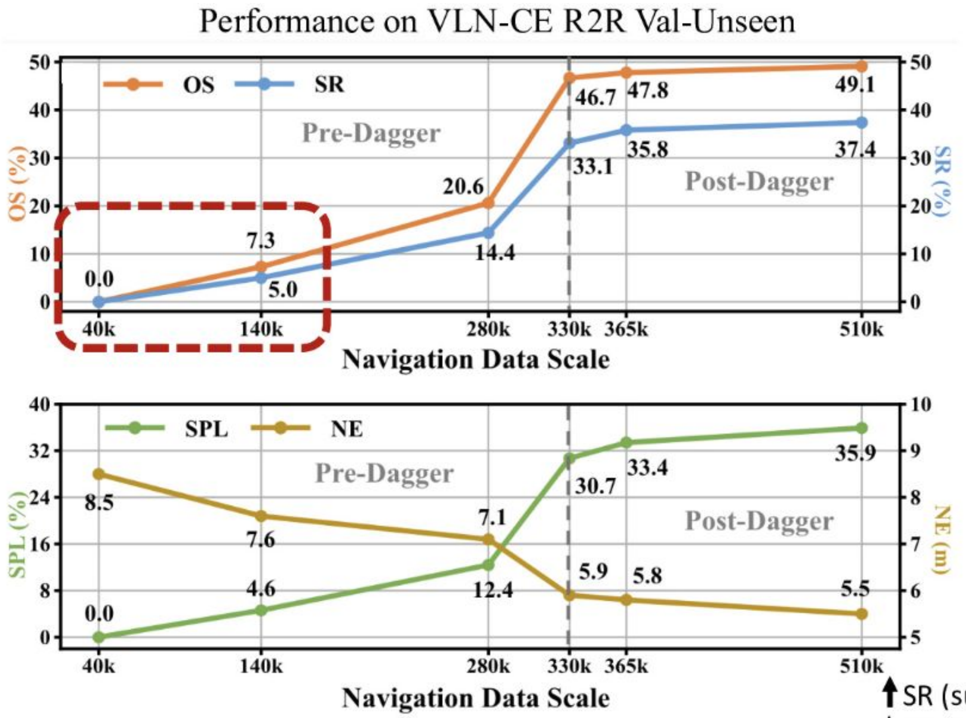
R2R train -> RxR val-unseen (cross dataset)

	Observation			VLN-CE RxR Val-Unseen				
	S.RGB	Depth	Odo.	TL	**NE↓**	**OS↑**	**SR↑**	**SPL↑**
LAW [73]	✓	✓	✓	4.01	10.87	21.0	8.0	8.0
CM2 [29]	✓	✓	✓	12.29	8.98	25.3	14.4	9.2
WS-MGMap [16]	✓	✓	✓	10.80	9.83	29.8	15.0	12.1
Seq2Seq [43]	✓	✓		1.16	11.8	5.02	3.51	3.43
CMA [43]	✓	✓		5.09	11.7	10.7	4.41	2.47
RGB-Seq2Seq	✓			4.43	11.2	12.2	0.0	0.0
RGB-CMA	✓			13.56	9.55	14.8	0.0	0.0
A^2Nav [17]	✓			–	–	–	16.8	6.3
Ours	✓			10.59	**8.41**	**34.5**	**23.8**	**21.2**

↑SR (success rate)
↑OS (oracle success rate)
↑SPL (success weighted by path length)
↓NE (Navigation error)

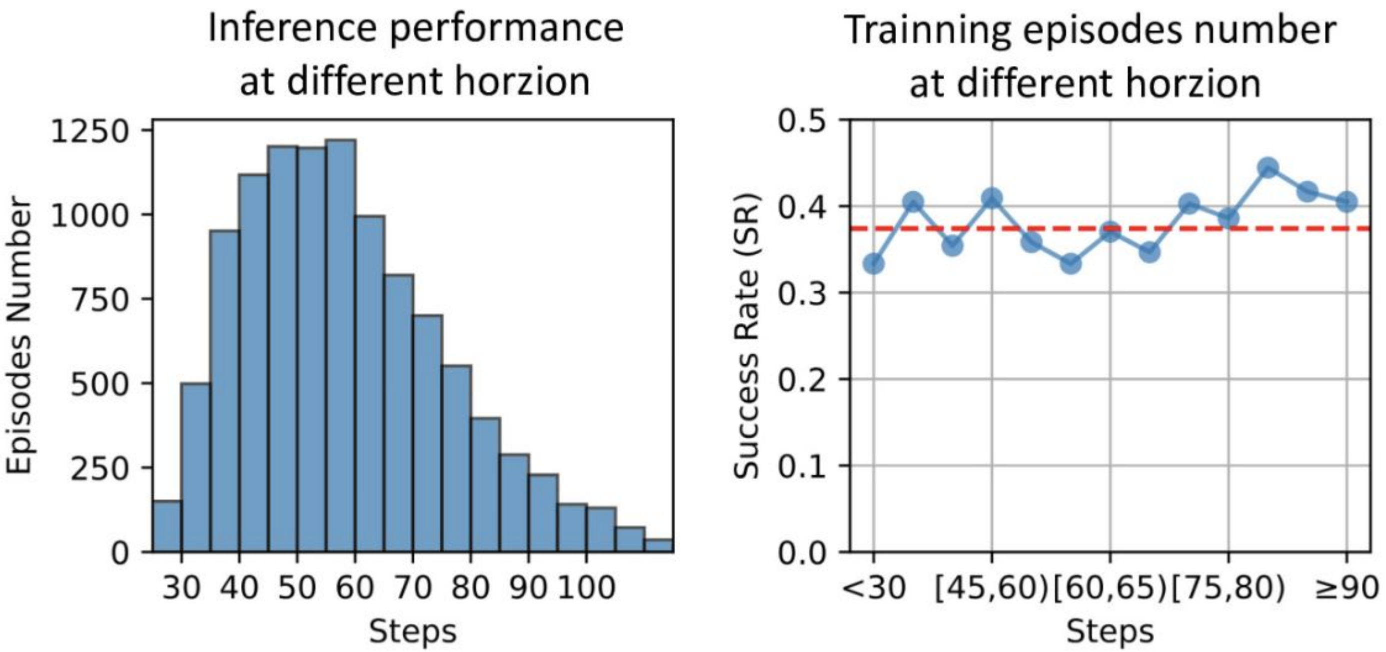
Our method consistently demonstrates state-of-the-art (SOTA) performance, significantly surpassing baseline metrics.

