EmbodiedNav: Spatial Cognition-Aware Embodied Navigation in Urban Spaces with Large Multimodal Model Agent

Anonymous ACL submission

Abstract

Asking an aerial agent to navigate to any location-goal in an unseen environment autonomously is one of the fundamental tasks in embodied artificial intelligence. However, embodied navigation in low-altitude urban spaces is understudied, with aerial agents facing significant perception and reasoning challenges due to the complexity of large-scale three-dimensional environments. To fill this gap, we propose EmbodiedNav, a large multimodal model-empowered agent for goal-oriented embodied navigation, inspired by three mechanisms of spatial cognition psychology: size constancy, viewpoint invariance, and spatial perspective-taking. Our formulation mimics how humans navigate using only RGB visual information, without relying on maps, odometers, or depth inputs. Experimental results demonstrate that our method significantly exceeds existing solutions and ablation analysis further reveals that the spatial cognition mechanism leads to more effective embodied navigation.

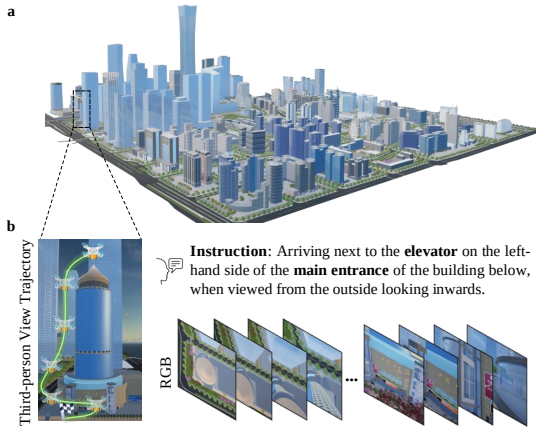

Figure 1: **a.** The urban simulator (Gao et al., 2024) is based on Beijing, a megacity in China. Goal-oriented embodied navigation cases distributed among different buildings are established. **b.** For each goal-oriented navigation case, the agent is required to autonomously reach its goal based on location instructions, navigating through an unfamiliar environment merely using RGB observations. An example trajectory is depicted by the green line, while flags represent the destination location.

1 Introduction

Embodied intelligence research (Gupta et al., 2021) requires robots to navigate effectively to a destination based on human goal-directed language instructions, such as "go to *some target location*". In this process, the robot utilizes first-person visual observation to identify spatial markers and, through step-by-step planning, navigates to the destination. Existing research (Liang et al., 2024) primarily focuses on indoor ground robots, with relatively little attention given to the more fundamental yet challenging task of aerial embodied navigation in urban spaces. A typical application scenario is drone-based emergency medical services (Roberts et al., 2023; Liu et al., 2023b), which aim to reduce response times by delivering critical medical supplies, such as automated external defibrillators and naloxone, to patients in urban areas during time-sensitive emergencies like cardiac arrest, stroke, drug overdose, seizures, and trauma.

Two remain critical challenges as follows:

- **Sensing in complex environment.** Our goal is to develop a navigation agent with enhanced generalization capabilities (Zhang et al., 2024), allowing it to navigate using only sequential RGB observations, much like humans, without relying on GPS, depth, or odometry data. A solution based on pure RGB inputs is particularly advantageous for drones, as it is hardware-friendly. However, this approach requires the agent to infer spatial information such as three-dimensional structures, distances, and orientations solely from RGB observations.

- **The gap between navigation actions and goals**

when navigating in urban spaces. Agents are required to translate goal instructions and sensory observations into actionable steps (Shah et al., 2023). These agents must possess fundamental capabilities in both vision and language comprehension, as well as an understanding of urban commonsense. Moreover, in the realm of low-altitude urban flight scenarios, the vastness of urban outdoor spaces demands reasoning based on an understanding of streets and building structures to navigate effectively (Gao et al., 2024).

Spatial cognition psychology is a field that explores psychological processes and mechanisms involved in human perception and understanding of spatial environments (Farzanfar et al., 2023; Brucato et al., 2023). The spatial capabilities are facilitated by the brain's ability to integrate continuous RGB visual inputs with prior knowledge and contextual information, forming a coherent representation of the world. The human visual system uses various cues to infer depth and maintain a stable perception of object size and shape. By mimicking these natural processes, embodied agents can be designed to navigate and interact with their environments merely with RGB sensing, enhancing their ability to perform tasks that require spatial awareness and adaptive learning in dynamic settings (Newcombe et al., 2023).

