Towards Learning a Generalist Model for Embodied Navigation

Duo Zheng[1,2], Shijia Huang[1], Lin Zhao[3], Yiwu Zhong[1], Liwei Wang[1]

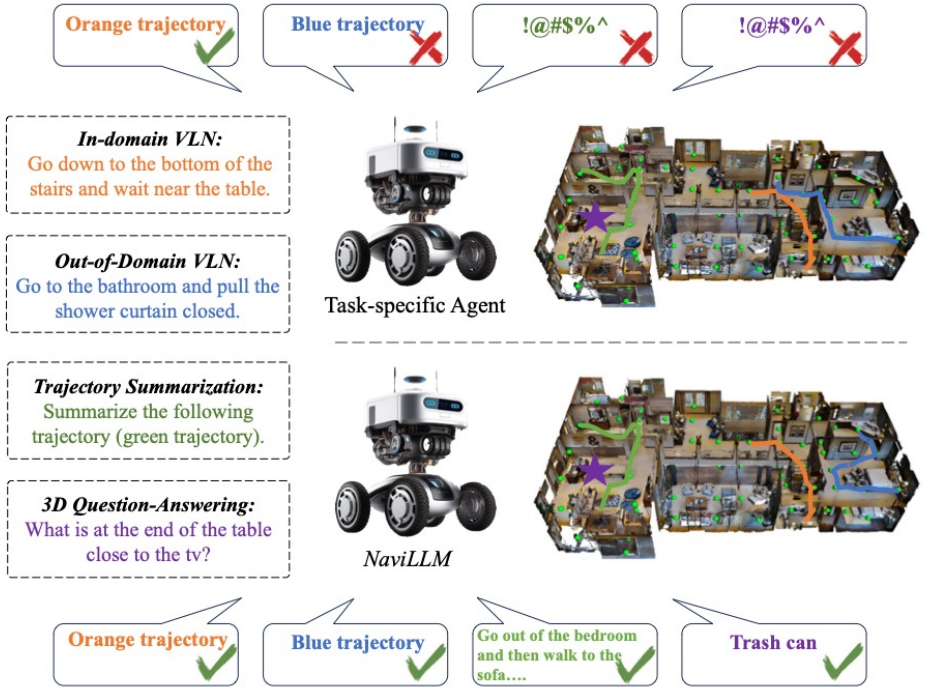
[1]The Chinese University of Hong Kong [2]Shanghai AI Laboratory
[3]Centre for Perceptual and Interactive Intelligence

Outline

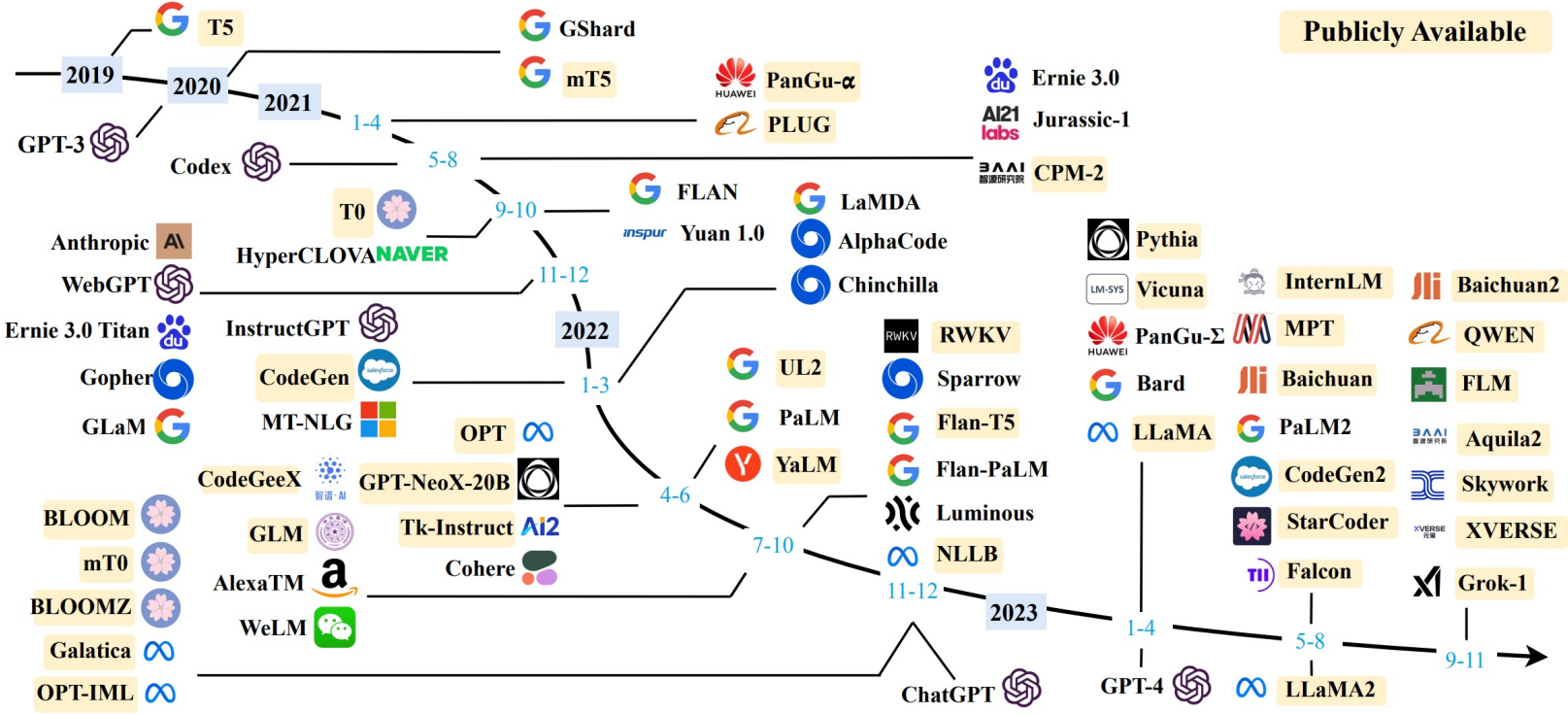
- Background
- Method
- Experimental Results
- Conclusion

