One Step at a Time: Long-Horizon Vision-and-Language Navigation with Milestones

Chan Hee (Luke) Song[1], Jihyung Kil[1], Tai-Yu Pan[1], Brian M. Sadler[2], Wei-Lun Chao[1], Yu Su[1]
The Ohio State University[1], Army Research Laboratory[2]

Highlights

- Propose ***M-Track***, a model-agnostic milestone-based task tracker that guides the agent and monitor its progress during long-horizon Vision-and-Language (VLN) task
- ***M-Track*** leads to a notable 45% and 70% relative improvement in unseen success rate over two competitive base models

Introduction

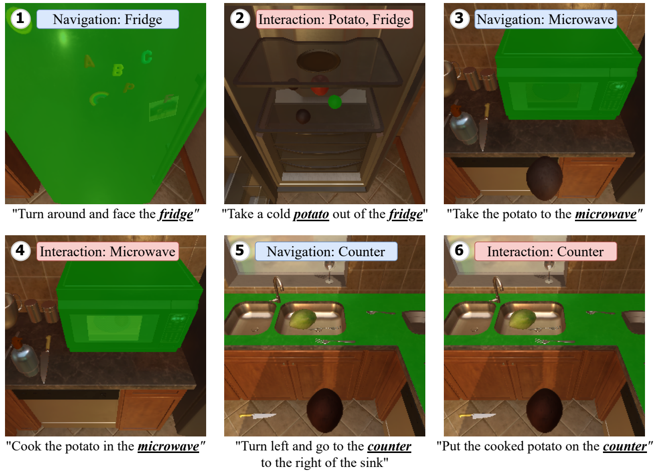
- VLN Challenges
  - An agent must comprehend the language instruction, ground it into the partially-observable environment with only visual perception, and plan and perform navigation and interaction actions in the environment to complete the task.
  - Significant challenge arises when the task horizon becomes substantially longer. That is, a task is so complex that it essentially consists of multiple “subtasks ”that need to be completed sequentially to fulfill the whole task.

M-Track

- Equips VLN agents with an explicit task tracker, which keeps track of the agent’s progress within a subtask and guides it for when to move on to the next.
- ***Milestone builder*** extracts the milestone (i.e., the necessary completion condition) of each subtask from the corresponding language instruction.
- ***Milestone checker*** tries to ground (i.e., identify and localize) the extracted target objects in the perceived environment using an object detection model and checks if the agent is close enough to them and/or is about to interact with them — to decide if the agent is completing the current subtask and ready to move on.

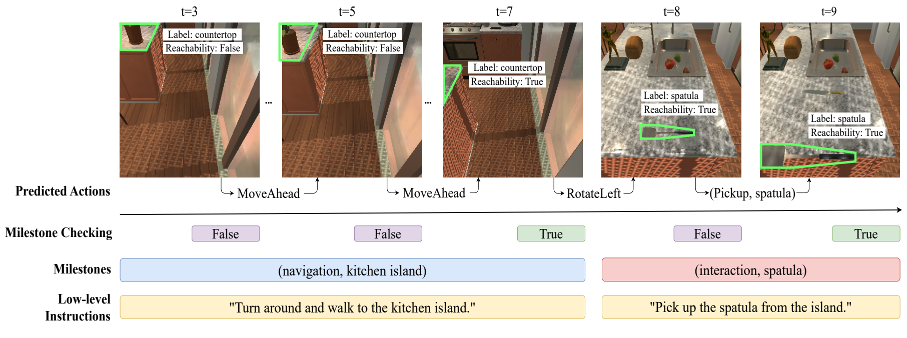
Illustration of our *M-Track* Approach

Goal: "Put a hot potato on the counter to the right of the sink"


1 Navigation: Fridge
2 Interaction: Potato, Fridge
3 Navigation: Microwave
"Turn around and face the ***fridge***"
"Take a cold ***potato*** out of the ***fridge***"
"Take the potato to the ***microwave***"
4 Interaction: Microwave
5 Navigation: Counter
6 Interaction: Counter
"Cook the potato in the ***microwave***"
"Turn left and go to the ***counter*** to the right of the sink"
"Put the cooked potato on the ***counter***"

M-Track Pipeline

- Milestones are extracted from the current low-level instruction by milestone builder
- After an action is predicted, milestone checker examines, based on objects with reachability information (text in images) from its object detector, if the resultant state satisfies the milestone.