Meanwhile, recent advancements in embodied navigation have shown impressive results (Schumann et al., 2024; Huang et al., 2023a), particularly with agents that utilize large pre-trained model-based planners (Schumann et al., 2024; Zhou et al., 2023a). However, while LMM can describe the content of images observed by the agent, they fall short in understanding the spatial relationships between ego and environment or the agent's intent to alter these relationships through actions. Therefore, th navigation agent requires more sophisticated designs that extend beyond the capabilities of current LMM.

Actually, the key to addressing the above challenges is to improve the spatial cognition abilities in urban spaces. Therefore, we propose an embodied agent called **EmbodiedNav**, which is inspired by the human's spatial cognition mechanisms. Specifically, EmbodiedNav introduces three spatial cognition mechanisms—size constancy, viewpoint invariance, and spatial perspective-taking. These mechanisms facilitate reasoning about spatial relationships from RGB perceptions, such as egocentric relative distances, orientations, and understanding of three-dimensional objects. Building on this, we further developed an LLM-empowered agent that integrates perception, planning, skill, and action execution modules through a memory module, enabling the mapping from RGB observations to actions. To summarize, this work makes the following contributions:

- We developed a LMM-empowered agent's three spatial cognition abilities—size constancy, viewpoint invariance, and spatial perspective-taking—through memory retrieve and observation changes in perspective. To the best of our knowledge, this work is the first to highlight the principles of spatial cognition in agents, mimicking human-like spatial perception and action using only RGB input, without additional information such as depth, odometer data, or maps.

- We deduce the agent's five functional modules: perception, planning, skill, action, and memory, as shown in Figure 2. By stimulating spatial imagination, planning performance is enhanced. Additionally, skills translate each step of the plan into a sequence of actions.

- Experimental results demonstrate that the proposed agent provides a viable solution for urban goal-oriented embodied navigation tasks compared to existing LMMs and indoor navigation frameworks.

2 Problem Formulation

The objective of the goal-oriented embodied navigation problem is to reach the specified goal-location p_L, given through a natural language instruction I. The process involves following an algorithm π to guide the agent through a sequence of observations and actions to reach the target locationp_L. At each time step t, the agent obtains an RGB observation o_t. The agent takes action a_t based on π:

$$a_t = \pi\left(o_t, I\right) \tag{1}$$

The action a_t can be formed by the arbitrary combination of discrete control commands: turn-left, turn-right, move-forward, move-left, move-right, move-back, move-up, move-down, adjust-camera-gimbal-upwards, and adjust-camera-gimbal-downwards.

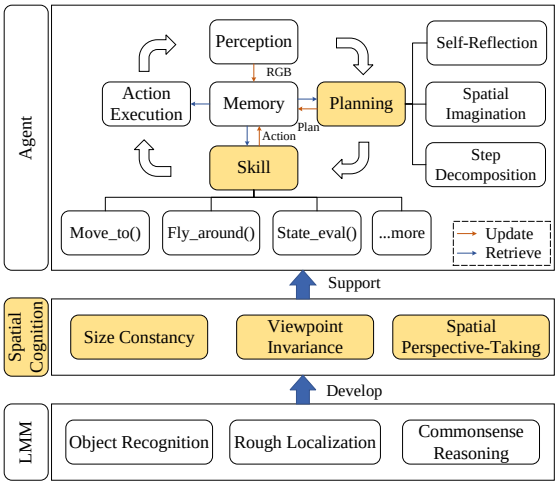

Figure 2: Overview of EmbodiedNav framework based on foundational capabilities from LMM, integrating three spatial cognition principles, i.e. size constancy, viewpoint invariance, and stereoscopic perception. These elements inform the design of the agent's key modules—perception, planning, skill, action, and memory—enabling effective navigation within 3D urban spaces.

In the urban space, the real position p_t of the agent, which can not be observed by itself, is changed based on physical dynamics rules:

$$p_t = f\left(a_t, p_{t-1}\right) \tag{2}$$

After T steps, the navigation is successful if the agent stops within a Euclidean distance of ε meters from the target location p_L.

$$\|p_T - p_L\| \leq \varepsilon \tag{3}$$

We aim for the agent to reach the target location in diverse scenarios. Assuming there are scenarios $i = 1, 2, ..., N$, the objective can be formally stated as follows:

$$\max_{\pi} \frac{1}{N} \sum_{i=1}^{N} 1\left(\left\|p_T^{(i)} - p_L^{(i)}\right\| \leq \varepsilon\right) \tag{4}$$

where $1\left(\cdot\right)$ is the indicator function, which is 1 if the condition inside is true and 0 otherwise.