Embodied Navigation

- An agent located in 3D environment is required to navigate according to various forms of instructions and provide textual responses based on user queries.



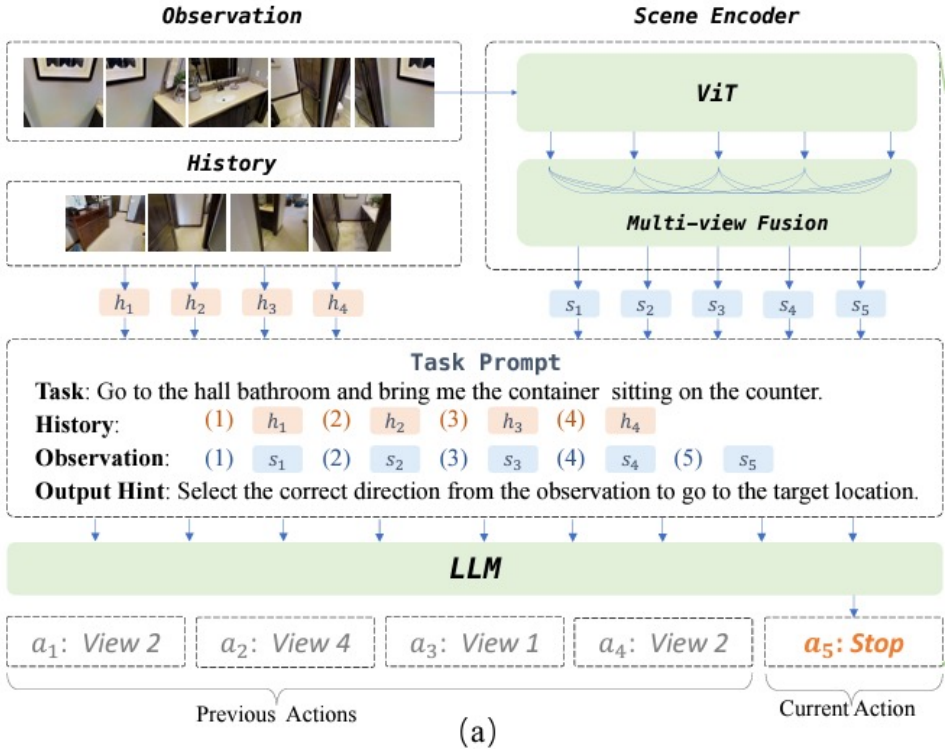
The Era of Large Language Models



Reference: A Survey of Large Language Models, arXiv 2023

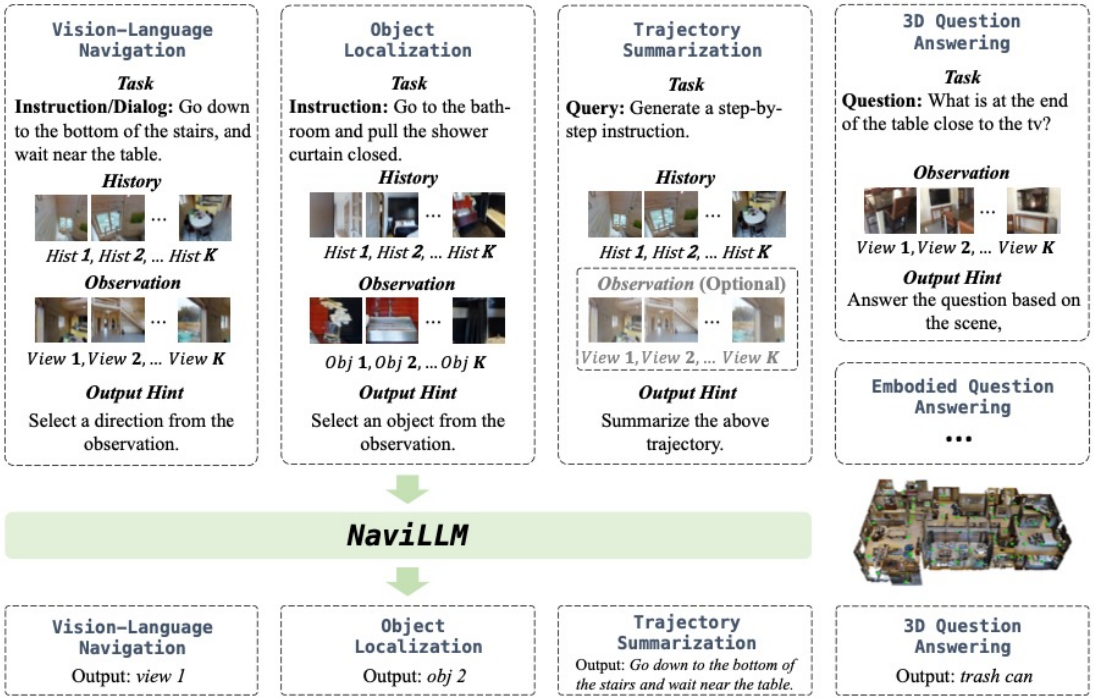
NaviLLM

- NaviLLM is an embodied model grounded in LLM, comprising two modules, i.e., a scene encoder and an LLM.



Schema-based Instruction

- The schema is designed to be a unified format that can adapt to different data sources and enable flexibility for wide span of tasks.



Comparison with SoTA Methods

	CVDN		SOON		R2R		REVERIE		ScanQA	
	Val-U	Test	Val-U	Test	Val-U	Test	Val-U	Test	Val	Test
Separate Model For Each Task										
PREVALENT [25]	3.15	2.44	-	-	53	51	-	-	-	-
HOP [44]	4.41	3.24	-		57	59	26.11	24.34	-	-
HAMT [11]	5.13	5.58	-	-	61	60	30.20	26.67	-	-
VLN-BERT [27]	-	-	-	-	57	57	24.90	23.99	-	-
GBE [61]	-	-	13.34	9.23	-	-	-	-	-	-
DUET [12]	-	-	22.58	21.42	60	58	33.73	36.06	-	-
Meta-Explore [31]	-	-	-	25.80	62	**61**	34.03	-	-	-
AZHP [23]	-	-	-	-	61	60	**36.63**	35.85	-	-
VLN-SIG [32]	5.52	5.83	-	-	62	60	-	-	-	-
VLN-PETL [45]	5.69	6.13	-	-	60	58	27.67	26.73	-	-
BEV-BERT [1]					**64**	60	36.37	**36.41**	-	-
3D-LLM [28]	-	-	-	-	-	-	-	-	20.5	19.1
Unified Model For All Tasks										
