Demo Abstract: Embodied Aerial Agent for City-level Visual Language Navigation Using Large Language Model

Weichen Zhang[1,4], Yuxuan Liu[1], Xuzhe Wang[1], Xuecheng Chen[1], Chen Gao[2,3], Xinlei Chen[1,4,5†]
{zhangwc23,yuxuan-l21,wang-xz21,chenxc21}@mails.tsinghua.edu.cn
chgao96@tsinghua.edu.cn;chen.xinlei@sz.tsinghua.edu.cn
[1]Shenzhen International Graduate School, [2]Department of Electronic Engineering, [3]BNRist, Tsinghua University;
[4]Pengcheng Laboratory; [5]RISC-V International Open Source Laboratory

ABSTRACT

As unmanned aerial vehicles (UAVs) become more prevalent in smart cities, their capacity for visual language navigation (VLN) is garnering increasing interest. VLN in cities has significant applications in delivery, rescue, and security patrol, among other fields. One of the most representative tasks is to navigate to specific locations following the language instructions. While some current methods have achieved notable results in indoor settings, challenges persist outdoors, including agents' inaccurate spatial understanding and ambiguous language instructions. In this work, we explore an embodied navigation agent design, in which a fine-grained spatial verbalizer and a history path memory are proposed to guarantee accurate VLN in open 3D urban environments.

KEYWORDS

Visual language navigation, urban, embodied navigation agent

1 INTRODUCTION

In urban environments, VLN by drone is gaining widespread attention for its application on public safety [1] and crowdsensing [2, 3]. While current outdoor navigation typically relies on map-based methods, some important locations are not present on maps. For example, in drone-assisted firefighting operation, drones must find the nearest water sources or fire hydrants–entities are not marked on the map–to replenish their water supplies. Besides, constant and accurate GPS signal may not be always available when UAVs are near urban buildings [4]. Therefore, it's crucial for drones to navigate to targets through perception and reasoning without maps.

Existing VLN methods such as NavGPT [5] are deployed within limited, discrete environments, represented by a navigation graph. However, navigation in three-dimensional city space requires drones to move continuously toward distant targets indicated by language instructions. Thus, urban VLN still meets the following challenges: **1) complex spatial understanding.** Objects can appear in arbitrary positions in three-dimensional space [6]. The UAV needs to fully comprehend complicated spatial relations for accurate action control such as accurate landing [7]. **2) ambiguous instruction for long-term navigation.** Occasionally, UAVs are required to execute long-term navigation tasks with limited instruction, such as "navigating to the nearest park". The UAV struggles to derive a clear destination and path from such vague descriptions.

In this demonstration, we developed an aerial VLN agent leveraging a large language model (LLM) for urban navigation. To tackle the two challenges, we design a navigation prompter enhancer (NPE) to augment the aerial agent's spatial perception and path routing capabilities. To the best of our knowledge, it is the first embodied navigation agent for large-scale, three-dimensional continuous spaces.

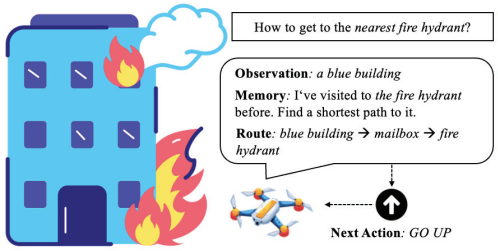

Figure 1: Navigation process of embodied agent

2 SYSTEM DESIGN

Fig. 2 presents our system architecture. Our system comprises a multimodal perception module and an action decision module based on commonsense reasoning. The agent first perceives the environment through visual and lingual information. Then, the NPE extracts richer spatial information from the perception results. Besides, NPE leverages the memory layer storing historical trajectories to provide more navigation clues for ambiguous navigation tasks.

Multimodal perception: To thoroughly perceive its surroundings, the drone captures a panoramic image by rotating itself and a language model serves as a landmark extractor to identify landmarks within the navigation instructions. A vision-language model (VLM) namely GroundingDINO is then employed to detect landmarks in each image. If the same landmark appears in multiple views, the perspective in which the landmark has the highest score is designated as the landmark's observation viewpoint.

Navigation prompt enhancer: NPE is designed to fill the gap between the perception results and the reasoning input.

Fine-grained spatial verbalizer: When the UAV observes landmarks, the spatial information of landmarks is converted to text for the next action inference. However, the naive spatial description including only viewpoint information is inadequate for the LLM to discern the relative positioning. For example, a road detected in the front view might be positioned centrally or to the left, yet only a central positioning cues the drone to move along the road. To bridge this gap, we propose a fine-grained spatial verbalizer (FSV). FSV splits each viewpoint of the UAV into nine distinct bins, each representing a more detailed spatial description, such as 'top left' for the upper-left corner. Hence, the spatial information of the landmarks is described by the viewpoint as well as their precise locations within each view.

