ROBOTICS
SCIENCE AND SYSTEMS
JULY 15–19, 2024 | DELFT UNIVERSITY OF TECHNOLOGY, NETHERLANDS

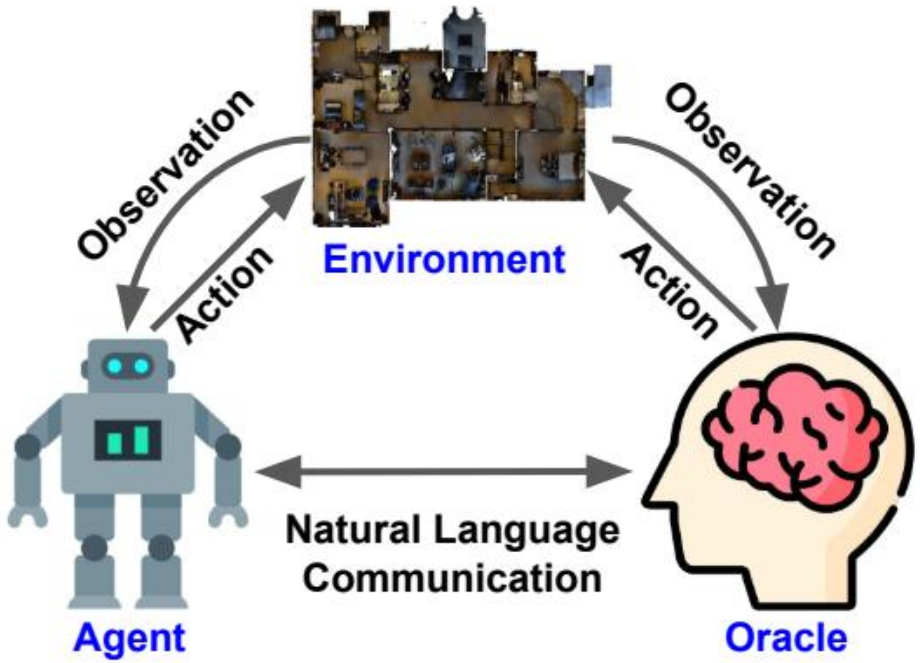
NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation

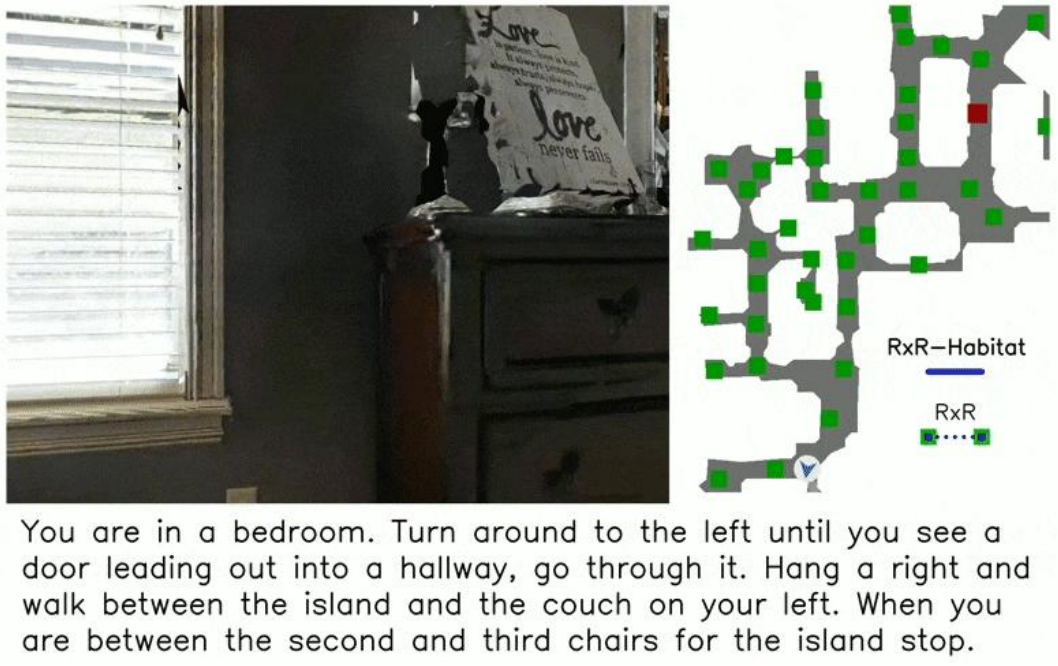
Jiazhao Zhang*, **Kunyu Wang***, **Rongtao Xu***, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang†, He Wang†

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''





[1] Gu, Jing et al. "Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions." *Annual Meeting of the Association for Computational Linguistics* (2022).

Key Insight 1 – VLM-Driven Real-World VLN

- Leverage the power of **foundational VLMs** to extend VLN to **real-world** applications, using **pretrained** large model, and **co-tuning** with **web-based data**.

Walk forward until you reach a white pail and stop.