MT-RCM+Env [54]	4.65	3.91	-	-	49	40	-	-	-	-
NaviLLM	**6.16**	**7.90**	**29.24**	**26.26**	59	60	35.68	32.33	**23.0**	**26.3**

- ❖ Our method delivers SoTA results with a single model.
- ❖ Significant improvement on CVDN can be credited to the utilization of LLM and multi-task training.
- ❖ Better Performance on tasks with complex instructions.

Generalization Ability on Unseen Tasks

	CVDN		SOON				REVERIE			
	TL	**GP↑**	TL	**OSR↑**	**SR↑**	**SPL↑**	TL	**OSR↑**	**SR↑**	**SPL↑**
DUET (R2R)	21.12	3.38	26.83	7.64	4.66	2.84	7.88	29.11	24.91	20.00
DUET (REVERIE)	76.13	3.30	33.72	20.86	10.24	6.06	-	-	-	-
DUET (SOON)	48.61	2.40	-	-	-	-	38.10	43.45	10.91	3.64
NaviLLM	26.37	**4.46**	28.66	**33.11**	**19.81**	**14.29**	18.96	**51.47**	**28.10**	**21.04**

- Our method generalizes to out-of-domain VLN tasks.
- *NaviLLM* can combine the learned navigation and question-answering ability to solve more complex tasks, e.g., MP3D-EQA.

Conclusion

- We propose the first generalist model for embodied navigation, enabling a wide spectrum of capabilities required for embodied navigation.
- We unify various tasks in a single model by adapting LLM and introducing schema-based instruction.
- Our single model achieves SoTA results on CVDN, SOON, and ScanQA, with a significant margin of 29% compared to the previous SoTA on CVDN. Furthermore, it also exhibits strong generalizability on unseen tasks.