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00033

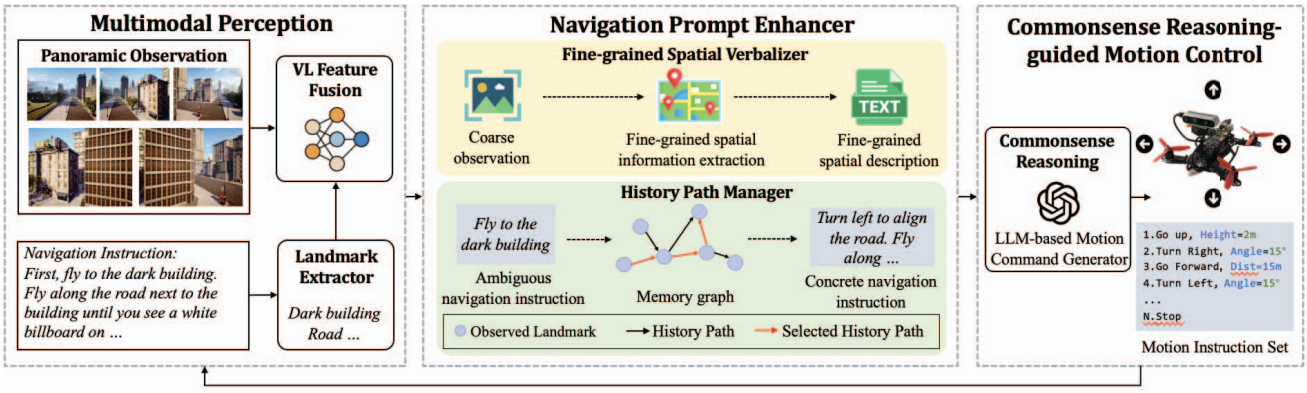

Figure 2: The overall architecture of the VLN agent

History Path Memory: Occasionally, navigation instructions may include ambiguous descriptions, such as "fly to the nearest park". The UAV fails to derive a navigable path from such ambiguous instructions. Therefore, we have designed a history path memory (HPM) module. The HPM stores the agent's historical trajectories and instructions in a graph, with nodes representing previously encountered landmarks and edges storing specific navigation instructions between landmarks. For any two nodes, a navigable path can be retrieved by the shortest path algorithm. Thus, the agent can derive a feasible path from the HPM by current observed landmarks and the target landmark for ambiguous instructions.

Commonsense reasoning-guided motion control: Once the fine-grained observation description and well-described navigation instructions are obtained, a formatted prompt is designed for LLM to predict the next action for the UAV. The complete text prompt is composed of UAV's discrete action space, current observations, navigation instruction, and history actions. The goal of LLM is to predict the next action which is converted to a control command for the UAV to execute.

3 DEMONSTRATION DESCRIPTION

As shown in Fig. 3, we demonstrate an aerial visual navigation task within a simulated environment modeled after Guomao area in Beijing. Initially, the UAV is positioned at a point where the target is not visible. Subsequently, the system predicts the next-step action solely based on visual observations and navigation instructions until the agent determines it has reached its destination. Moreover, the system also outputs the rationale in each step so that participants can understand the reasoning process.

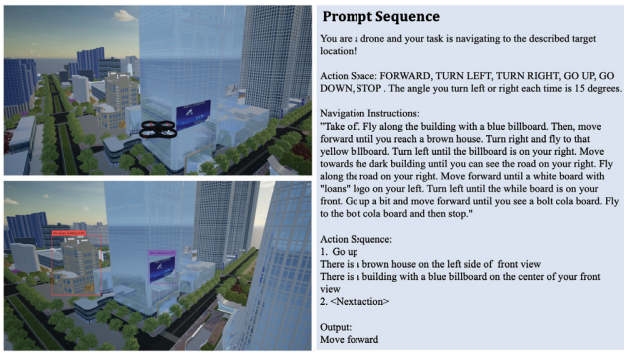

Figure 3: Navigation in simulated city environment

ACKNOWLEDGMENTS

This paper was supported by the National Key RD program of China No. 2022YFC3300703, the Natural Science Foundation of China under Grant No. 62371269. Shenzhen 2022 Stabilization Support Program No. WDZC20220811103500001, and Tsinghua Shenzhen International Graduate School Cross-disciplinary Research and Innovation Fund Research Plan No. JC20220011. The Major Key Project of Peng Cheng Laboratory (PCL) under Grants PCL2023A09.

REFERENCES

[1] Zuxin Li, Fanhang Man, Xuecheng Chen, Baining Zhao, Chenye Wu, and Xinlei Chen. Tract: Towards large-scale crowdsensing with high-efficiency swarm path planning. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pages 409–414, 2022.

[2] Xuecheng Chen, Haoyang Wang, Zuxin Li, Wenbo Ding, Fan Dang, Chengye Wu, and Xinlei Chen. Deliversense: Efficient delivery drone scheduling for crowdsensing with deep reinforcement learning. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pages 403–408, 2022.

[3] Yuxuan Liu, Xinyu Liu, Fanhang Man, Chenye Wu, and Xinlei Chen. Fine-grained air pollution data enables smart living and efficient management. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 768–769, 2022.

[4] Haoyang Wang, Xuecheng Chen, Yuhan Cheng, Chenye Wu, Fan Dang, and Xinlei Chen. H-swarmloc: Efficient scheduling for localization of heterogeneous mav swarm with deep reinforcement learning. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 1148–1154, 2022.

[5] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*, 2023.

[6] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15384–15394, 2023.

[7] Chenyu Zhao, Haoyang Wang, Jiaqi Li, Fanhang Man, Shilong Mu, Wenbo Ding, Xiao-Ping Zhang, and Xinlei Chen. Smoothlander: A quadrotor landing control system with smooth trajectory guarantee based on reinforcement learning. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pages 682–687, 2023.