Towards Learning a Generalist Model for Embodied Navigation

Supplementary Material

1. Dataset Statistics

We provide training data statistics in Table 3.

CVDN. In the CVDN dataset, the training section includes 4,742 questions spanned across 57 different environments. The unseen validation subset comprises 907 questions located in 10 unique environments. Meanwhile, the test subset consists of 1,384 questions over 16 environments.

SOON There are 2,780 instructions over 34 houses in the training set, while the unseen validation set hosts 339 instructions from 5 houses. The test set, on the other hand, contains 1,411 instructions derived from 14 houses.

R2R is comprised of a training set with 14,039 instructions drawn from 61 unique houses, a validation unseen set (val-unseen) containing 2,349 instructions from 11 houses, and a test set which includes 4,173 instructions from 18 houses.

REVERIE. For the REVERIE dataset, the training set encompasses 10,466 instructions across 60 houses, the unseen validation set (val-unseen) consists of 3,521 instructions from 10 houses, and the test set includes instructions from 16 houses, totaling 6,292.

ScanQA. The training set holds 25,563 questions spread across 562 scenes, the validation set consists of 4,675 questions within 71 scenes, and the test set without objects (test w/o objects) includes 6,179 questions from 97 scenes.

EQA. For EQA, we filter out the data in the validation set with inaccurate endpoints, retaining 849 entries for testing.

2. Hyperparameter Details

We follow previous works [11, 31] to set up the hyperparameters, i.e., the view number n is set to 36 for all tasks, and the history round t is set to 15, 15, 30, 20 for R2R, REVERIE, CVDN and SOON, respectively.

3. Ablation for Schema Elements

As shown in Table 1, we observe a significant drop when removing these elements. This reveals the effectiveness of the proposed schemas in distinguishing different tasks.

4. Detailed Comparison on Benchmarks

We provide the detailed comparison of CVDN, SOON, R2R, and REVERIE in Table 2, 4, 5, and 6, respectively. We present the results of the top 5 teams on the ScanQA leaderboard up to Nov. 2023 in Table 7, and our method wins second place on the leaderboard.

O	T	CVDN	SOON	R2R	REVERIE	ScanQA	Sum↑
-	-	3.51	20.68	37.51	23.93	21.86	107.49
✓	-	5.07	24.52	50.29	28.27	21.50	129.65
✓	✓	5.40	26.32	55.07	30.56	23.10	140.45

Table 1. The ablation of schemas. We train all models for 2,500 steps and report the results on validation unseen sets. O: output hint. T: task hint.

	Val-Unseen	Test
Seq2Seq [52]	2.10	2.35
PREVALENT [25]	3.15	2.44
HOP [44]	4.41	3.31
MT-RCM [54]	4.36	-
MT-RCM+Env [54]	4.65	3.91
HAMT [11]	5.13	5.58
VLN-SIG [32]	5.52	5.83
VLN-PETL [45]	5.69	6.13
NaviLLM (Ours)	6.16	7.90

Table 2. Detailed comparison with state-of-the-art methods on CVDN.

5. Instruction Templates

Based on the proposed schema-based instruction, we present the instruction templates as shown in Table 8.

	CVDN	SOON	R2R	REVERIE	ScanQA	LLaVA	R2R†	REVERIE†
scene	57	34	61	60	562	-	60	60
inst.	4.7k	27.8k	14k	10.5k	25.6k	23k	1070k	20k

Table 3. Training data statistics. **scene** and **inst.** indicate the number of 3D scenes and instances, respectively. † means augmentation datasets provided by [11].

	Val-Unseen				Test Unseen			
	TL	**OSR↑**	**SR↑**	**SPL↑**	TL	OSR↑	SR↑	SPL↑
GBE [61]	28.96	28.54	19.52	13.34	21.45	28.96	19.52	13.34
DUET [12]	36.20	50.91	36.28	22.58	41.83	43.00	33.44	21.42
AZHP [23]	39.33	56.19	40.71	26.58	-	-	-	
Meta-Explore [31]	-	-	-	-	-	48.7	39.1	25.8
NaviLLM (Ours)	27.72	53.24	38.33	29.24	28.80	45.78	35.04	26.26

Table 4. Detailed comparison with state-of-the-art methods on the val-unseen split of the SOON dataset.

	Val Unseen				Test Unseen			
	TL	**NE↓**	**SR↑**	**SPL↑**	TL	**NE↓**	**SR↑**	**SPL↑**
PREVALENT [25]	10.19	4.71	58	53	10.51	5.30	54	51
HOP [44]	12.27	3.80	64	57	12.68	3.83	64	59
HAMT [11]	11.46	2.29	66	61	12.27	3.93	65	60
VLN-BERT [27]	12.01	3.93	63	57	12.35	4.09	63	57
DUET [12]	13.94	3.31	72	60	14.73	3.65	69	59
Meta-Explore [31]	13.09	3.22	72	62	14.25	3.57	71	61
AZHP [23]	14.05	3.15	72	61	14.95	3.52	71	60
VLN-SIG [32]	-	-	72	62	-	-	72	60
VLN-PETL [45]	11.52	3.53	65	60	12.30	4.10	63	58
NaviLLM (Ours)	12.81	3.51	67	59	13.21	3.71	68	60

Table 5. Detailed comparison with state-of-the-art methods on the R2R dataset. **NE** means navigation error.