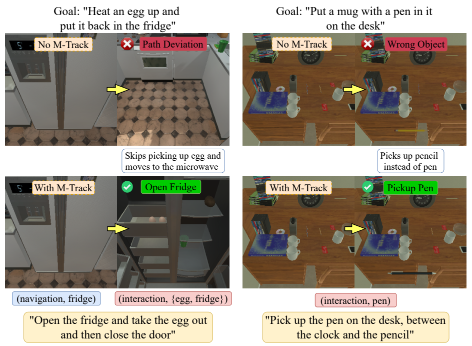
Experiments

- ***M-Track*** **– Main results**

Model	Test Unseen			Test Seen		
	SR	PLWSR	GC	SR	PLWSR	GC
MOCA [31]	5.30	2.72	14.28	22.05	15.10	28.29
LAV [26]	6.38	3.12	17.27	13.35	6.31	23.21
EmBERT [33]	7.52	3.58	16.33	31.77	23.41	39.27
E.T. [27]	8.57	4.10	18.56	38.42	**27.78**	45.44
LWIT [25]	9.42	5.60	20.91	30.92	25.90	40.53
HiTUT [40]	13.87	**5.86**	20.31	21.27	11.10	29.97
ABP [15]	15.43	1.08	24.76	**44.55**	3.88	**51.13**
HLSM [2]	**16.29**	4.34	27.24	25.11	6.69	35.79
LSTM-L	7.48	1.31	11.39	14.12	5.34	20.61
LSTM-L + M-TRACK	12.68	3.88	19.45	20.13	8.78	26.34
VLN⟳BERT-L	10.21	3.64	20.57	22.99	8.10	30.87
VLN⟳BERT-L + M-TRACK	14.79	4.97	**29.25**	26.21	9.82	38.19

- ***M-Track*** **– Ablation study**

Model	Component						Val Unseen		Val Seen	
	RL	ALFRED-OD	Binary	Passive	Proactive	GT	SR	GC	SR	GC
LSTM							1.82	3.09	9.26	11.09
LSTM-L	✓						7.80	10.01	12.92	18.22
	✓	✓					9.26	13.34	15.61	20.94
	✓	✓	✓				11.70	15.39	18.41	24.87
	✓	✓		✓			14.87	20.21	21.45	28.65
LSTM-L + M-TRACK	✓	✓			✓		**15.73**	**21.59**	**22.31**	**29.64**
	✓	✓				✓	20.36	30.79	25.12	31.41
VLN⟳BERT							3.66	7.19	14.51	20.11
VLN⟳BERT-L	✓						9.34	19.45	17.68	27.71
	✓	✓					10.09	21.33	20.48	29.03
	✓	✓	✓				13.90	24.91	24.26	32.99
	✓	✓		✓			15.12	26.45	26.34	35.12
VLN⟳BERT-L + M-TRACK	✓	✓			✓		**16.34**	**32.06**	**27.68**	**38.15**
	✓	✓				✓	24.38	39.34	31.95	46.27


Goal: "Heat an egg up and put it back in the fridge"
Goal: "Put a mug with a pen in it on the desk"
No M-Track
Path Deviation
No M-Track
Wrong Object
Skips picking up egg and moves to the microwave
Picks up pencil instead of pen
With M-Track
Open Fridge
With M-Track
Pickup Pen
(navigation, fridge)
(interaction, {egg, fridge})
(interaction, pen)
"Open the fridge and take the egg out and then close the door"
"Pick up the pen on the desk, between the clock and the pencil"

- M-Track keeps the agent on track to not skip the current subgoal (left)
- M-Track with proactive checking ensures that the agent interact with the correct object (right)