3 Benchmark Setup in City Spaces

Dataset The environment is built based on the simulation environment in (Gao et al., 2024). In total, there are $N = 1,089$ delivery trajectories were established. As shown in Figure 1(b), each delivery trajectory i consists of a starting point $p_0^{(i)}$, an endpoint $p_L^{(i)}$ along with the corresponding textual instruction, and the ground truth route. The routes are obtained by manually controlling the drones. Through the construction of these realistic scenarios, we can evaluate agent's ability to navigate within open-world urban environments and accurately locate delivery destinations.

Metrics We utilize three standard metrics to evaluate the navigation performance: Success Rate (SR), Success Weighted by Path Length (SPL), and Distance to Goal (DTG) (Anderson et al., 2018; Yu et al., 2023; Zhou et al., 2023a). SR indicates the proportion of delivery episodes where the agent successfully reaches the target location within a specific margin of error. Similar to Eq. 4, it is calculated using $\text{SR} = \frac{1}{N}\sum_{i=1}^{N} s_i$, where N is the number of delivery episodes and s_i represents the success of the i-th delivery, where it takes a value of 1 for success and 0 for failure. As a metric that considers both navigation precision and efficiency, SPL comprehensively takes into account the SR and the corresponding ratio of the optimal path length l_i to the actual delivery path length g_i. The calculation formula is represented as $\text{SPL} = \frac{1}{N}\sum_{i=1}^{N} s_i \frac{l_i}{\max(l_i, g_i)}$. DTG is computed by $\text{DTL} = \frac{1}{N}\sum_{i=1}^{N} d_i$, where d_i denotes average distance from the agent's final location to the destination.

4 Methodology

We first develop three foundational mechanisms of the embodied agent's spatial cognition: size constancy, viewpoint invariance, and spatial perspective-taking. These enable EmbodiedNav to establish spatial awareness relying solely on RGB observations and memory. Subsequently, these abilities are integrated into the agent's design to facilitate the mapping from perception to action.

4.1 Spatial Cognition Mechanism

In spatial cognition, humans understand relative distances and relationships in a three-dimensional world without depth cues primarily through two mechanisms: The ability to infer spatial relationships by observing changes in the visual scene following actions (action-driven perception) and the capacity for spatial imagination of three-dimensional objects. These abilities correspond to three specific spatial cognitive skills, as shown in Figure 3. Through memory retrieval of relevant perceptions and designed prompts, these skills can be developed and enhanced.

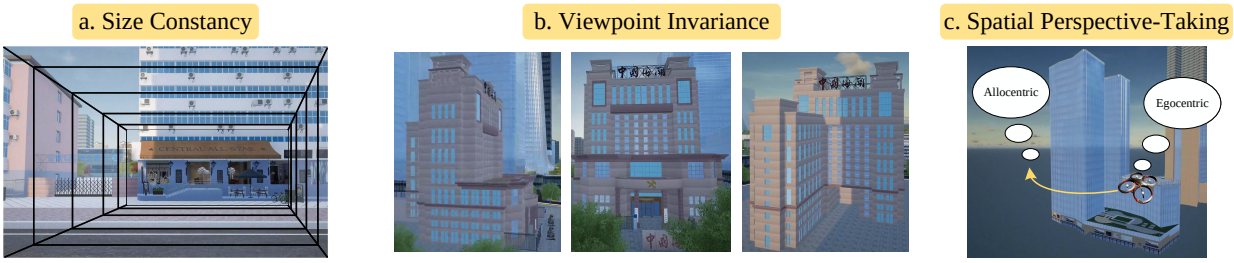

Figure 3: Three key mechanisms for spatial cognition. Size constancy indicates an object's perceived size remains constant despite changes in distance. Viewpoint invariance means an object/urban element is consistently recognized from different angles. Spatial perspective-taking means the agent can understand and imagine relationships between different perspectives.

4.1.1 Size Constancy

Size constancy is the perceptual ability to perceive the size of an object as constant despite changes in its distance from the observer. This phenomenon can be described mathematically by considering the relationship between the actual size of an object, the size of the image it projects on the retina, and the viewing distance. The perceived size P of an object is determined by the equation:

$$P = k \cdot \left(\frac{S \cdot D}{d} \right), \tag{5}$$

where: P is the perceived size, k denotes a scaling constant, S is the actual size of the object, D represents the perceived distance from the observer to the object, and d is the distance from the observer's eye to the image on the retina.

