2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

NavSpace: How Navigation Agents Follow Spatial Intelligence Instructions

Haolin Yang*[12], Yuxing Long*[12], Zhuoyuan Yu[12], Zihan Yang[1], Minghan Wang[1], Jiapeng Xu[1], Yihan Wang[1], Ziyan Yu[1], Wenzhe Cai[3], Lei Kang[1], and Hao Dong†[12]
[1]CFCS, School of Computer Science, Peking University [2]PrimeBot [3]Shanghai AI Lab
*Equal contribution, † Corresponding author

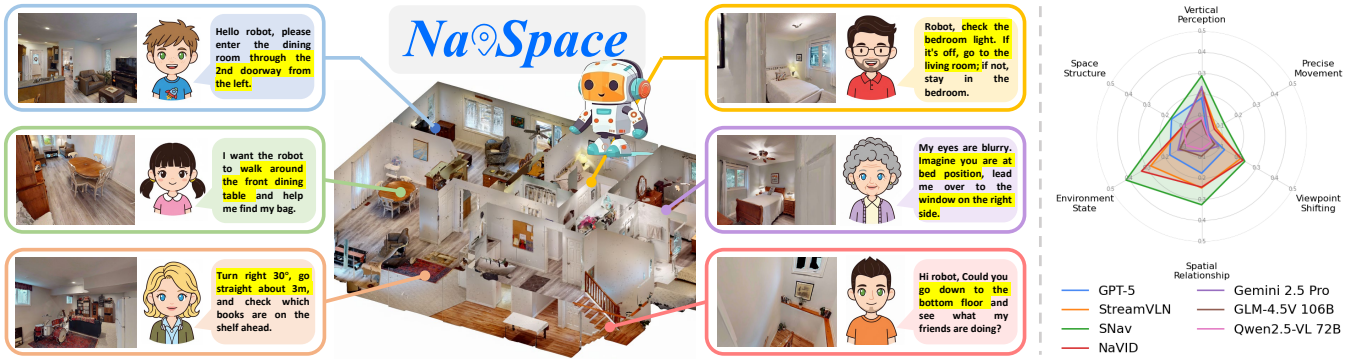

Fig. 1: **(Left) Everyday navigation instructions that require spatial intelligence.** To execute these instructions, a navigation agent must perceive and reason about space layout, scale, agent–object relative orientations, and environmental state. As the first benchmark to evaluate navigation agents' spatial intelligence, NavSpace collects navigation instructions covering the above six types of spatial-intelligence capabilities. **(Right) Evaluation results on NavSpace about navigation agents driven by multimodal large models and navigation models.** We further propose SNav model to serve as a strong baseline.

***Abstract*— Instruction-following navigation is a key step toward embodied intelligence. Prior benchmarks mainly focus on semantic understanding but overlook systematically evaluating navigation agents' spatial perception and reasoning capabilities. In this work, we introduce the NavSpace benchmark, which contains six task categories and 1,228 trajectory–instruction pairs designed to probe the spatial intelligence of navigation agents. On this benchmark, we comprehensively evaluate 22 navigation agents, including state-of-the-art navigation models and multimodal large language models. The evaluation results lift the veil on spatial intelligence in embodied navigation. Furthermore, we propose SNav, a new spatially intelligent navigation model. SNav outperforms existing navigation agents on NavSpace and real robot tests, establishing a strong baseline for future work.**

I. Introduction

Building navigation agents that can follow human instructions to move within environments is a key step toward realizing embodied intelligence. Owing to their user-friendly human–machine interaction, instruction navigation methods have been widely studied in recent years. Visual Language Navigation (VLN) tasks such as R2R [1], R4R [2], and RxR [3] require an agent to move to a specified location based on navigation actions and landmarks described in the instruction. Object Goal Navigation (ObjNav) [4] tasks require a robot to explore the environment and search for the target object named in the instruction. Demand Driven Navigation (DDN) [5] tasks present an abstract human need; the agent must understand that need and perform semantic reasoning to complete the navigation.

Although existing evaluation tasks have driven progress in instruction-following navigation, they concentrate on benchmarking agents' multimodal understanding of language and visual semantics and do not systematically assess spatial perception and reasoning. Yet, as illustrated in Figure 1, navigation tasks that demand spatial intelligence are common in everyday life. The navigation agent should accurately perceive spatial scales, subject–object spatial relations, and environmental structures, and correctly infer navigation actions. No prior benchmark has widely evaluated navigation agents' perceptual and reasoning abilities in space. Consequently, the spatial intelligence of both navigation models and multimodal large language models (MLLMs) on embodied navigation tasks remains unclear, and methods for improving these capabilities are underexplored.

Therefore, we introduce a novel benchmark, NavSpace. We begin by conducting a questionnaire survey to identify key categories of spatial intelligence essential for navigation tasks. The six most frequently selected categories include **Vertical Perception**, **Precise Movement**, **Viewpoint Shifting**, **Spatial Relationship**, **Environment State**, and **Space Structure**. To enable large-scale data collection for these categories, we design a large model assisted platform and a annotation pipeline: *Trajectory Collection*: annotators teleoperate agents to navigate within the photo-realistic scenes to record navigation trajectories; *Instruction Annotation*: annotators compose navigation instructions based on the requirements and the information analyzed by MLLM; and *Human Cross-Validation*: a separate annotator replays the trajectory to ensure instruction accuracy and consistency. Following this pipeline,

we collect a total of 1228 navigation trajectory-instruction pairs for NavSpace benchmark.