	Val Unseen				Test			
	TL	**OSR↑**	**SR↑**	**SPL↑**	TL	**OSR↑**	**SR↑**	SPL↑
Seq2Seq [3]	11.07	8.07	4.20	2.84	10.89	6.88	3.99	3.09
HOP [44]	18.85	36.24	31.78	26.11	16.38	33.06	30.17	24.34
HAMT [11]	14.08	36.84	32.95	30.20	13.62	33.41	30.40	26.67
VLN-BERT [27]	16.78	35.02	30.67	24.90	15.68	32.91	29.61	23.99
DUET [12]	22.11	51.07	46.98	33.73	21.30	56.91	52.51	36.06
AZHP [23]	22.32	53.65	48.31	36.63	21.84	55.31	51.57	35.85
VLN-PETL [45]	14.47	37.03	31.81	27.67	14.00	36.06	30.83	26.73
NaviLLM	15.34	52.27	42.15	35.68	15.16	51.75	39.80	32.33
*NaviLLM*w/p pretrain	16.04	53.74	44.56	36.63	16.39	56.21	43.49	34.45

Table 6. Experimental results on REVERIE val-unseen and test sets.

Rank	Team	**EM↑**	**BLEU-1↑**	**BLEU-4↑**	**ROUGE↑**	**METEOR↑**	**CIDEr↑**
1	mare	30.82	34.41	17.75	41.18	15.60	79.35
2	*NaviLLM* (Ours)	24.77	38.72	12.49	37.95	15.55	74.90
3	bubble-bee	23.78	33.26	14.71	34.37	13.86	67.62
4	Optimus-prime	23.01	30.15	11.85	32.76	12.92	62.62
5	ML	21.37	32.69	11.73	32.41	13.28	62.8

Table 7. Top 5 Teams on the ScanQA Leaderboard as of Nov 2023

Dataset	Example
CVDN	Find the described room according the given dialog. Target: The goal room contains a fireplace. Question: where should I go? Answer: If you turn around and keep heading then there is an open door on your right. Question: okay where to? Answer: Go through the open doors on the right.
SOON	Find the described target. Target: I want to find a ceramic, rectangular and white sink, which is set in the washroom. The sink is under the mirror and next to the toilet. The washroom is inside the office, which is on the first floor and next to the living room.
R2R	Navigate following the instruction. walk towards the faucet sculpture, continue past it on left side and walk to the right of the marble cylinder to the doorway. Walk through the doorway to the right of the long rug, around the foot of the bed, and through the doorway on the left. Stop in front of the towel rack.
REVERIE	Go to the location to complete the given task. Task: Proceed to the office and turn on the ceiling fan.
R2R-Summ	Predict the fine-grained instruction based on your previous history and current location. Fine-grained instructions contain commands for each individual step.
SOON-Summ	Generate the target you want to find based on your previous history and current location. Describe both the target and its surroundings.
REVERIE-Summ	Generate the task you need to complete based on your previous history and current location.
ScanQA	Please answer questions based on the observation.
REVERIE-OG, SOON-OG	Select the target object from the candidate objects based on the instruction and history.

(a) Task

Dataset	Example
Any Datasets	Following is the History, which contains the visual information of your previous decisions.

(b) History

Dataset	Example
CVDN, SOON, R2R, REVERIE	Following is the Candidate, which contains several directions you can go to at the current position, candidate (0) is stop.
R2R-Summ, SOON-Summ REVERIE-Summ	Following is the Observation, which contains panoramic views at your current location.
REVERIE-OG, SOON-OG	Following is the Object, which contains several objects that you could see at the current viewpoint, option (0) indicates not exist.
ScanQA	The following is the Observation, which includes multiple images from different locations.

(c) Observation

Dataset	Example
CVDN	Understand the dialog in the Instruction and infer the current progress based on the History and dialog. Then select the correct direction from the candidates to go to the target location.
SOON	Nearby areas and objects can assist you in locating the desired room and object. Select the correct direction from the candidates to go to the target location.
R2R	Compare the History and Instruction to infer your current progress, and then select the correct direction from the candidates to go to the target location.
REVERIE	Explore the scene to find out the targeted room and object. Then select the correct direction from the candidates to go to the target location.
R2R-Summ	Please generate the step-by-step instruction.
SOON-Summ	Please predict both the target you want to find and its surroundings.
REVERIE-Summ	Please predict the task you need to complete.
REVERIE-OG, SOON-OG	Following is the Object, which contains several objects that you could see at the current viewpoint, option (0) indicates not exist.

(d) Output Hint

Table 8. All instruction templates used in our method. R2R-Summ, SOON-Summ, and REVERIE-Summ are the trajectory summarization datasets created from R2R, SOON, and REVERIE, respectively. REVERIE-OG and SOON-OD are employed for object localization.