Based on the relationship between perceived size and distance, an embodied agent can extract RGB observations before and after an action from memory, allowing the LMM to compare the relative size of the viewpoint/goal in the visual field. A prompt example is:

> *Given [RGB observation 1] (e.g., the outermost view in Figure 3(a) and [RGB observation 2] (e.g., the innermost view in Figure 3(a), has the area occupied by the cafe in the image significantly increased or decreased?*

Furthermore, the agent can use this inverse relationship to determine the relative proximity between the egocenter and the visual target. This is crucial for the agent's sense of distance in the skill module.

4.1.2 Viewpoint Invariance

Viewpoint invariance refers to the ability to recognize an object from different angles or perspectives, as shown in Figure 3(b). This cognitive ability is crucial for understanding and interacting with the spatial environment. Let O represent an object, and $V_1, V_2, \ldots, V_n$ represent different viewpoints. An object I is said to exhibit viewpoint invariance if there exists a function f such that:

$$I = f(O, V_i) = f(O, V_j) = \text{const} \quad \forall i, j. \tag{6}$$

The agent requires sequential observations with a certain degree of overlap as cues to determine whether an object in the current view is the same as an object in past observations. This explains the importance of the memory module and why the action generation in skills needs to incorporate historical observations. Memory stores contextual information about spatial environments, such as the typical arrangement of features and objects. This context helps maintain viewpoint invariance by providing additional cues for recognition and orientation. This capability of viewpoint invariance is closely related to the agent's sense of direction in skills such as navigation and loop closure.

4.1.3 Spatial Perspective-Taking

Spatial perspective-taking is a cognitive process that involves the ability to imagine scenes from allocentric perspectives different from the egocentric perspective, as depicted in Figure 3(c). This requires the mental transformation of spatial information to understand how objects and their relationships appear from another perspective. To enable the embodied agent to understand the three-dimensional nature of objects and how their perspective affects perception, the corresponding visualization and descriptive prompts can be em-

ployed. The visualization approach involves generating scenarios with an imagination generator, as described in the planning module. In prompt design, we can incorporate descriptions related to three-dimensional and volumetric attributes. For example:

> *The building containing the goal is a rectangular prism with observable surfaces: the top surface and four side surfaces. The optimal perceptual angle for the top surface is a top-down view. For the side surfaces, the best perception angles are either a horizontal view or a 45-degree overhead view ...*

The ability of spatial perspective-taking is crucial for macro-scale scene understanding in the planning module.

4.2 EmbodiedNav Agent

Given the aforementioned spatial cognition capabilities, we design the LLM agent with five modules: perception, planning, skill, action, and memory. The **perception** module obtains an RGB observation at each time step; the **planning** module outputs the latest plan; the **skill** module converts steps in the plan into a combination of skills, which then outputs the agent's action; the **action** is finally executed. Throughout this process, the output of each module is updated in the **memory** module, providing information to the planning, skill, and action execution modules.

4.2.1 Spatial Planning

The role of the spatial planning module is to output a high-level plan to approach the goal location gradually, as shown in Figure 4. It begins by retrieving historical sequences of RGB observations and plans from the memory module. The self-reflection module then determines whether the plan needs adjustment. If no adjustment is needed, it evaluates whether the current step in the plan is complete. If complete, it outputs the next step; otherwise, it continues with the current step.

If the plan requires creation or modification, spatial imagination is activated. Imagination is needed because the goal-location is not directly observable; the agent must imagine its spatial relationship with the currently observed urban elements. The focus is on the relationship between the goal and the scene, which consists of abstract spatial regions not easily identifiable in the view. An imagination generator is used to create visualizations of common spatial relationships between the goal and buildings, enhancing LMM's reasoning. Through images and prompts, the agent gains spatial perspective-taking, allowing it to understand that buildings have multiple sides in 3D space. To find the final goal, the agent must move around obstructions and approach the goal's vicinity, observing from a specific camera pan angle. This imagination narrows the search area, enabling quick goal localization. Additionally, the agent considers the relationship between easily identifiable urban elements or objects and the goal, such as a main entrance facing a road, which aids in determining the search direction and speeding up navigation.

Using the results of this imagination, navigation is broken down into feasible steps. The plan is stereoscopic, involving vertical and horizontal movements and camera angle adjustments for effective perception. The final output is the updated plan and the current step to execute.

4.2.2 Skill Curation

To facilitate mapping steps in the plan to actions, we extract common steps into skills, including: moving to an abstract spatial location relative to the scene/building ($move_to()$), flying around a target ($fly_around()$), adjusting the view ($adj_view()$), and approaching an object/element in view ($approach()$). These skills also provide feasibility constraints for step decomposition in planning. The skills are grounded in spatial cognition principles, enabling spatial awareness such as relative distance and scene-object consistency, thereby establishing action mapping, as shown in Figure 5.