On NavSpace, we conducted a comprehensive evaluation of 22 existing navigation agents, covering lightweight navigation models, navigation large models, open-source MLLMs, and proprietary MLLMs. The evaluation included state-of-the-art instruction navigation models such as StreamVLN, as well as flagship MLLMs like GPT-5 and Gemini Pro 2.5. Through both quantitative and qualitative experiments, we derived several key insights: the importance of spatial intelligence benchmarks for navigation, the limitations of MLLMs in embodied navigation tasks, the advantages of navigation large models over lightweight ones, and promising directions for enhancing the spatial intelligence of navigation agents.

We further investigate methods for improving agents' spatial intelligence. In particular, we explore generating spatially intelligent navigation instructions from open-source datasets, and leveraging these instructions to inject spatial perception and reasoning capabilities into navigation models. Building on this approach, we propose SNav, a spatially intelligent navigation large model that serves as a strong baseline for NavSpace.

In this work, our main contributions are:

- We introduce the first spatial intelligence benchmark NavSpace for instruction navigation. NavSpace stems from questionnaire surveys and manually collects 1,228 high-quality trajectory-instruction pairs.
- On NavSpace benchmark, we comprehensively evaluate 22 navigation agents in total, which include navigation models and multimodal large language models. Several key insights are derived from the evaluation results.
- We propose SNav, a spatially intelligent navigation model, that surpasses existing models and establishes a strong baseline for NavSpace and real robot tests.

II. Related Work

A. Instruction Navigation Benchmarks

Since the emergence of the Visual Language Navigation (VLN) task, research on instruction-following navigation has proliferated. After R2R [1], subsequent works such as R4R [2] focused on models' ability to follow longer instructions, while RxR [3] examined the impact of multilingual instructions on navigation models. CVDN [6] shifted attention to human–model interaction via dialogue. Object Goal Navigation [4] emphasized models' ability to search for objects in indoor environments. In recent years, researchers have largely moved toward topics like human demand [5], crowded environment [7], and multimodal instructions [8]. However, spatial-perception intelligence, one fundamental capability of navigation models, has not yet been evaluated, compared, or analyzed by any benchmark for existing instruction-following models.

B. Navigation Large Models

Massive internet-scale multimodal data have significantly driven the development of multimodal large models. Pre-trained multimodal models such as GPT-5, Qwen2.5-VL [9], and LLaVA-Video [10] demonstrate strong capabilities in language understanding and visual perception. This has inspired researchers in the navigation field to fine-tune multimodal large models to build end-to-end navigation models. NaVid [11], NaVILA [12], and CorrectNav [13] train multimodal large models for the visual-language navigation task, while StreamVLN [14] and Uni-NaVid [15] further extend the instruction navigation task to object goal navigation. Although these models already possess basic instruction-following navigation capabilities, their performance on the NavSpace benchmark shows that when instructions primarily require spatial awareness of the scene, they fail to complete navigation tasks effectively. This indicates that the spatial intelligence of current large navigation models still needs improvement.

III. NavSpace Benchmark

A. Task Definition

The task definition of NavSpace follows classical instruction navigation tasks [16]. Given a language instruction L_{nav} from NavSpace, the navigation agent should predict the next navigation action $a_{t+1} \in A$ at time step t based on observation $\{O_1, O_2, \ldots, O_t\}$. If the agent chooses to stop, its distance to the destination must be below a predefined threshold.

B. Benchmark Construction

As shown in Figure 3, we design a four-stage pipeline to construct navigation trajectories and instructions for NavSpace benchmark.

Questionnaire Survey. We designed a two-part survey to identify which navigation instructions best reflect spatial intelligence. In part one, respondents read a detailed definition of spatial intelligence [17] and confirmed their comprehension; in part two, they were shown 17 candidate instruction types that might require spatial intelligence and asked to select up to six that best matched the definition and seemed reasonable. We collected 512 responses and, to ensure reliability, retained only 457 with completion times exceeding three minutes for analysis. The six most frequently selected categories were Vertical Perception, Precise Movement, Viewpoint Shifting, Spatial Relationship, Environment State, and Space Structure. We then collected navigation trajectories and instructions based on these categories. Each category is introduced in Section III-C.

Trajectory Collection. To collect navigation trajectories, we built a data-collection platform based on the Habitat 3.0 [18] simulator and HM3D [19] scenes. The system consists of a front-end annotation webpage and a back-end server that interfaces with the simulator and stores the data. After logging in, annotators teleoperate the agent with the keyboard while viewing first-person RGB observations. Annotation officially begins once the annotator has familiarized themself with the scene layout (after moving at least 200 steps). Before recording, the platform specifies the instruction category the annotator should follow. When the annotator clicks "Start Recording Trajectory" button, the platform records the agent's first-person RGB frames, navigation actions, and coordinates

