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 pre-training, 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 pre-training 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.12 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.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Language Encoder 1</th>
<th>Route Encoder 2</th>
<th>Language Decoder 3</th>
<th>Route Decoder 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>self_att feed/foward</td>
<td>cross_att self_att feed/foward</td>
<td>cross_att self_att feed/foward</td>
<td>cross_att self_att feed/foward</td>
</tr>
<tr>
<td></td>
<td>×9</td>
<td>×1</td>
<td>×4</td>
<td>×4</td>
</tr>
</tbody>
</table>

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 \( E^r \), Language Encoder \( E^l \), Language Decoder, \( D^l \), and Route Decoder \( D^r \).

1. Initialize \( E^r, E^l, D^l, D^r \).
2. for iteration \( i \in [1, \ldots, N] \) do
3. Sample batch \( B \subset \mathcal{H} \)
4. Sample a pretraining task \( T \) from \{IG, IF\}
5. \( L \leftarrow 0 \)
6. if \( T \) is IG then
7. for \( (R, X) \in B \) do
8. \( \bar{x}_{t-1} \) = \( E^l(R) \)
9. \( x_{t-1} = D^l(\{\bar{x}_{t-1}, [\bar{r}_{1:t-1}]\}) \)
10. \( q_t = D^r(\{\bar{x}_{t-1}, [\bar{r}_{1:t}]\}) \)
11. Calculate \( \partial \mathcal{L} \)
12. Calculate \( \partial \mathcal{L} \)
13. Update \( E^r, E^l, D^l \)
14. else if \( T \) is IF then
15. for \( (R, X) \in B \) do
16. \( \bar{r}_{t-1}, O_t \) = \( E^r([\bar{r}_{1:t-1}, O_t]) \)
17. \( x_{t} = E^l([\bar{r}_{1:t-1}, O_t]) \)
18. \( \bar{x}_{t} = D^l([\bar{r}_{1:t}, O_t], [\bar{x}_{1:t}]]) \)
19. \( p_t = D^r([\bar{r}_{1:t}, O_t], [\bar{x}_{1:t-1}]]) \)
20. Calculate \( \partial \mathcal{L} \)
21. Update \( E^r, E^l, D^l \)
22. Calculate \( \partial \mathcal{L} \)
23. return \( E^r, E^l, D^l, D^r \)

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.
### 3. More Ablative Study on Route Encoder

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

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 · · · 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.

![Image](Image 514x246 to 522x254)

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

### 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.
**GT Instruction:** Leave the closet and take a right into the hallway. In the hall walk straight and stick left passing a cabinet on your left. Once past the cabinet go into the first room on your left.

**LANA \text{mt}: SUCCESS**

![Map](image1.png)

**LANA \text{st}: FAILURE**

![Map](image2.png)

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

**GT Instruction:** Walk to the left of the table and chairs down the hallway. Turn left and walk into the bedroom on the left. Walk around the bed and into the bathroom and stop.

**Navigation Report generated by LANA\text{mt}**

- **Step 1-3:** Turn around and walk into the room. stop in front of the table.
- **Step 3-8:** Exit the room and walk through the hallway. Turn left and walk into the bedroom. Stop in the doorway.
- **Step 8-10:** Walk past the bed and exit the bedroom. Walk into the bathroom and wait near the sink.

![Map](image3.png)

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


