LANA: A Language-Capable Navigator for Instruction Following and Generation
Supplementary Material

Xiaohan Wang Wenguan Wang Jiayi Shao Yi Yang
CCAI, Zhejiang University
https://github.com/wxh1996/LANA-VLN

This document provides more details of our approach and additional experimental results, organized as follows:

- §1 Implementation Details of LANA.
- §2 Additional Quantitative Results on REVERIE [7].
- §3 More Ablative Study on Route Encoder.
- §4 Additional Qualitative Results.
- §5 Discussion about Social Impact and Limitations.

1. Implementation Details of LANA

We employ an additional Instruction Trajectory Matching (ITM) task following previous efforts [2] during pretraining, which predicts whether a pair of instruction and trajectory is aligned. The three tasks IF (Instruction Following), IG (Instruction Generation) and ITM (Instruction Trajectory Matching) are sampled with a ratio IG:IF:ITM=4:1:2. We present the pseudo-code of the pretraining procedure in Algorithm 1 (ITM is omitted for simplicity). For finetuning, the instruction following task is optimized with Reinforcement Learning (RL) and Imitation Learning (IL). IL utilizes the same loss in Eq.13 while RL is implemented based on the Asynchronous Advantage Actor-Critic (A3C) algorithm [6]. During finetuning, the sampling ratio for IG and IF is set to IG:IF=2:5; the ITM task is abandoned. Following the common practice [2, 5, 8], we concatenate the object features with the panoramic features and add an object grounding loss for the instruction following task on REVERIE [7]. The detailed architecture of LANA is shown in Table A.

	Language Encoder $\mathcal{E}^l$	Route Encoder $\mathcal{E}^r$	Language Decoder $\mathcal{D}^l$	Route Decoder $\mathcal{D}^r$
Layer	[self_att, feedforward, -] ×9	[cross_att, self_att, feedforward] ×1	[cross_att, self_att, feedforward] ×4	[cross_att, self_att, feedforward] ×4

Table A: Detailed model architecture of LANA (§1).

Algorithm 1 The pseudo-code of pre-training for LANA.

Arguments: The labeled dataset $\mathcal{H}=\{(R,X)\}$, the maximum iteration N, Route Encoder $\mathcal{E}^r$, Language Encoder $\mathcal{E}^l$, Language Decoder, $\mathcal{D}^l$, and Route Decoder $\mathcal{D}^r$.

```
Initialize E^r, E^l, D^r, D^l
for iteration i ∈ [1, ..., N] do
    Sample batch B ⊂ H
    Sample a pretraining task T from {IG, IF}
    L ← 0
    if T is IG then
        for (R, X) ∈ B do
            [r̄_{1:T}] = E^r(R)
            [x̄_{1:l-1}] = E^l([x_{1:l-1}])
            q_l = D^l([x̄_{1:l-1}], [r̄_{1:T}])
            Estimate L^g                         ▷ Defined in Eq.12.
            L ← L + L^g
        Calculate ∂L
        Update E^r, E^l, D^l
    else if T is IF then
        for (R, X) ∈ B do
            [r̄_{1:t-1}, Ō_t] = E^r([r_{1:t-1}, O_t])
            [x̄_{1:L}] = E^l(X)
            p_t = D^r([r̄_{1:t-1}, Ō_t], [x̄_{1:L}])
            Estimate L^f                         ▷ Defined in Eq.13.
            L ← L + L^f
        Calculate ∂L
        Update E^r, E^l, D^r
return E^r, E^l, D^r, D^l
```

---

2. Additional Quantitative Results on REVERIE

The synthetic samples in the PREVALENT dataset are created with a speaker trained on R2R [1]. A recent work DUET [3] collected a new augmented dataset by synthesizing instructions with a speaker model trained on the REVERIE dataset [7]. We report additional quantitative results of LANA trained with this dataset in Table B. Remarkably, this training strategy boosts the performance by a large margin on REVERIE [7]. LANA achieves better navigation performance than DUET [3] with the same training set, demonstrating the algorithmic advantages of our approach.