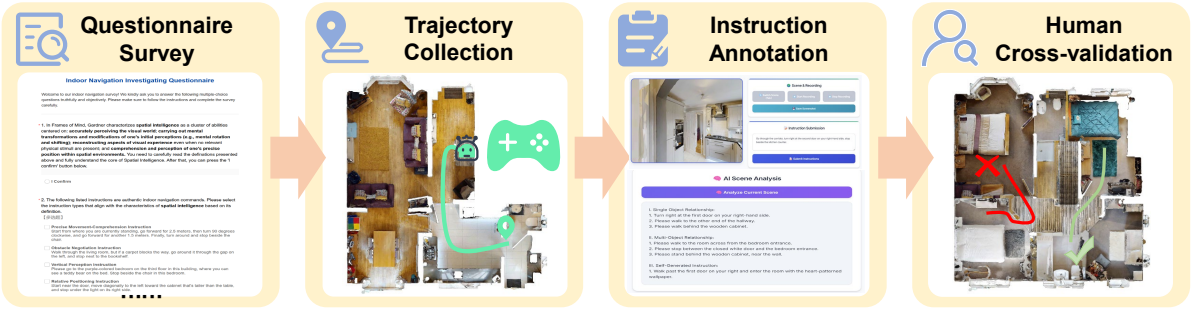

Fig. 2: **Construction pipeline of NavSpace.** (1) *Questionnaire Survey*: identify which forms of navigation instruction best reflect spatial intelligence. (2) *Trajectory Collection*: teleoperate agents in a simulated environment to record trajectories. (3) *Instruction Annotation*: use large-model–assisted analysis to create navigation instructions requiring spatial-intelligence. (4) *Human cross-validation*: manually review and validate the annotated instructions to ensure correctness and executability.

in real time; recording ends when the annotator clicks "Stop Recording Trajectory" button.

Instruction Annotation. After recording a complete navigation trajectory, the annotator can invoke GPT-5 to analyze the collected trajectory. MLLM's textual inputs include the target instruction type, the discrete navigation actions and position coordinates, and visual inputs consisting of the agent's first-person observations sampled along the trajectory. With this information, the MLLM analyzes the rooms, areas, and objects encountered and generates candidate navigation instructions for annotators to review. The human annotator must write the final navigation instructions following the annotation requirements.

Human Cross Validation. To ensure that annotated instructions are executable, we ask annotators to cross-validate them. Specifically, each instruction must be executed by a different annotator who has not seen it, remotely controlling an agent in Habitat to navigate. If the annotator successfully reaches the intended destination, the instruction is considered valid; otherwise, it is discarded and re-annotated.

Figure 4 visualizes the statistics about NavSpace.

C. NavSpace Instruction Categories

Vertical Perception. This category assesses the model's capability to determine its vertical position within indoor environments. These instructions may include explicit floor references tied to the building's structure, such as *"Go to the second floor, walk through the corridor, and stop by the bed in the bedroom at the end of the corridor."* This requires the model to identify the current floor and the target floor for effective route planning. Besides, instructions might use relative terms instead of concrete numbers, like *"Go to a higher floor, pass the sofa next to the staircase, and stop beside the television in the bedroom ahead."* The model must correctly interpret relative height changes to locate the target and success. In other cases, explicit numbers or relative terms may be omitted entirely, as in *"Go to the topmost floor and stop at the bedroom doorway next to the staircase."* or *"Stop halfway up the stairs beside the picture frame."* The challenge lies in the model's ability to infer vertical positioning from context (*e.g.*, "*topmost floor*," or "*halfway*"). Success is measured by arriving within 3.0 meters of the target location.

Precise Movement. This category tests an agent's ability to precisely understand the detailed distances and angles specified in the instruction and accurately interpret them into navigation actions. The agent should be aware of the space scales. For example, "*From the door, turn right 180°, go straight 1 m, turn left 90° and go 5 m, then turn 90° clockwise and go 7.5 m, then stop.*" The agent must correctly carry out each specified rotation and translation. Because the controller has no backward-action primitive, any "walk backward" instruction must be implemented by rotating 180° and moving forward. The success radius is defined as 1.0m.

Viewpoint Shifting. This category mainly tests a navigation agent's ability to switch viewpoints between subjects and objects. It requires the agent to possess spatial imagination and spatial transformation capabilities. Unlike previous work [20], NavSpace places extra emphasis on the long-term memory and history-aware reasoning: the agent must correctly reason over its entire movement history, even after many relocations. One typical instruction is *"Imagine you are the television in front of you. Move toward your front-left, follow the hallway to the end, and stop at the white door."* The agent must adopt the television's perspective, realize that the television's front-left corresponds to the agent's own right-hand side, and then navigate accordingly to the target. The success radius is defined as 2.0m.

Spatial Relationship. This category focuses on perceiving sequential order and relative spatial relationships among multiple objects or rooms. It may involve cross-room navigation with instructions like *"Walk down the hallway, turn left at the third door on your left, and stop next to the chair in the bedroom,"* which test counting and ordering skills. It also assesses spatial reasoning with multiple objects, such as *"Go downstairs to the living room and stop between the two brown sofas,"* which require identifying object locations and understanding inter-object relations to determine where to move or stop. Success is defined as arriving within a 2.0m radius of the target.