4.2.3 Memory Update and Recall

As previously mentioned, the capabilities of spatial cognition and viewpoint invariance are derived from the comparison and synthesis of continuous RGB observations. The agent, built upon foundational LMM capabilities and spatial cognition, relies on memory for its planning, skill, and action execution modules. The memory is organized into four categories: RGB observations, plan, skill, and action, along with their reasoning processes at each time moment. Each module retrieves the necessary components from memory as inputs as needed. Memory enables individuals to remember paths, directions, or waypoints required to reach

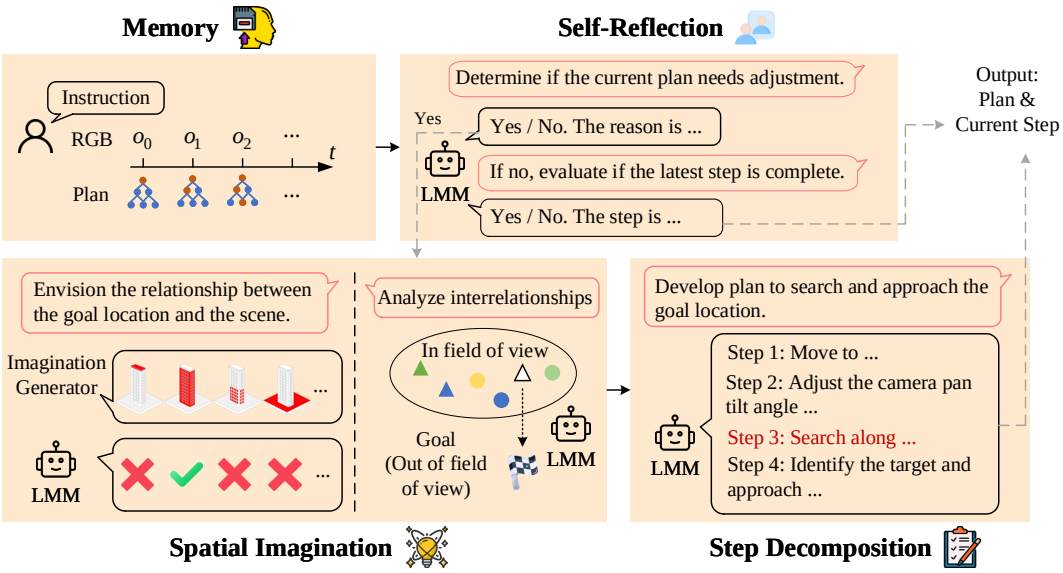

Figure 4: The design of the planning module primarily consists of three submodules: self-reflection, spatial imagination, and step decomposition. Self-reflection determines whether the plan needs adjustment and assesses the progress of the current step based on the latest observations. If a plan needs to be created or adjusted, spatial imagination is performed first. Through the imagination generator, possible three-dimensional spatial relationships are visualized to aid LMM reasoning. Additionally, interrelationship reasoning is conducted to determine which approximate region in view is most likely spatially adjacent to the target location. Based on the reasoning results, navigation is broken down into different steps, which can be executed through a combination of skills. Finally, the module outputs the current updated plan and the steps that need to be executed.

a destination, helping in both habitual and novel navigation tasks.

5 Experiment

5.1 Experimental Setup

Implementation Details: To ensure fairness, all LMM-empowered methods utilize the same foundation model. The parameters for capturing RGBD observations and pose from the drone are set to the default configuration in the AirSim platform (Shah et al., 2017).

Baselines: The compared methods include **Random**, **Action Sampling**, **AG-GPT4** (Vemprala et al., 2023), **NavGPT** (Zhou et al., 2023a), **ZSON** (Majumdar et al., 2022), and **SayNav** (Rajvanshi et al., 2023). For the goal-oriented aerial embodied navigation task in urban spaces, no direct solutions are available for direct comparison. Thus, we not only construct three common benchmarks for comparison purposes (Liu et al., 2023a), but we also adapt existing vision-language navigation approaches (Zhou et al., 2023a) and indoor object navigation (Gadre et al., 2023; Rajvanshi et al., 2023) to our scenario.

5.2 Overall Performance

All navigation cases are evenly divided into three categories based on distance: short, middle, and long. Generally, the longer the navigation distance of a case, the more complex and challenging the decision-making process becomes. The results, presented in Table 1, lead us to the following three observations.