Vision Language Model

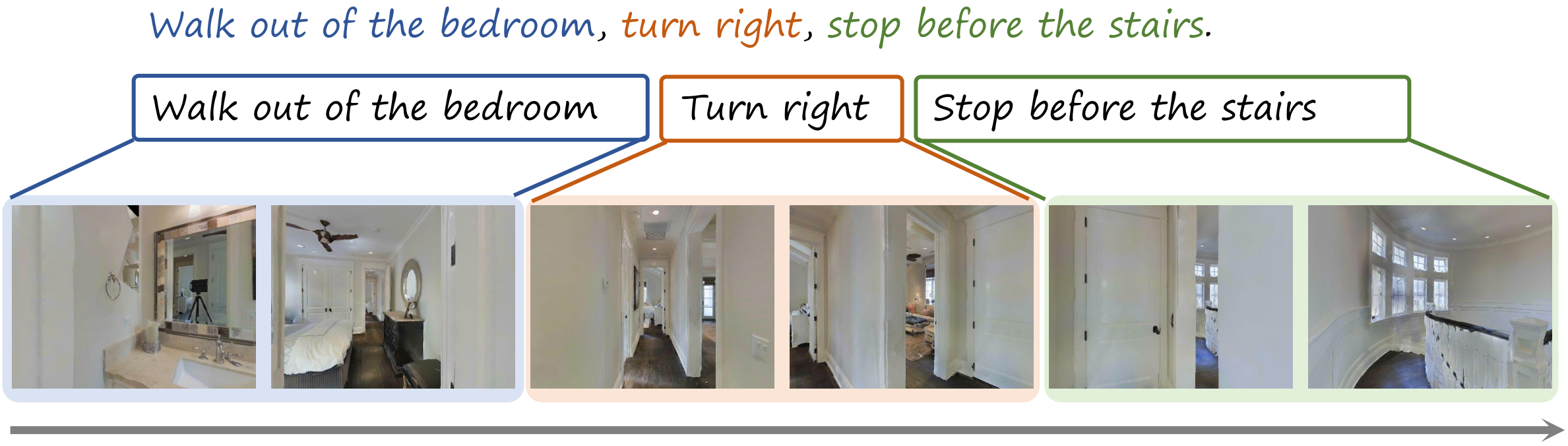
Actions

- Move forward
- Turn left
- Turn right
- Stop

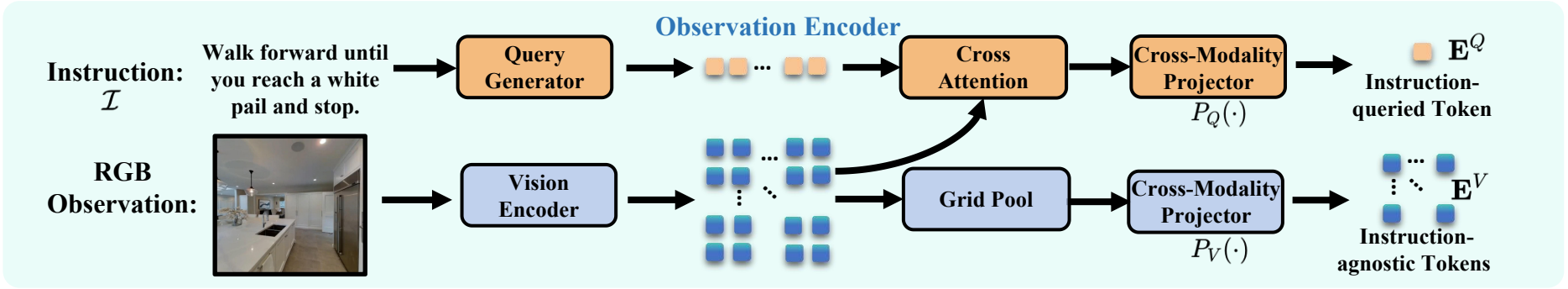
510k Navigation Data + 763k Web-based Data → Total 1.2M training data