Environment State. This category requires the agent to accurately perceive environment states during navigation and make correct decisions about future actions based on those states. A representative format of this category is

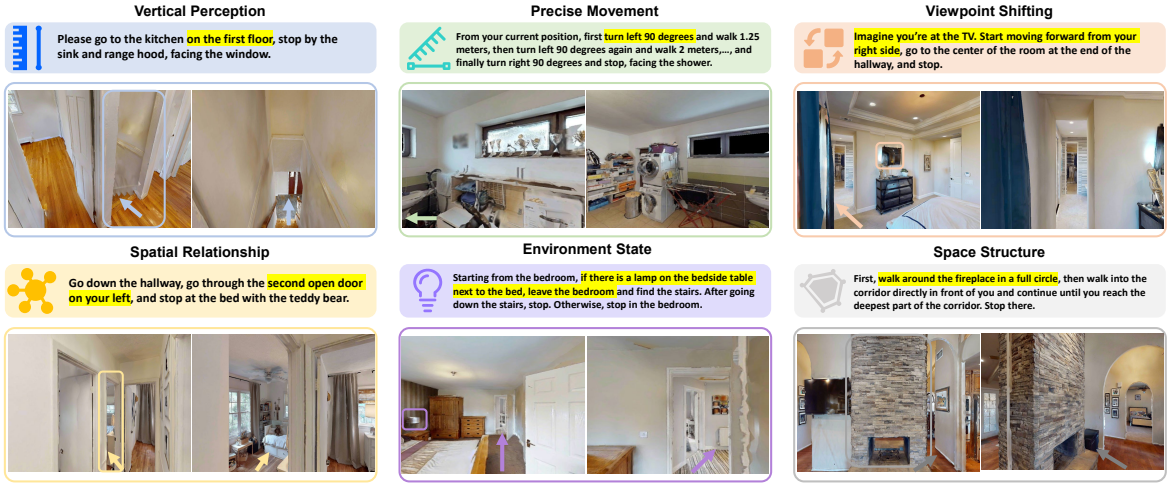

Fig. 3: **Instruction Categories in NavSpace.** These six categories were determined based on the questionnaire survey results. Every navigation trajectory and instruction was collected manually from HM3D scene datasets through our designed platform.

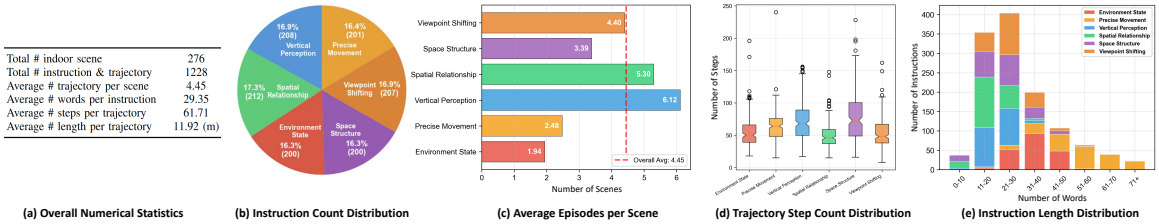

Fig. 4: **Visualization of NavSpace Statistics.**

"*if... otherwise...*". An example instruction is "*Walk through the hallway to the foyer and wait beside the storage cabinet; if you see the keys, stop, otherwise go to the front door and check.*" The success radius is defined as 2.0m.

Space Structure. This category needs the agent to understand the spatial layout and perform navigation behaviors following the instructions, such as circling, making round trips, and moving to locations at distance extremes. For example, instructions may require circling an object for a whole round, such as *"Walk around the eight-person dining table once"* to assess the model's ability to grasp an object's dimensions and shapes. Others demand back-and-forth paths, like *"Go to the sofa in the room at the end of the hallway and then return,"* testing return navigation. Still others identify extreme locations (*e.g.*, nearest or farthest), as in *"Go upstairs to the room on your right and stop by the farthest sofa"*. Success is reaching within 1.0m of the target.

IV. SNav Model

A. Model Details

The architecture of the SNav incorporates three fundamental components: the Vision Encoder $v(\cdot)$, the Projector $p(\cdot)$, and the Large Language Model (LLM) $f(\cdot)$. In processing an RGB video input, the Vision Encoder generates visual feature representations from sampled frames, denoted as $Z_v = v(\{I_1, I_2 ... I_t\})$. These representations are subsequently transformed by the MLP Projector into the LLM's semantic space, yielding a sequence of visual tokens $H_v = p(Z_v)$. The LLM $f(\cdot)$ then performs auto-regressive predictions by integrating these visual tokens H_v with texual tokens X, which are derived from the task instruction L. For implementation, SigLIP [21] serves as the Vision Encoder, a 2-layer MLP [22] functions as the Projector, and Qwen2 [23] acts as the LLM.

The SNav model is initialized from LLaVA-Video 7B [10]. Then we follow the previous work [13] to conduct navigation finetuning through co-training with three tasks. These tasks include Navigation Action Prediction, Trajectory-based Instruction Generation, and General Multimodal Data Recall. After this, we obtain the vanilla SNav model.

B. Spatial Intelligence Enhancement

To improve the spatial intelligence of the vanilla SNav model, we designed pipelines for creating navigation data that require spatial perception and reasoning (Figure 5 Left) and finetune vanilla SNav with these instruction-trajectory pairs (Figure 5 Right). Data creation pipelines are detailed in the following.

