LLM as Copilot for Coarse-grained Vision-and-Language Navigation - Supplementary

Yanyuan Qiao[1], Qianyi Liu[2,3], Jiajun Liu[4,5], Jing Liu[2,3], and Qi Wu[1⋆]

[1] Australian Institute for Machine Learning, The University of Adelaide
[2] Institute of Automation, Chinese Academy of Sciences
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences
[4] CSIRO Data61
[5] The University of Queensland
{yanyuan.qiao,qi.wu01}@adelaide.edu.au, liuqianyi2022@ia.ac.cn,
jiajun.liu@csiro.au, jliu@nlpr.ia.ac.cn

1 Modification of Base Agent

We use AutoVLN [2] as the base agent. It mainly consists of two modules: the topological mapping module which builds a topological map during navigation, and the global action planning module which predicts the next location on the map or a stop action to end the navigation. For more details on the base agent, please refer to the paper of AutoVLN [2].

To facilitate the navigation guidance from the LLM, we modify both the base agent's language embedding and the action prediction process. Specifically, if the computed confusion score surpasses the threshold, the guidance information R_{guide} given by the LLM is passed through the language encoder of the agent to generate the guidance embedding E_R of the guidance. Then, E_R is concatenated with the pre-embedded instruction embedding E_I as the final language embedding E_L, which is used to implement cross-attentions with topological graphs and visual observations respectively. For action prediction, we equip the agent with the capability of self-reviewing. Detailedly, the initial predicted logits over the candidates are used to compute the confusion score rather than to select the next action. If the confusion score is lower than the threshold, which means the agent is confident in the initial prediction, the agent will select the next action according to the logits. Otherwise, the agent will ask LLM for help and make new predictions with the help of the guidance information.

2 Additional Ablation Study

Different LLMs We also conducted experiments on the other LLM of Llama2-7B [3] to validate the generalization of our proposed method. As shown in Table 1, the performance of the Llama2-7B model is competitive with the Vicuna-7B-1.5 model [4].

⋆ Corresponding author: Qi Wu

Table 1: Comparison of using Llama2-7B and Vicuna-7B-1.5.

Method	Navigation			Grounding	
	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
Llama2-7B	61.35	56.01	43.08	38.09	29.34
Vicuna-7B-1.5	62.62	57.40	43.63	38.88	29.75

3 More Qualitative Examples

Example of LLM-generated Guidance As shown in Figure 1, we provide more examples of different types of LLM-generated guidance. In sub-figure (a), though the agent has arrived at the target location, it is still confused to make the prediction. Then, the LLM tells the agent about the current location and the target location to help it make the final decision. In sub-figure (b), the LLM points out that the prior task is to find the office with a tall plant, then the agent will go to the office with the plant first and find the pictures later. In sub-figure (c), the agent could refer to the description of the wine pantry and know that it should reach a place that has shelves filled with wine. In sub-figure (d), based on the previous observation, the agent can know that it has gone down the stairs, and then when it chooses a direction, it will choose candidate 3 with the kitchen instead of candidates 1 and 2 with stairs.

Visualization of Navigation Trajectory We also visualize navigation trajectories predicted by our VLN-Copilot and the other method AutoVLN [2] on the REVERIE dataset. As shown in Figure 2, our agent could successfully reach the dining room and stop near the chair according to the instructions, while AutoVLN's agent deviated from the correct direction after Step 3, it did not go to the target destination and walked the wrong room. In addition, the trajectory visualization is shown in Figure 3. In step 2, the agent seeks help from the LLM, receiving guidance: "Your current location is a lobby. The destination is level 3." This indicates the agent needs to go upstairs rather than stay on the current floor. The agent then returns, demonstrating how LLM helps correct its route when necessary.

4 Datasets

REVERIE contains 21,702 instructions, the average length of each instruction is 18 words. The path length is about 4∼7 steps. The dataset has 10,567 panoramas and 4,140 target objects, divided into 489 categories. On average, each target viewpoint has 7 objects with 50 bounding boxes. REVERIE follows the same train/validation/test split strategy as the R2R dataset. The training set contains 59 scenes with over 2,353 objects and 10,466 instructions. The validation set contains 63 scenes, 953 objects, and 4,944 instructions. The test set contains 16 scenes, 834 objects, and 6,292 instructions.

Fig. 1: Examples of the LLM-generated guidance.

5 Evaluation Metrics

(1) Trajectory Length (TL) measures the average length of all the predicted navigation trajectories in meters. (2) Success Rate (SR) measures the ratio of successful tasks, of which the agent's stop location is less than 3 meters away from the target location. (3) Oracle Success Rate (OSR) measures the ratio of tasks of which one of its trajectory viewpoints can observe the target object within 3 meters. (4) Success weighted by Path Length (SPL) [1] trades-off SR against TL, which measures both the accuracy and efficiency of navigation. (5) Remote Grounding Success rate (RGS) measures the ratio of tasks that successfully locate the target object. (6) RGS weighted by Path Length

Fig. 2: Visualization of the predicted trajectory of our method VLN-Copilot and AutoVLN [2].

(RGSPL) is RGS weighted by Path Length. SPL is the key metric for navigation and RGSPL is the key metric for object grounding. (7) Goal Progress (GP) measures the difference between the completed distance (from the start to the target) and the left distance to the goal.

References

1. Anderson, P., Chang, A.X., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., Zamir, A.R.: On evaluation of embodied navigation agents. CoRR **abs/1807.06757** (2018)
2. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Learning from unlabeled 3d environments for vision-and-language navigation. In: ECCV (2022)

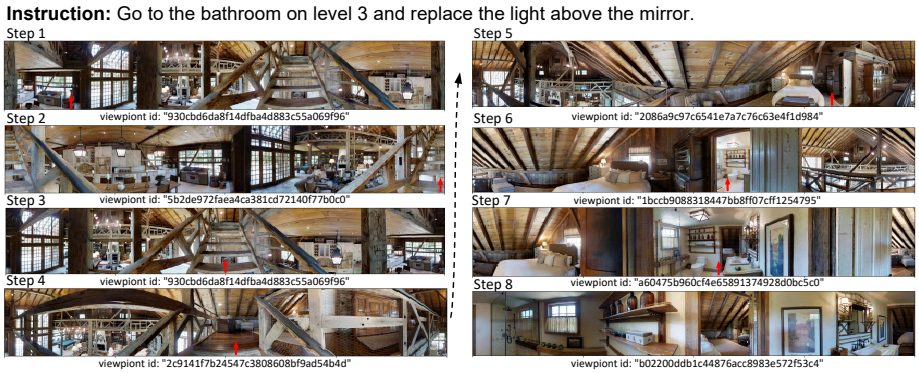

Fig. 3: Trajectory Visualization.

3. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
4. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. arXiv preprint arXiv:2306.05685 (2023)