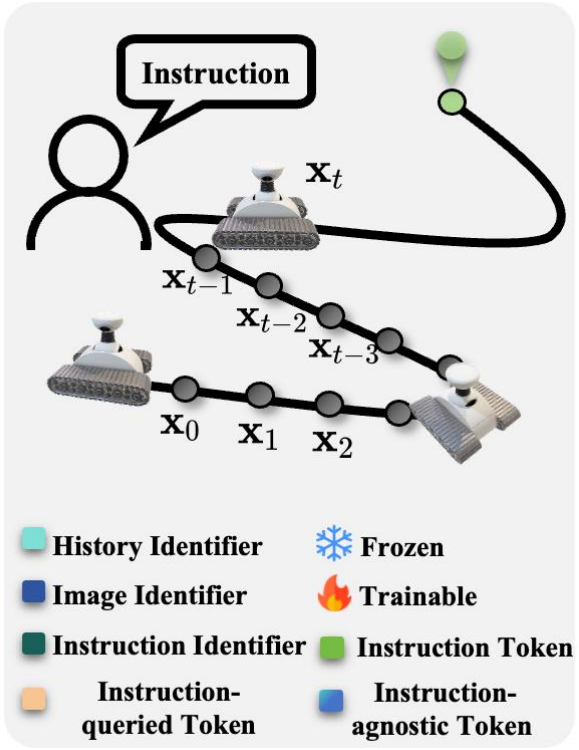
Key Insight 2 – Video-Based VLN Agent

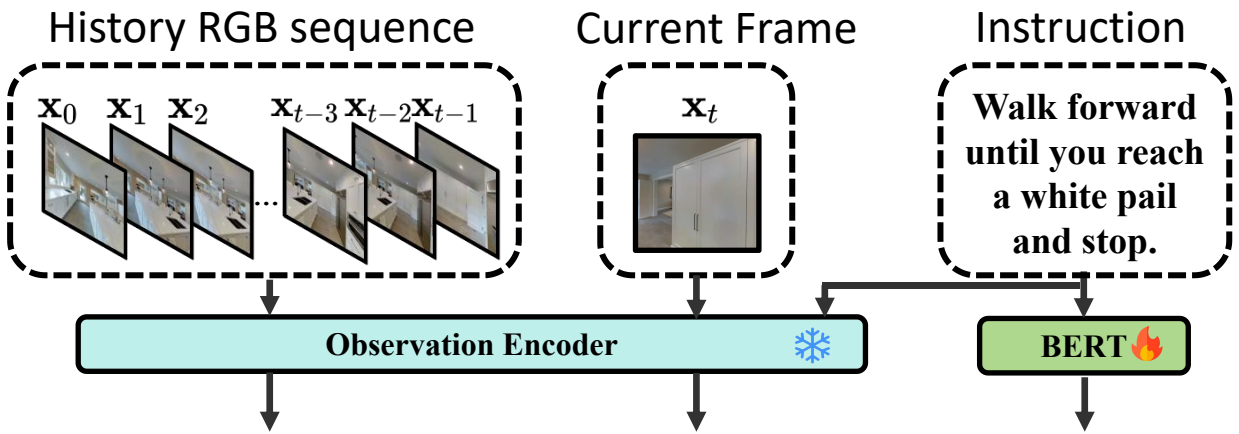
- Navigate in a **human-like** manner, relying **solely** on real-time **video streams** from a monocular camera, **without the need for maps, odometers, or depth inputs**.



Pipeline







[1] Li, Yanwei et al. “LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models.” ArXiv abs/2311.17043 (2023)

Text:

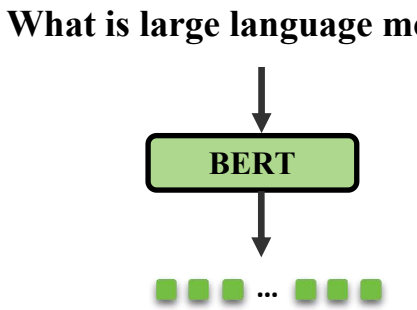
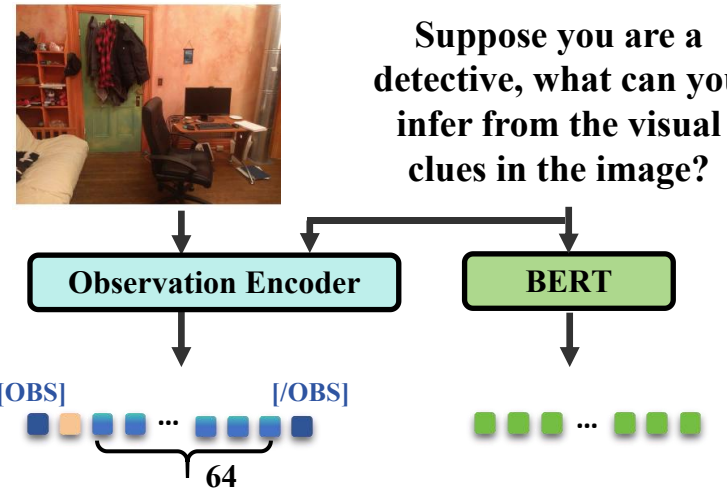
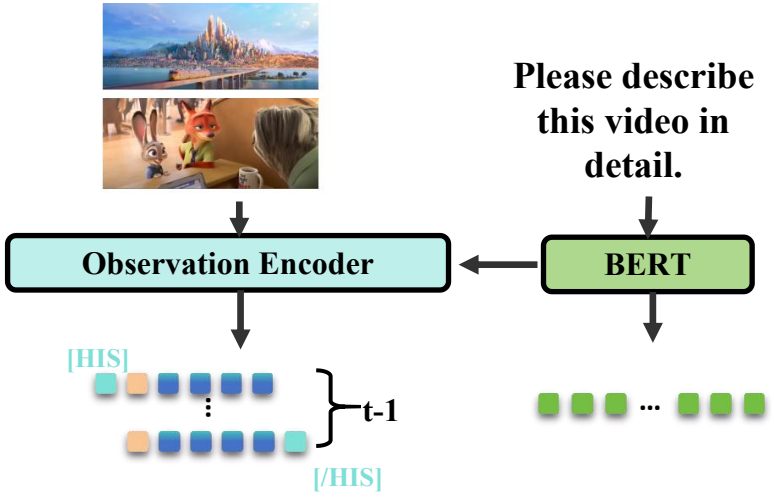