Cross-floor Navigation. We select the R2R trajectories that are likely to cross floors by thresholding the height difference between the start and end. For each selected trajectory, we place the agent at the start position in the Habitat and follow a shortest-path planner to the goal while recording RGB camera observations. We label a trajectory as floor-crossing if GPT-5 detects stairs in at least three recorded frames. Following the floor-segmentation method HOV-SG [24], we assign floor labels to the start and end points, and combine these with Habitat's total-floor count to produce vertical-space annotations. With these annotations, GPT-5 can restyle raw instructions like "*Walk up the stairs ...*" into "*Walk up to the top floor ...*" or "*Walk up to the third floor ...*"

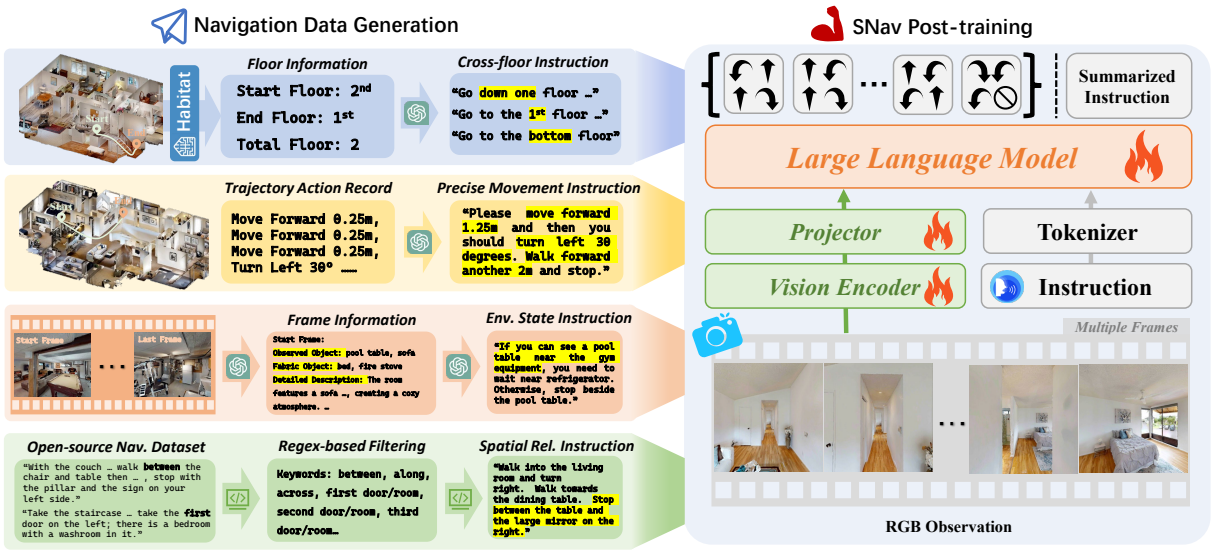

Fig. 5: **Framework of SNav model.** (Left) We propose a set of pipelines to create 4 types of spatially intelligent navigation instructions from existing scene data and instruction navigation data. (Right) With these generated data, we further finetune an end-to-end navigation foundation model to obtain a navigation large model SNav with enhanced spatial intelligence.

Precise Movement. We randomly sample start and goal points in MP3D scenes and use the shortest-path planner in the Habitat simulator to compute a path. After filtering trajectory steps, we can obtain trajectories of the desired length (*e.g.*, 20–60 steps). We follow each path in the simulator and record the discrete navigation actions (*i.e.*, turn left 30°, turn right 30°, move forward 0.25m, and stop) . By merging consecutive actions of the same type, we produce concise movement descriptions such as "*move forward 3 m, turn right 60°, move forward 2 m*". Finally, GPT-5 paraphrases these movement descriptions into natural-language navigation instructions. For example, "*Please walk forward 3 meters first, then turn right 60°, and then continue forward 2 meters*".

Environment State Inference. We first extract start–end point pairs and their corresponding navigation instructions from the R2R dataset. For each pair, we use a shortest-path planner to generate the trajectory and save the RGB frames observed along that path. We then query GPT-5 with the first and last frames of each trajectory to infer three information: observable objects, unobservable objects, and a detailed description. Given that this category of instructions often follows an "*if…otherwise…*" structure, we design a group of templates that combine these multimodal observations with the original instructions to create new instructions. Two template patterns are: (1) "*Original_instruction; if [visible_object in last frame] then stop at [last-frame stop location], otherwise go to [scene description inferred from first frame]*", and (2) "*If [fabricated_object detected in first frame] then stop where you are, otherwise follow Original_instruction and stop at [last-frame stop location].*" We instantiate five template categories covering the common if/otherwise cases, and use GPT-5 to rewrite all instructions according to these templates to generate our training data.

Spatial Relationship. We applied regular expressions to the instructions in the R2R dataset to select those containing ordinal phrases (*e.g.*, "*first room*", "*first door*", "*second room*", "*second door*", "*third room*", "*third door*"). We also identified instructions that express multi-object relations by searching for words such as "*between*", "*along*", and "*across*".

V. Experiments

A. Evaluation Setup

Environment and Metrics. NavSpace takes Habitat 3.0 [18] as the simulator to conduct the evaluation. Evaluation scenes are selected from the HM3D datasets. At each step, the agent can only select one action: move forward 0.25m, turn left 30°, turn right 30°, or stop. Following previous instruction navigation benchmarks, we employ the following widely used evaluation metrics: Navigation Error (NE), Oracle Success Rate (OS), Success Rate (SR).