- **The action space in urban simulator is large and suitable for embodied navigation.** Both the random and action sample methods exhibit SR and SPL scores close to 0 in middle-distance and long-distance groups. This indicates that the task encompasses a vast action space. The agent fails to reach or even approach the destination without understanding the instructions, visual perceptions, and their alignment.
- **The direct application of large multimodal models for generating drone control commands yields unsatisfactory performance.** The success rate of AG-GPT4 is also below 21%, and its DTG is higher than that of the Action Sampling method. This indicates that even one of the best LMMs lacks the capability to achieve end-to-end perception-action integration. Cur-

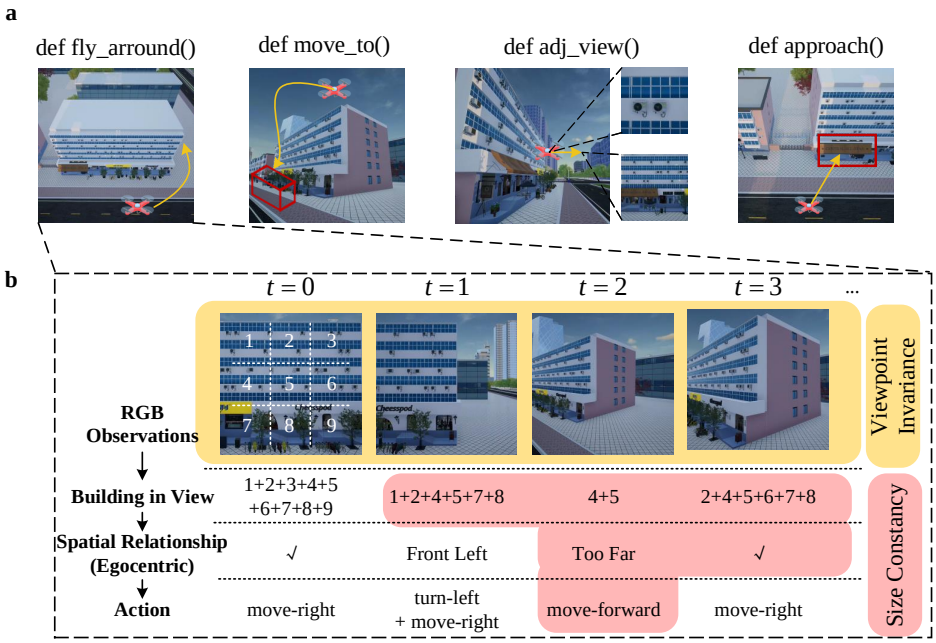

Figure 5: **a.** Examples of skill curation include flying around a target ($fly_around()$), moving to an abstract spatial location relative to the scene/building ($move_to()$), adjusting the view ($adj_view()$), and approaching an object/element in view ($approach()$). **b.** A detailed example of the $fly_around()$ skill demonstrates its functionality: flying around an area of interest while maintaining a relatively constant distance and ensuring effective camera perception. The example involves flying around a building in view. To accommodate the LMM, the field of view is divided into a grid to determine the rough location of the area of interest. At $t = 0$, the building fully appears in view, indicating the agent is at an appropriate distance from the area of interest, resulting in a move-right action. At $t = 1$, the building appears to the agent's front left; to re-center the building in view, the agent outputs turn-left and move-right actions. At $t = 2$, the building occupies a smaller area in the view, suggesting increased distance, which can impair the agent's perception, resulting in a move-forward command. The $t = 3$ RGB observation confirms the view adjustment goal is achieved. Throughout different viewpoints during the flight, the ability to consistently recognize the building of interest showcases the viewpoint invariance ability. Inferring distance based on the perceived area of the building in view illustrates the size constancy ability.

rent LMMs struggle to comprehend the three-dimensional spatial relationship between the ego and the environment, hindering their ability to make action decisions.

- **Spatial cognition capabilities are crucial for embodied navigation in urban spaces.** Embodied outperforms other methods across various difficulty scenarios, especially in the hard scenario where it achieves SR and SPL scores that are more than **twice** those of the other methods. Additionally, its DTG is significantly lower than that of other methods, indicating that the drone's final position is close to the goal location. This highlights the differences between indoor and outdoor scenes. The information density for scene perception is much lower in outdoor environments, leading to a strong coupling between perception and action, which imposes higher demands on the design of planning and skill modules. The integration of spatial cognitive psychology helps bridge the gap between basic vision-language capabilities of LMM and the agent's understanding of spatial relationships (proximity, direction, three-dimensional objects, etc.), enabling successful action output.

5.3 Ablation Study

We separately remove the prompts and processes related to the three types of spatial cognitive abilities, and the navigation performance is listed in the Table 2. Each group exhibits varying degrees of performance degradation. The removal of viewpoint invariance has the most significant impact because the agent relies on viewpoint to determine its position and orientation within the environment. In specific cases, we observed that without this

Table 1: Results of zero-shot location-goal embodied navigation.