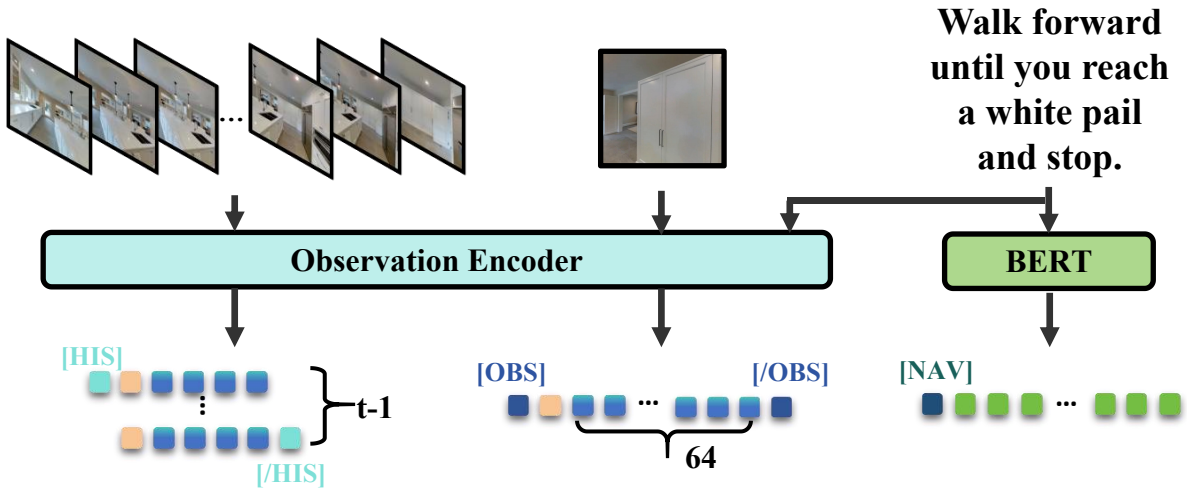
Image:



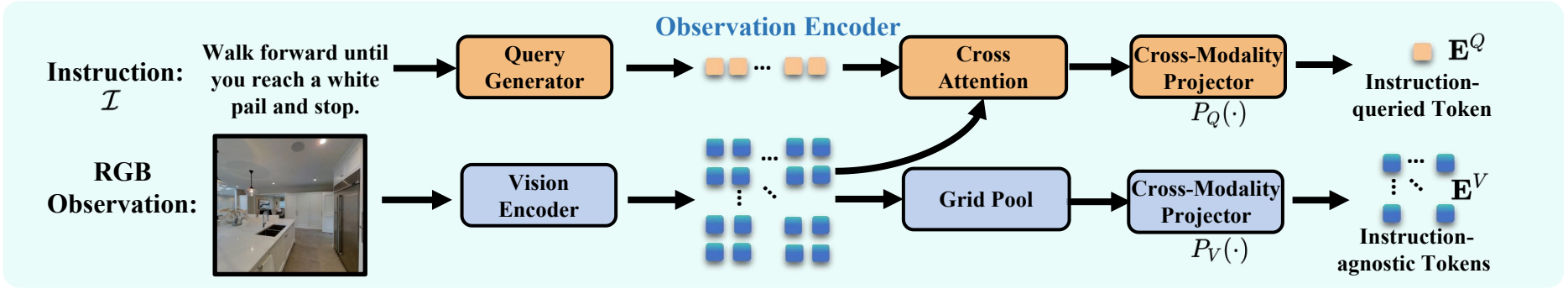
Video:

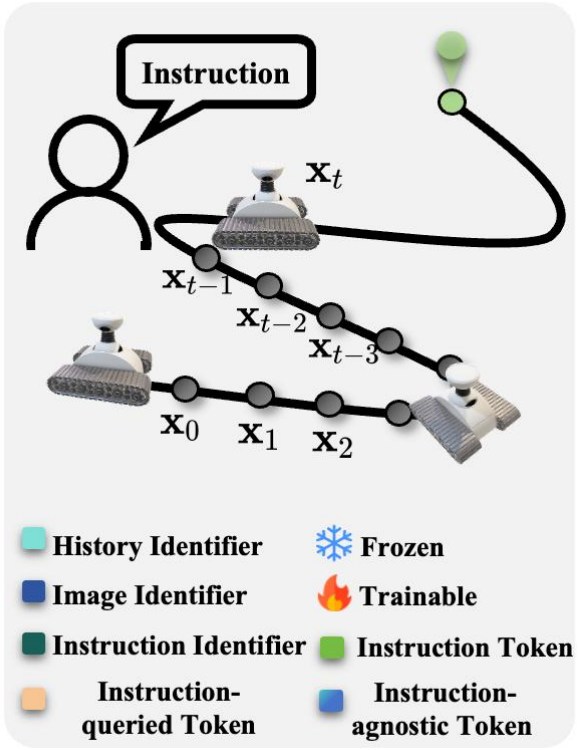


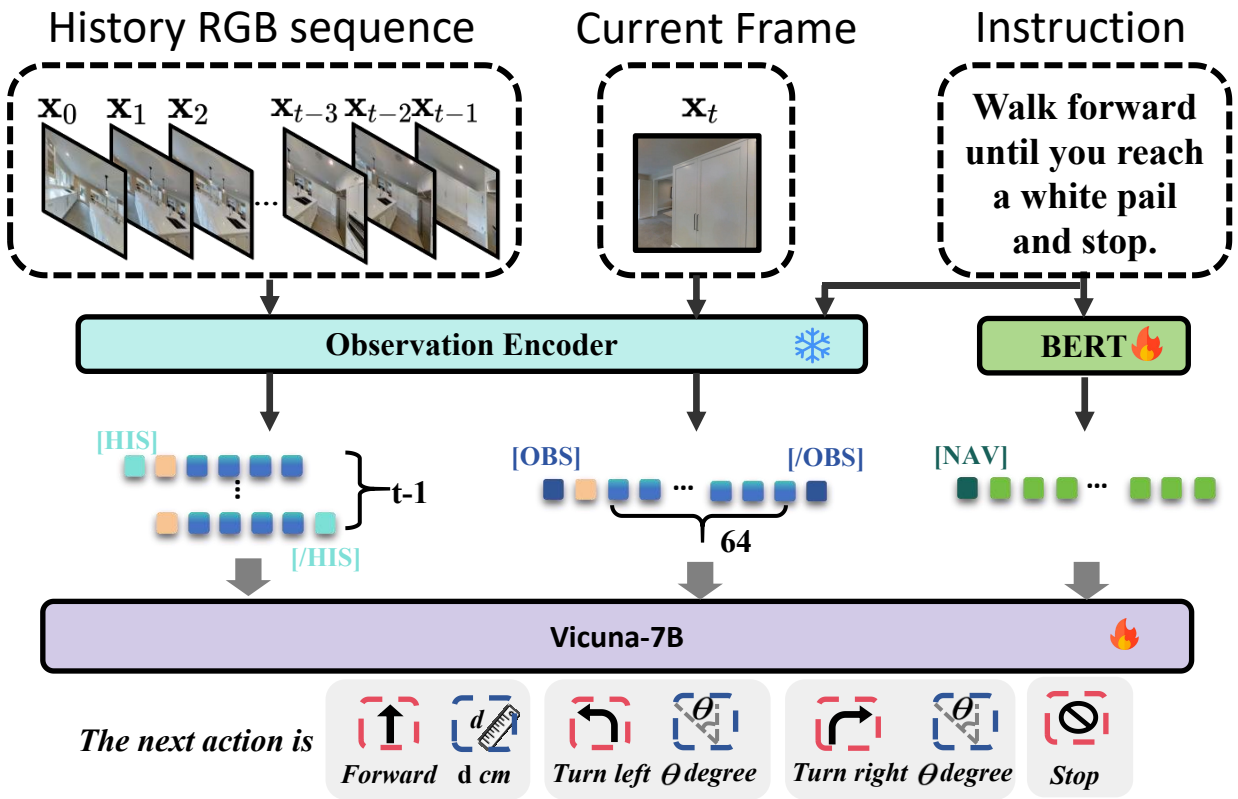
Navigation:



Pipeline

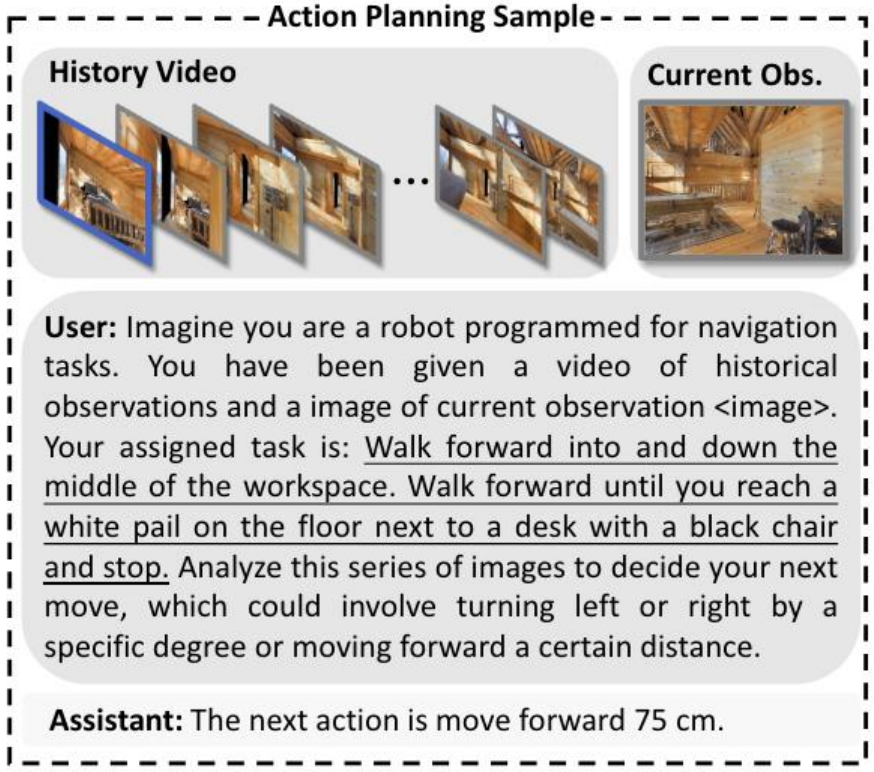

Observation Encoder
Instruction:
$\mathcal{I}$
Walk forward until you reach a white pail and stop.
Query Generator
Cross Attention
Cross-Modality Projector
$P_Q(\cdot)$
$\mathbf{E}^Q$
Instruction-queried Token
RGB Observation:
Vision Encoder
Grid Pool
Cross-Modality Projector
$P_V(\cdot)$
$\mathbf{E}^V$
Instruction-agnostic Tokens