Baseline Models. We conduct a comprehensive evaluation of existing multimodal large models and navigation models. These models can be categorized into the following five types.

- **Chance Level Baselines**: Chance Level (Random) is the peformance from random guessing among four navigation actions (25% for each). Chance Level (Frequency) refers to performing navigation actions based on the action occurrence frequencies observed in the trajectories of the NavSpace benchmark.
- **Open-source MLLMs**: We selected the multimodal large language models Qwen2.5-VL [9] and LLaVA-Video [10], which are widely used as backbone models in navigation. Besides, we also test GLM-4.5V and GLM-4.1V-Thinking [25] released recently.
- **Proprietary MLLMs**: We chose the latest GPT models recently released by OpenAI, including GPT-5 and GPT-5 Mini, as well as the previous-generation flagship model GPT-4o. In addition, we also tested Google's latest models, Gemini 2.5 Pro and Gemini 2.5 Flash.
- **Lightweight Navigation Models**: The lightweight navigation models we selected include waypoint predictor-based models, such as BEVBert [30] and ETPNav [29],

TABLE I: **Quantative performances on NavSpace.**

	Vertical Perception			Precise Movement			Viewpoint Shifting			Spatial Relationship			Environment State			Space Structure			Average		
	NE ↓	OS ↑	SR ↑	NE ↓	OS ↑	SR ↑	NE ↓	OS ↑	SR ↑	NE ↓	OS ↑	SR ↑	NE ↓	OS ↑	SR ↑	NE ↓	OS ↑	SR ↑	NE ↓	OS ↑	SR ↑
Chance Level Baselines																					
Chance Level (Random)	6.92	0.144	0.043	7.23	0.075	0.010	6.65	0.126	0.039	7.00	0.057	0.042	5.52	0.145	0.060	5.22	0.255	0.055	6.42	0.134	0.042
Chance Level (Frequency)	6.90	0.221	0.115	6.94	0.129	0.035	6.23	0.232	0.116	7.09	0.160	0.075	5.50	0.230	0.065	5.54	0.325	0.090	6.37	0.216	0.083
Open-source MLLMs																					
LLaVA-Video 7B [10]	6.09	0.139	0.077	6.36	0.119	0.035	6.59	0.092	0.068	6.02	0.165	0.113	6.59	0.090	0.065	5.07	0.310	0.045	6.12	0.153	0.067
GLM-4.1V-Thinking 9B [25]	6.85	0.173	0.077	6.35	0.095	0.020	6.43	0.135	0.082	5.72	0.198	0.113	5.12	0.205	0.070	5.33	0.265	0.030	5.97	0.179	0.065
GLM-4.5V 106B [25]	6.75	0.207	0.077	6.39	0.095	0.025	6.50	0.164	0.072	5.86	0.198	0.094	4.90	0.240	0.120	5.21	0.290	0.065	5.94	0.199	0.076
Qwen2.5-VL 7B [9]	6.29	0.111	0.063	5.96	0.109	0.025	6.29	0.082	0.077	5.44	0.142	0.094	5.20	0.195	0.085	4.77	0.305	0.105	5.66	0.157	0.075
Qwen2.5-VL 72B [9]	6.56	0.120	0.091	6.42	0.095	0.030	6.32	0.135	0.053	5.85	0.132	0.061	5.08	0.160	0.085	5.02	0.300	0.100	5.88	0.157	0.070
Proprietary MLLMs																					
GPT-4o	6.04	0.163	0.101	6.50	0.114	0.040	6.65	0.077	0.039	5.43	0.123	0.099	5.27	0.110	0.085	4.66	0.300	0.095	5.76	0.148	0.077
GPT-5 Mini	5.81	0.197	0.154	6.31	0.095	0.040	6.44	0.106	0.058	5.81	0.203	0.123	4.91	0.270	0.140	4.65	0.355	0.140	5.66	0.204	0.109
GPT-5	5.47	0.226	0.183	5.69	0.124	0.030	5.82	0.145	0.126	5.06	0.189	0.175	4.39	0.220	0.175	3.73	0.310	0.165	5.03	0.202	0.142
Gemini 2.5 Flash	6.30	0.115	0.038	6.32	0.114	0.040	6.51	0.106	0.048	5.55	0.099	0.075	4.82	0.170	0.115	4.73	0.265	0.075	5.71	0.145	0.065
Gemini 2.5 Pro	5.42	0.303	0.236	5.09	0.124	0.040	5.67	0.126	0.092	5.43	0.080	0.071	4.50	0.155	0.130	**3.99**	0.245	0.100	5.02	0.172	0.112
Lightweight Nav Models																					
Seq2Seq [16]	7.88	0.029	0.010	6.85	0.129	0.000	7.12	0.106	0.000	6.88	0.075	0.014	6.22	0.130	0.015	5.25	0.365	0.005	6.70	0.139	0.007
CMA [26]	6.60	0.019	0.005	5.56	0.134	0.000	5.81	0.135	0.014	6.12	0.123	0.028	5.42	0.175	0.055	5.11	0.390	0.005	5.77	0.163	0.018
HPN+DN [27]	6.62	0.106	0.087	5.59	0.154	0.035	5.04	0.174	0.106	4.97	0.142	0.113	4.68	0.210	0.130	5.28	0.110	0.040	5.36	0.149	0.085
VLN⟳BERT [26]	6.57	0.005	0.005	7.30	0.065	0.015	6.58	0.082	0.034	7.36	0.014	0.000	5.42	0.075	0.040	4.69	0.310	0.135	6.32	0.092	0.038
Sim2Sim [28]	6.72	0.005	0.005	7.46	0.060	0.060	6.73	0.087	0.087	7.45	0.009	0.000	5.64	0.070	0.070	4.86	0.310	0.165	6.48	0.090	0.065
ETPNav [29]	6.98	0.067	0.034	7.70	0.100	0.025	6.66	0.121	0.048	6.32	0.094	0.033	5.15	0.240	0.090	5.64	0.240	0.025	6.41	0.144	0.043
BEVBert [30]	6.60	0.082	0.043	6.33	0.070	0.020	6.30	0.159	0.072	6.14	0.094	0.038	5.41	0.195	0.065	5.28	0.265	0.040	6.01	0.144	0.046
Navigation Large Models																					
NaVid [11]	5.56	0.317	0.231	5.83	**0.219**	0.070	**4.97**	0.266	0.227	4.98	0.311	0.241	3.47	0.430	0.330	4.28	0.300	0.100	4.85	0.307	0.200
NaVILA [12]	6.71	0.038	0.034	7.26	0.025	0.025	6.64	0.063	0.053	6.73	0.066	0.038	5.58	0.130	0.080	5.09	0.205	0.130	6.34	0.088	0.060
StreamVLN [14]	6.00	0.351	0.231	5.59	0.189	0.080	5.42	0.271	0.213	5.02	0.311	0.245	3.88	0.375	0.280	4.44	0.355	0.100	5.06	0.309	0.192
SNav (Ours)	**5.30**	**0.365**	**0.288**	**4.68**	0.199	**0.124**	5.03	**0.304**	**0.237**	**4.47**	**0.354**	**0.325**	**3.17**	**0.520**	**0.415**	4.17	**0.460**	**0.170**	**4.47**	**0.367**	**0.260**
- Cross-floor Navigation	5.61	0.313	0.240	4.50	0.169	0.080	5.48	0.285	0.213	4.40	0.387	0.340	3.48	0.435	0.345	4.79	0.420	0.125	4.71	0.335	0.224
- Environment State	5.86	0.269	0.178	4.99	0.194	0.080	5.40	0.290	0.237	4.76	0.349	0.302	3.66	0.460	0.260	4.93	0.360	0.100	4.93	0.320	0.193
- Precise Movement	5.93	0.274	0.188	4.99	0.194	0.080	5.42	0.304	0.227	4.49	0.363	0.302	3.35	0.495	0.375	4.99	0.375	0.090	4.86	0.334	0.210
- Spatial Relationship	6.01	0.226	0.159	4.87	0.144	0.065	5.45	0.261	0.198	5.06	0.283	0.241	3.20	0.495	0.380	5.13	0.310	0.060	4.95	0.287	0.184