Methods	REVERIE val unseen						REVERIE test unseen					
	SR↑	**SPL**↑	OR↑	TL↓	RGS↑	RGSPL↑	**SR**↑	**SPL**↑	OR↑	TL↓	RGS↑	RGSPL↑
RCM [9] [CVPR2019]	9.29	6.97	14.23	11.98	4.89	3.89	7.84	6.67	11.68	10.60	3.67	3.14
VLN↻BERT [5] [CVPR2021]	30.67	24.90	35.02	16.78	18.77	15.27	29.61	23.99	32.91	15.86	16.50	13.51
AirBERT [4] [ICCV2021]	27.89	21.88	34.51	18.71	18.23	14.18	30.28	23.61	34.20	17.91	16.83	13.28
HAMT [2] [NeurIPS2021]	32.95	30.20	36.84	14.08	18.92	17.28	30.40	26.67	33.41	13.62	14.88	13.08
HOP [8] [CVPR2022]	30.39	25.10	35.30	17.16	18.23	15.31	29.12	23.37	32.26	17.05	17.13	13.90
DUET† [3] [CVPR2022]	46.98	33.73	51.07	22.11	32.15	23.03	52.51	36.06	56.91	21.30	31.88	22.06
LANA (ours)	34.00	29.26	38.54	16.28	19.03	16.18	33.50	26.89	36.41	16.75	17.53	14.25
LANA† (ours)	**48.31**	**33.86**	**52.97**	23.18	**32.86**	22.77	51.72	**36.45**	**57.20**	18.83	**32.95**	**22.85**

Table B: Additional quantitative results for **instruction following** on REVERIE [7]. † indicate the model is trained on the DUET dataset [3]. See §2 for details.

#	Route Encoder		Instruction Following				Instruction Generation					
	self_att	cross_att	**SR**↑	**SPL**↑	OR↑	TL↓	**SPICE**↑	Bleu-1↑	Bleu-4↑	CIDEr↑	Meteor↑	Rouge↑
1	✔		65.6	60.1	72.7	11.7	0.202	0.703	0.262	0.375	0.224	0.476
2		✔	64.8	59.8	72.8	11.9	0.222	0.715	0.281	0.430	0.233	0.486
3	✔	✔	**67.9**	**61.6**	**75.7**	12.0	**0.226**	**0.736**	**0.298**	**0.457**	**0.238**	**0.498**

Table C: Ablation study on R2R val unseen [1]. See §3 for details.

3. More Ablative Study on Route Encoder

In this section, we further study the efficacy of our route encoder design. Our route encoder $\mathcal{E}^r$ considers both previous action tokens $\{\boldsymbol{a}_t\}_t$ as well as past panoramic observations $\{\boldsymbol{O}_t\}_t$ (see Eq. 3). We therefore report two variants, whose route encoder 1) only performs the temporal self-attention over the previous action tokens $\{\boldsymbol{a}_t\}_t$, and 2) only adopts the cross-attention operation to perceive the historical panoramic observation $\{\boldsymbol{O}_t\}_t$.

The results on R2R val unseen [1] are summarized in Table C. As seen, integrating both the historical panoramic observation and previous action information yields the best performance.

4. Additional Qualitative Results

In this section, we provide more qualitative results for instruction following and generation. Fig. 1 visualizes the comparison between LANA$_{mt}$ and LANA$_{st}$ on instruction generation. We can observe that LANA$_{mt}$ generates more accurate and vivid instructions. Concretely, LANA$_{mt}$ is able to not only describe precise actions (*e.g.*, turn left, walk through), but also highlight crucial landmark (*e.g.*, office, bathroom, toilet).

Fig. 2 compares LANA$_{mt}$ with LANA$_{st}$ on instruction following. Given the challenging instruction "*Leave the closet* $\cdots$ *on your left*", LANA$_{mt}$ successfully take actions to reach the target location, while LANA$_{st}$ terminates the navigation at a wrong position. This intuitively demonstrates the effectiveness of the joint-training strategy.