Method	Short			Middle			Long		
	SR/% ↑	SPL/% ↑	DTG/m ↓	SR/% ↑	SPL/% ↑	DTG/m ↓	SR/% ↑	SPL/% ↑	DTG/m ↓
Random	7.6	6.2	56.3	0	0	61.6	0	0	147.7
Action Sampling	15.3	12.8	40.3	10.4	8.9	50.5	0	0	64.7
AG-GPT4	20.6	18.2	59.2	9.2	8.0	103.7	0	0	162.6
NavGPT	22.4	18.3	47.4	15.6	12.5	50.1	7.7	5.3	131.5
ZSON	20.4	17.3	48.0	14.1	12.4	48.0	4.5	3.4	150.7
SayNav	19.8	18.7	52.1	12.1	11.0	55.7	5.8	5.2	100.6
EmbodiedNav	**52.1**	**45.9**	**26.0**	**41.1**	**35.8**	**39.1**	**33.5**	**25.6**	**69.7**

Table 2: Results of Ablation Study

Ablation	SR/% ↑	SPL/% ↑	DTG/m ↓
w/o Size Constancy	20.8	14.6	78.0
w/o Viewpoint Invariance	12.0	10.1	122.4
w/o Spatial Perspective-Taking	37.5	25.9	54.4

ability, the agent, whose goal was to navigate to Building A, would become disoriented and mistakenly identify Building B as the search area. The lack of size constancy manifests in failed cases as either being too far from the goal to observe it or being too close to building exteriors, resulting in a restricted field of view. Spatial perspective-taking is primarily reflected in the planning process; without this ability, errors include decreased logical coherence in planning, an expanded search area, and significantly longer search distances.

6 Related Work

6.1 Large Pre-trained Models in Embodied AI

The application of large pre-trained models in robotics has sparked widespread research interest due to its ability to enable robots to communicate and think like humans (Vemprala et al., 2023; Driess et al., 2023). Current research focuses on leveraging vision-language understanding and commonsense reasoning abilities of large pre-trained models to handle tasks like navigation (Lin et al., 2023; Shah et al., 2023), manipulation (Huang et al., 2023b), task planning (Zhao et al., 2024; Wang et al., 2023), and human-robot interaction (Qiao et al., 2023). In this work, we explore large pre-trained models to construct embodied agent for goal-oriented aerial navigation in urban spaces.

6.2 Route-Oriented Vision-Language Navigation

In this task, the agent must comprehend route instructions presented in text format (e.g., "Go towards a park bench, take a left at the stop sign, then stop at the trunk"), and subsequently integrate visual observations to execute the corresponding actions (Dorbala et al., 2022). Research (Shah et al., 2023; Schumann et al., 2024; Zhou et al., 2023a) utilize LLMs to extract landmarks and actions from the instructions. LLMs then proceed with sequence planning to determine the next action for the robot. This task bears some resemblance to location-goal navigation, as both involve the comprehension of textual instructions and the integration of visual information for navigation.

6.3 Zero-shot Object Navigation

This task is to navigate an agent to a specific goal object within an unknown environment (Gadre et al., 2023; Majumdar et al., 2022; Zhou et al., 2023b). Prior works (Majumdar et al., 2022; Gadre et al., 2023) recognize the object goal based on contrastive language-image pre-training model (CLIP) (Radford et al., 2021). In (Rajvanshi et al., 2023), LLMs directly generates high-dimensional action instructions, which are then combined with object grounding algorithms to achieve navigation. Existing studies primarily focus on indoor environments for ground robots. Urban low-altitude environments demand higher spatial perception and reasoning abilities from the agent.

7 Conclusion

This work takes a pioneering step for goal-oriented embodied navigation in urban spaces. We propose an LMM-empowered agent based on three spatial cognition mechanisms: size constancy, viewpoint invariance, and spatial perspective-taking to build its embodied capabilities, which help to overcome the critical challenges in urban spaces. The experiment results illustrate the effectiveness of the proposed agent from different perspectives.

Limitations

Our EmbodiedNav agent shows promising advancements in aerial navigation, but there are key limitations.

- The system's reliance on repeated interactions with a large-scale multimodal model results in high computational costs, hindering deployment on resource-constrained edge devices. Future work should focus on developing a lightweight version of the model to enable efficient deployment on edge devices.

- The agent's current use of discrete control commands limits precise object-oriented navigation, particularly in complex urban environments. Future research should enhance control algorithms for continuous control, allowing for more precise and adaptive navigation in real-world scenarios.