Instruction
$\mathbf{x}_t$
$\mathbf{x}_{t-1}$
$\mathbf{x}_{t-2}$
$\mathbf{x}_{t-3}$
$\mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{x}_2$
History Identifier
Image Identifier
Instruction Identifier
Instruction-queried Token
Frozen
Trainable
Instruction Token
Instruction-agnostic Token

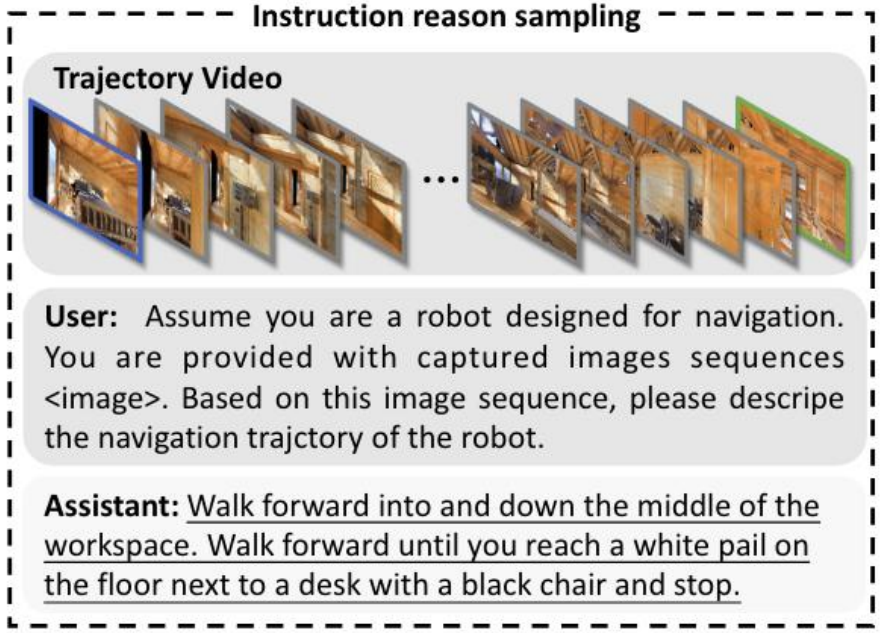

History RGB sequence
Current Frame
Instruction
$\mathbf{x}_0$ $\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_{t-3}$ $\mathbf{x}_{t-2}$ $\mathbf{x}_{t-1}$
$\mathbf{x}_t$
Walk forward until you reach a white pail and stop.
Observation Encoder
BERT
[HIS]
t-1
[/HIS]
[OBS]
[/OBS]
64
[NAV]
Vicuna-7B
The next action is
Forward d cm
Turn left θ degree
Turn right θ degree
Stop

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)

[1] Krantz, Jacob et al. “Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments.” European Conference on Computer Vision (2020).

Data collection

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

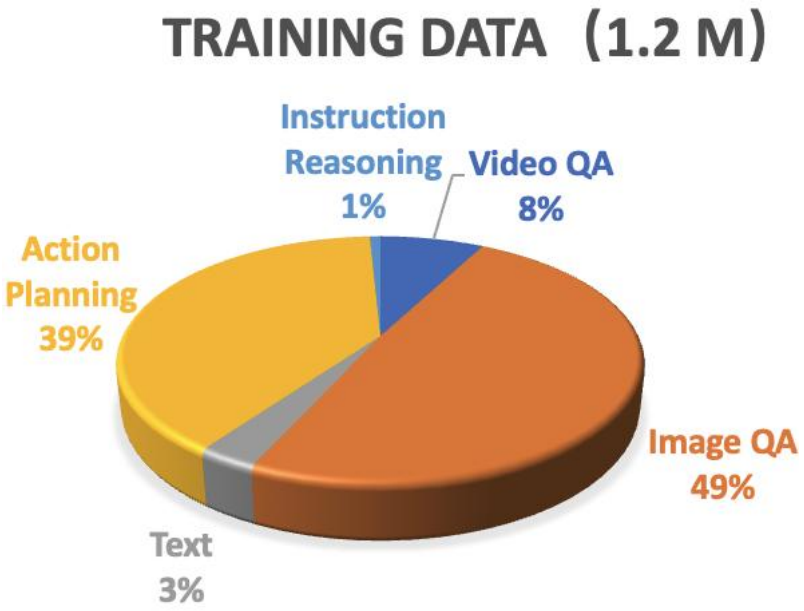
320k Action Planning Samples
10k Instruction Reasoning Sample

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

DAGGER

180k Action Planning Samples

(Total 1.2M training samples)



[1] Ross, Stéphane et al. "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning." International Conference on Artificial Intelligence and Statistics (2010).