Fig. 3 shows the real-time behavioral description provided by LANA$_{mt}$. The generated report keeps the monitor updated on the navigation process, and reveals its inner decision mode. For examples, the route descriptions generated for Step 1-3 and Step 3-8 can vividly explain to human how LANA$_{mt}$ executes the complex command "*Walk to the left of the table and chairs down the hallway*" – first "*Walk into the room. Stop in front of the table*," then "*exit the room and walk through the hallway*". This case reveals the advantages of LANA in interpretability and human-robot communication.

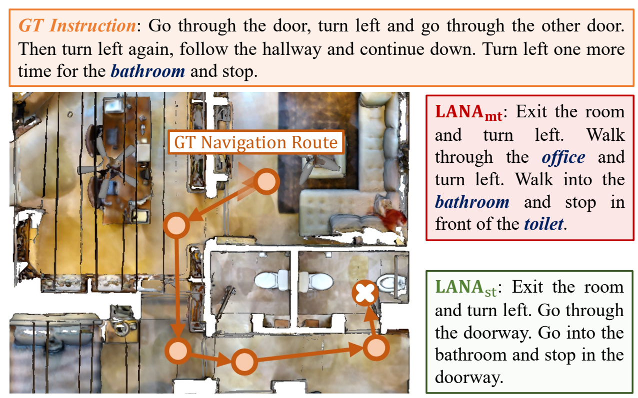

Figure 1: Visual comparison results between LANA$_{mt}$ and LANA$_{st}$ for the instruction generation task (§4). The start and end points of a navigation route are respectively denoted by and .

5. Discussion

Social Impact. A language-capable navigator can find much broader application scenarios compared with previous "dumb" ones and can be more deeply involved into human daily life. It can also serve as a guide robot to assist people who are low-vision or blind.

Limitations. The agent is developed in virtual simulated environments. If the algorithm is deployed on a real robot in a real dynamic environment, the collisions during navigation can potentially cause damage to persons and assets. More work should be done to practice real-world deployment, *e.g.*, introducing hard constraints to the action space