as well as waypoint predictor-free models like CMA [26] and Seq2Seq [16]. The model parameters of these lightweight navigation models are less than 100M. They usually only complete one type of instruction navigation task, like VLN.

- **Navigation Large Models**: The navigation large models that have been open-sourced so far include NaVid [11], NaVILA [12], and StreamVLN [14]. They are all 7B-parameter multimodal models fine-tuned for instruction navigation tasks. They predict navigation actions end-to-end from solely past RGB observations.

B. Performances on NavSpace

Multimodal Large Language Models. From Table I, NavSpace is extremely challenging for Open Source MLLMs. The average success rate of all open-source MLLMs falls below 10%, which is similar to Chance Level (Frequency). Proprietary MLLMs generally outperform Open Source MLLMs. Among the Proprietary MLLMs, GPT-5 demonstrates significantly better performance than other models. However, overall, the average success rate of all Proprietary MLLMs is still below 20%. This suggests that existing MLLMs are hardly capable of serving as navigation agents for spatial intelligence navigation tasks.

Navigation Models. From the evaluation results, lightweight navigation models such as BEVBert and ETPNav are almost incapable of executing navigation instructions that require spatial intelligence. The navigation large language model shows better performance on NavSpace compared to lightweight navigation models. Existing navigation large models, such as NaVid and StreamVLN, surpass GPT-5 in terms of average success rate, and have preliminarily demonstrated spatial intelligence capabilities for navigation.

SNav. As shown in Table I, our model SNav outperforms powerful navigation models (*i.e.*, StreamVLN and NaVid) and state-of-the-art MLLMs (*i.e.*, GPT-5 and Gemini 2.5 Pro) on the NavSpace benchmark, serving as a strong baseline model. Ablation study at the bottom of Table I demonstrates that our proposed instruction-generation pipelines help SNav improve the spatial intelligence.