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

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 [14]	✓	✓	✓	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 SOTA performance, significantly surpassing baseline metrics.

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 [13]	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**

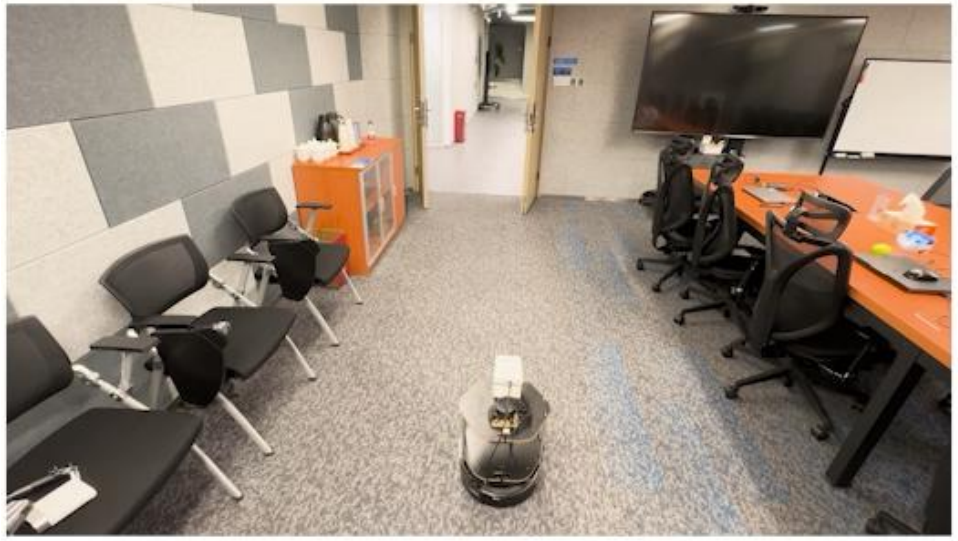
Example:

Real-world experiment – Simple Instruction

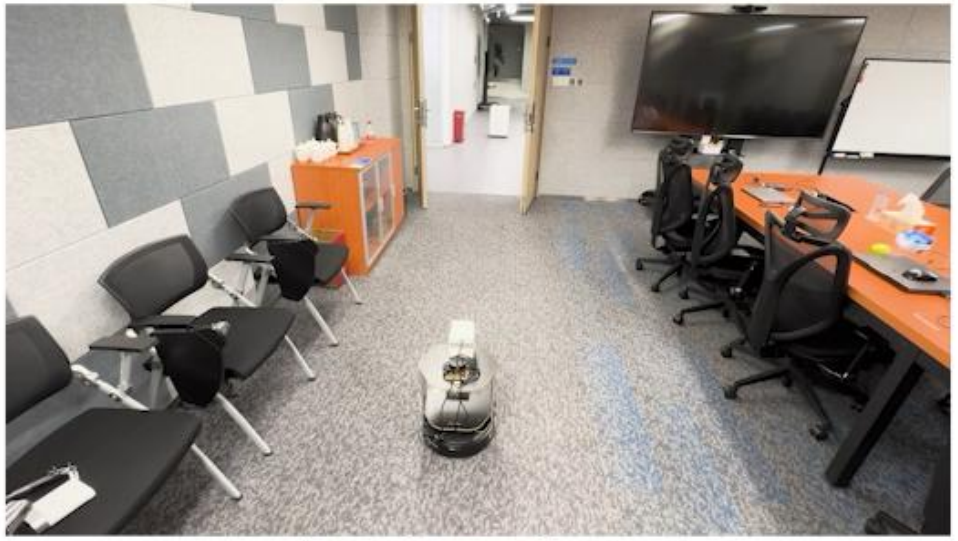
Simple Instruction following

Speed up x10

Walk towards the door then stop.



Walk towards the white box then stop.