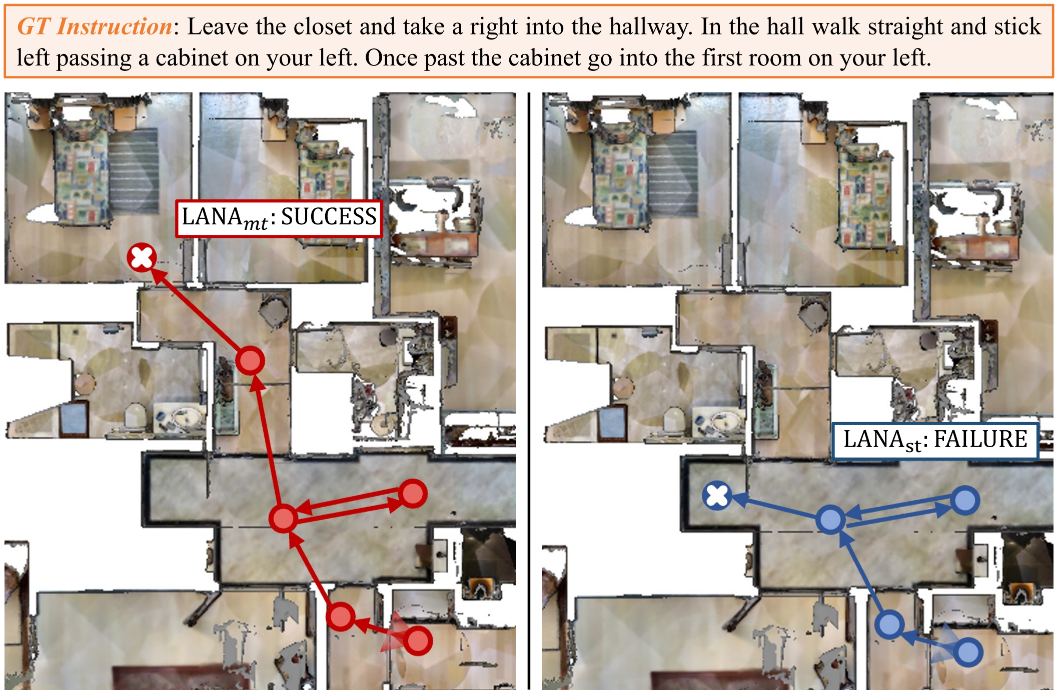

Figure 2: Visual comparison results between LANA$_{mt}$ and LANA$_{st}$ for the instruction following task (§4). The start and end points of a navigation route are respectively denoted by and .

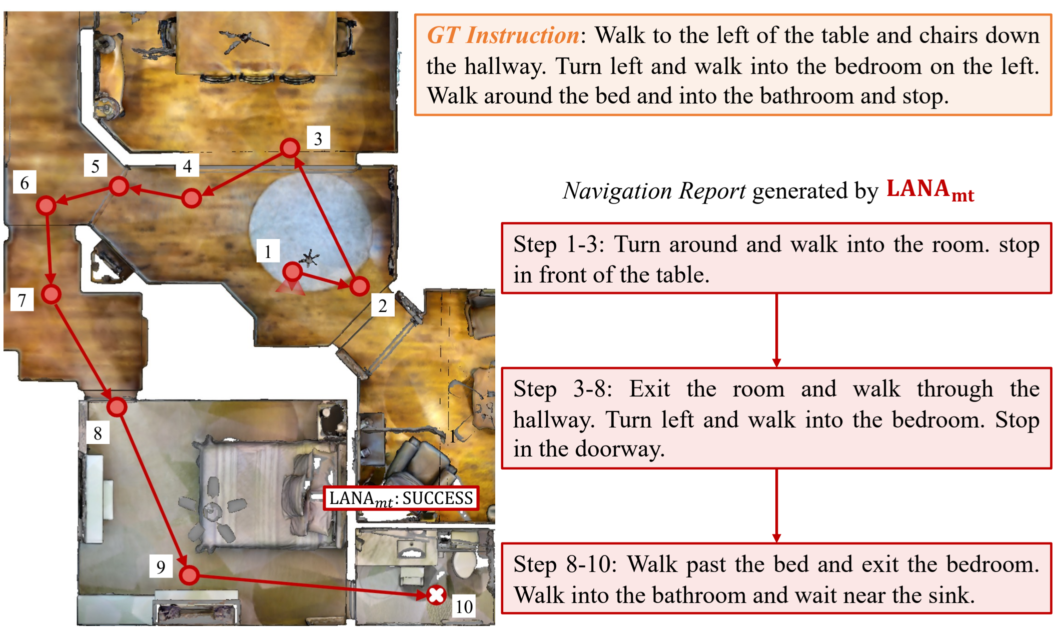

Figure 3: Step-by-step navigation behavioral explanation (§4). The start and end points of a navigation route are respectively denoted by and . LANA is able to interpret its navigation behavior using natural language. For example, at step 2-4, LANA$_{mt}$ enters the room and then exits the room because LANA$_{mt}$ intends to find the table mentioned in the instruction.

to avoid collisions, and including additional experiments to study the risk of potential damage. In addition, the generated route description, though informative and readable for human, is a kind of post-hoc interpretation. It cannot perfectly and exactly explain the inner decision mode of the agent.

Future work. In the future, in addition to investigating the efficacy of our approach in other navigation tasks (*e.g.*, object-goal navigation, audio-goal navigation), we will design more compact architecture to jointly learn the two tasks.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 1, 2

[2] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *NeurIPS*, 2021. 1, 2

[3] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, 2022. 1, 2

[4] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *ICCV*, 2021. 2

[5] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language bert for navigation. In *CVPR*, 2021. 1, 2

[6] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016. 1

[7] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 2020. 1, 2

[8] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop: History-and-order aware pre-training for vision-and-language navigation. In *CVPR*, 2022. 1, 2

[9] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, 2019. 2