Experiments



↑SR (success rate)
↑OS (oracle success rate)
↑SPL (success weighted by path length)
↓NE (Navigation error)

Experiments

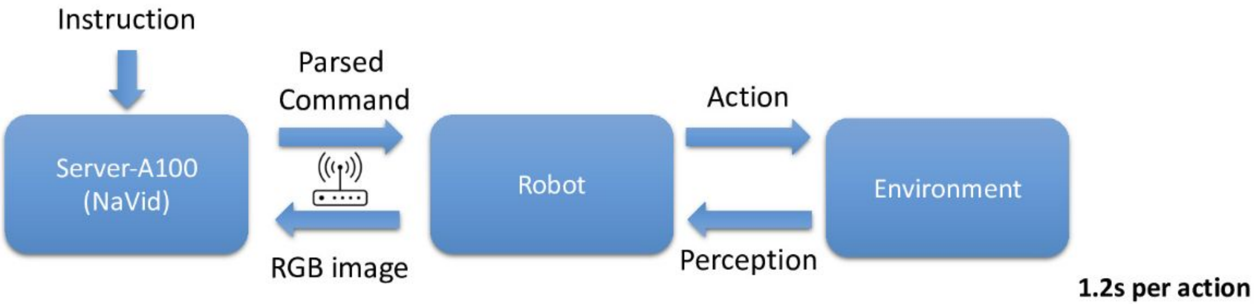


The 'Steps' of the x-axis indicate the oracle actions required by the instructions.

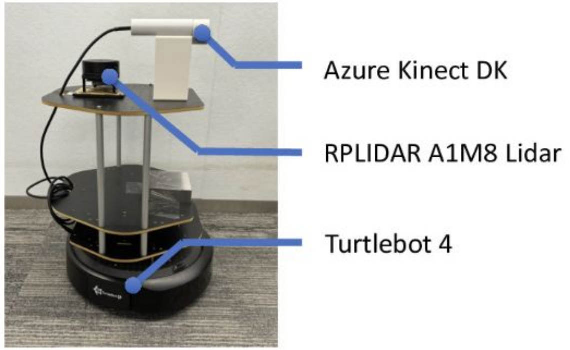
Experiments

Real-World Experiments

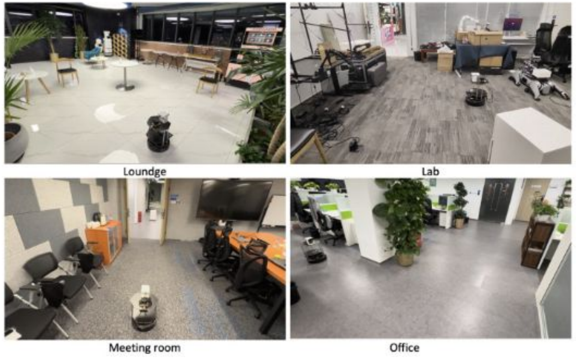
Pipeline



Robot



Environment



Experiments

R2R train -> Real world (Sim-to-real)

	Meeting Room				Office				Lab				Lounge			
	Simple I.F.		Complex I.F.		Simple I.F.		Complex I.F.		Simple I.F.		Complex I.F.		Simple I.F.		Complex I.F.	
	SR↑	NE↓	SR↑	NE	SR↑	NE↓	SR↑	NE↓	SR↑	NE↓	SR↓	NE↓	SR↑	NE↓	SR↑	NE↓
Seq2Seq [42]	4%	4.45	0%	7.21	0%	4.28	0%	6.92	0%	4.58	0%	6.61	0%	5.95	0%	6.82
CMA [42]	0%	4.27	0%	7.30	8%	4.62	0%	5.71	4%	4.35	0%	5.67	0%	4.63	0%	5.46
WS-MGMap [15]	52%	1.18	24%	2.20	60%	0.96	20%	2.94	44%	1.85	12%	3.18	48%	1.66	32%	2.88
Ours	**92%**	**0.55**	**56%**	**0.98**	**84%**	**0.63**	**48%**	**0.71**	**76%**	**0.83**	**40%**	**1.89**	**88%**	**0.72**	**44%**	**1.37**

Example:

Experiments

Experiments

Go straight to the wall, then turn left and walk to the door, then stop.

Walk towards the chair then turn right, and move to the door, then stop.

Go straight and move close to the plant, then turn right facing the door, then walk to the door and stop.

Conclusion & Limitations

Takeaway Messages

➢ NaVid navigates in a human-like manner, requiring solely an on-the-fly video stream from a monocular camera as input, without the need for maps, odometers, or depth inputs.

➢ We collect 510K VLN video sequences from simulation environments and 763K real-world caption samples to achieve cross-scene generalization.

➢ With more high-quality data and a better architecture, video-based VLM could be a promising pathway to achieve VLN.

CS586 (25 Spring) : Paper Presentation

NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation

Thank you.

20244076 Zhaoyan Wang

20244050 Xiangchen Liu

KAIST