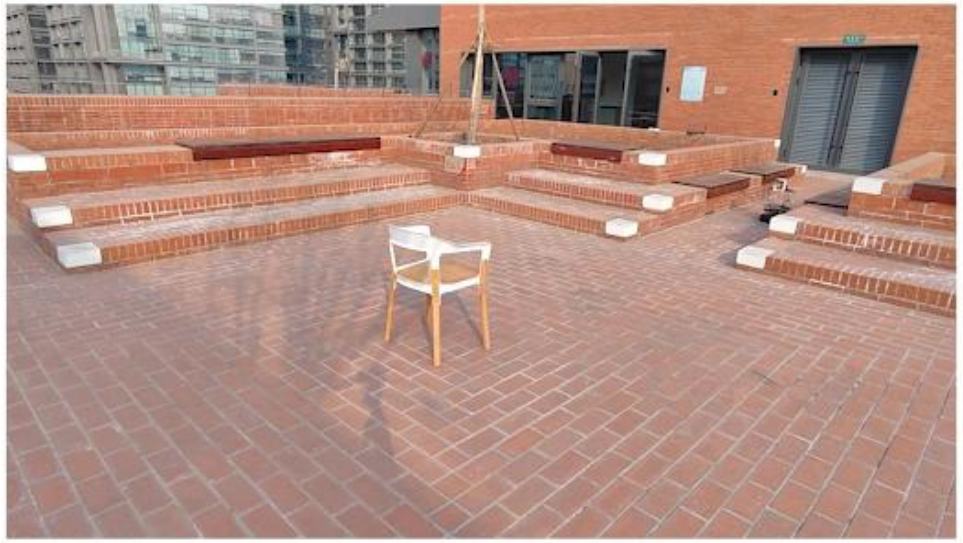


Real-world experiment - Outdoor Scenes

Simple Instruction following

Speed up x10

Walk forward to the chair then turn right, and stop at the stairs.



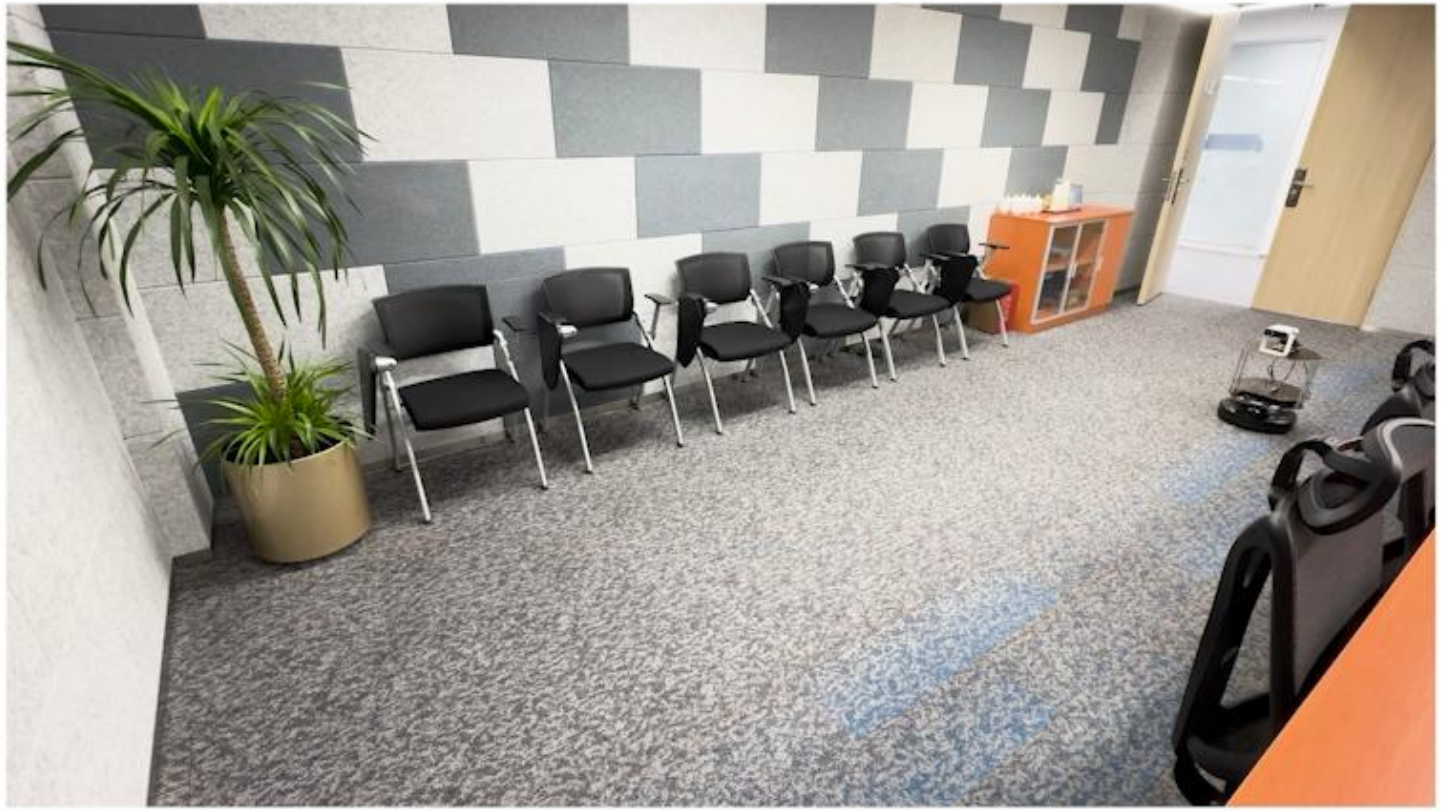
Walk forward to the chair then turn left and stop at the stairs.

Real-world experiment – Complex Instruction

Complex Instruction following

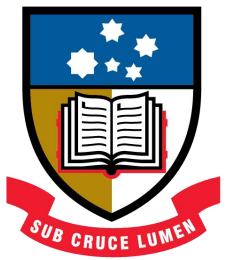
Speed up x10

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



Thanks for your attention

Project page: https://pku-epic.github.io/NaVid/



Contact me at zhngjizh@gmail.com, hewang@pku.edu.cn.