References

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. 2018. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

Maria Brucato, Andrea Frick, Stefan Pichelmann, Alina Nazareth, and Nora S Newcombe. 2023. Measuring spatial perspective taking: Analysis of four measures using item response theory. *Topics in Cognitive Science*, 15(1):46–74.

Vishnu Sashank Dorbala, Gunnar Sigurdsson, Robinson Piramuthu, Jesse Thomason, and Gaurav S Sukhatme. 2022. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Delaram Farzanfar, Hugo J Spiers, Morris Moscovitch, and R Shayna Rosenbaum. 2023. From cognitive maps to spatial schemas. *Nature Reviews Neuroscience*, 24(2):63–79.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. 2023. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181.

Chen Gao, Baining Zhao, Weichen Zhang, Jun Zhang, Jinzhu Mao, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, Xinlei Chen, and Yong Li. 2024. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. 2021. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):5721.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2023a. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. 2023b. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*.

Xiwen Liang, Liang Ma, Shanshan Guo, Jianhua Han, Hang Xu, Shikui Ma, and Xiaodan Liang. 2024. Cornav: Autonomous agent with self-corrected planning for zero-shot vision-and-language navigation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12538–12559.

Jinzhou Lin, Han Gao, Rongtao Xu, Changwei Wang, Li Guo, and Shibiao Xu. 2023. The development of llms for embodied navigation. *arXiv preprint arXiv:2311.00530*.

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

Xiangguo Liu, Qiuhuan Yuan, Guoying Wang, Yuan Bian, Feng Xu, and Yuguo Chen. 2023b. Drones delivering automated external defibrillators: A new strategy to improve the prognosis of out-of-hospital cardiac arrest. *Resuscitation*, 182:109669.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. 2022. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35:32340–32352.

Nora S Newcombe, Mary Hegarty, and David Uttal. 2023. Building a cognitive science of human variation: Individual differences in spatial navigation.

Hong Qiao, Ya-Xiong Wu, Shan-Lin Zhong, Pei-Jie Yin, and Jia-Hao Chen. 2023. Brain-inspired intelligent robotics: Theoretical analysis and systematic application. *Machine Intelligence Research*, 20(1):1–18.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from

natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Abhinav Rajvanshi, Karan Sikka, Xiao Lin, Bhoram Lee, Han-Pang Chiu, and Alvaro Velasquez. 2023. Saynav: Grounding large language models for dynamic planning to navigation in new environments. *arXiv preprint arXiv:2309.04077*.

Nathan B Roberts, Emily Ager, Thomas Leith, Isabel Lott, Marlee Mason-Maready, Tyler Nix, Adam Gottula, Nathaniel Hunt, and Christine Brent. 2023. Current summary of the evidence in drone-based emergency medical services care. *Resuscitation Plus*, 13:100347.

Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. 2024. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18924–18933.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2017. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*.

Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. Chatgpt for robotics: Design principles and model abilities. Technical Report MSR-TR-2023-8, Microsoft.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. 2023. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3554–3560. IEEE.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and Wang He. 2024. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2024. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36.

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

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. 2023b. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, pages 42829–42842. PMLR.

A Appendix

A.1 Details of Baselines

- **Random**. The drone randomly selects actions (e.g., forward, backward, upward, downward, left, and right) at each location until it meets the maximum number of steps or reaches the goal location. This approach effectively showcases the solution space's magnitude.
- **Action Sampling**. Action Sampling agents employ action sampling based on the action distribution derived from the proposed urban drone delivery dataset in the simulator, thus emphasizing the role of spatial chain planning.
- **Action Generation with GPT-4 (AG-GPT4)** (Vemprala et al., 2023). As one of the most powerful multimodal large models, *gpt-4o* can continuously receive instructions and RGBD observations as inputs and generates discrete drone commands. This approach is used to showcase the performance of directly applying multimodal large models and evaluate their embodied capabilities.
- **NavGPT** (Zhou et al., 2023a) is a purely LLM-based instruction following navigation agent by performing zero-shot sequential action prediction for navigation. This method primarily aims to show the performance of vision-language navigation algorithms when transferred to the location-goal problem.
- **ZSON** (Majumdar et al., 2022) presents a scalable zero-shot approach for open-world object-goal navigation using image-goal navigation training.
- **SayNav** (Rajvanshi et al., 2023) employs a novel grounding mechanism to build a 3D scene graph of the explored environment. This graph is used as input to LLMs to generate high-level navigation plans, which are then executed by a pre-trained low-level planner. ZSON and SayNav demonstrate the performance of indoor object navigation methods.