TABLE II: **Real world experiment results.**

	Precise Movement	Viewpoint Shifting	Spatial Relationship	Environment State	Space Structure	Average
NaVILA	0/10	0/10	1/10	0/10	2/10	6%
NaVid	1/10	2/10	2/10	1/10	1/10	14%
SNav	**3/10**	**4/10**	**4/10**	**1/10**	**4/10**	**32%**

C. Real World Test

In the real-world test, we compare our method against two leading navigation large models, NaVid and NaVILA, across office, campus, and outdoor environments. The test covers five categories of spatially intelligent navigation instructions (excluding vertical perception). Our experimental platform is the AgiBot Lingxi D1 quadruped, which is equipped with a monocular RGB camera and motion-control APIs. Upon receiving a navigation instruction, the robot transmits the RGB observation to the navigation model hosted on a remote server with an NVIDIA A100 GPU; the model then predicts actions and calls the D1 motion API to execute them. Real-robot results are summarized in Table II and demonstrations are shown in Figure 6.

VI. Discussions

Do existing spatial intelligence benchmarks truly reflect a model's capability in embodied navigation? We observed a clear phenomenon: MLLMs (*i.e.*, LLaVA-Video, Qwen2.5-VL, and GPT-4o) that perform reasonably well on existing spatial intelligence benchmarks such as VSI-Bench [31], SpatialBench [32], and MindCube [20] are almost unable to complete the navigation tasks in NavSpace. This may be because the existing benchmarks are static evaluations, where a model only needs to predict a deterministic numerical answer or choose from multiple options based on the given

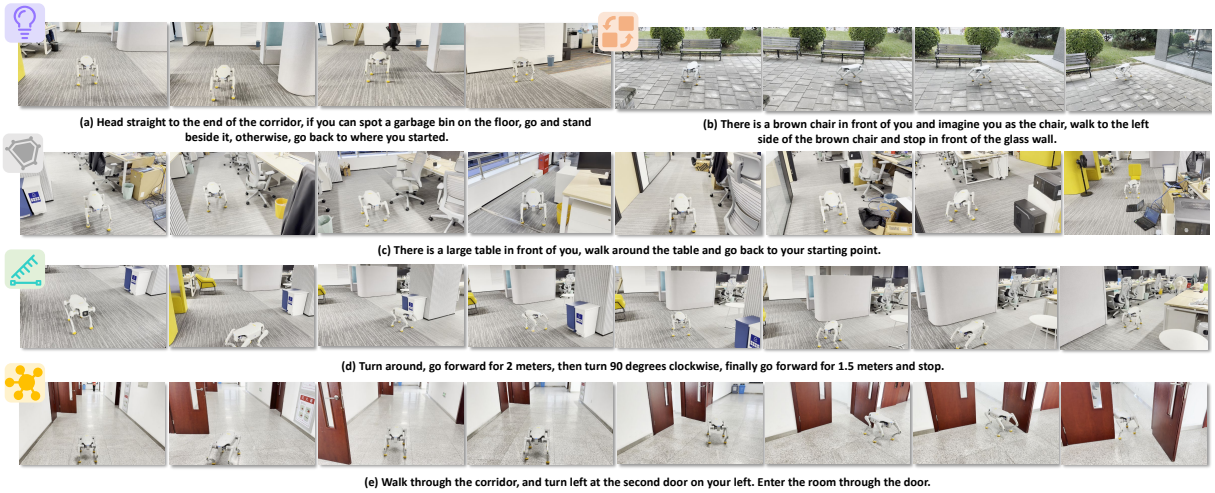

Fig. 6: **Qualitative results from the real-world deployment of SNav.** The evaluated instructions cover five categories proposed in NavSpace. The test environment includes the office, the campus building, and the outdoor area.

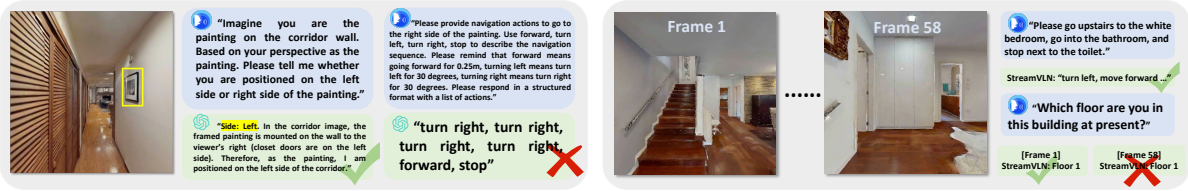

Fig. 7: **Case study about GPT-5 (Left) and StreamVLN (Right) on NavSpace.**

observations. In contrast, our benchmark requires the model to take dynamic actions in the scene based on its spatial perception and reasoning. For embodied tasks, translating spatial perception into precise movement is more important than one-off perceptual judgments. Therefore, our benchmark better captures the core demands of embodied navigation.

Do current MLLMs demonstrate emergent spatial intelligence for embodied navigation? To investigate why MLLMs perform poorly on NavSpace, we query GPT-5 with questions requiring spatial intelligence to re-perceive its erroneous trajectories. From case analysis, we found that GPT-5 sometimes can correctly answer questions about precise distance, viewpoint shift, or environmental state. However, when it predicts concrete navigation actions, the actual actions are inconsistent with its initial perception. One example of viewpoint shift is shown in Figure 7 (Left). As actions are executed, GPT-5's intermediate perceptions also sometimes contradict its original observations. Overall, beyond limitations in spatial reasoning, errors in reasoning from perception to action, and inconsistencies in perception across multiple frames are the main causes of MLLM's low success rate on NavSpace. These findings indicate that even the current flagship MLLMs have not yet demonstrated emergent spatial intelligence in embodied navigation.

Can lightweight navigation models effectively execute spatial intelligence navigation instructions? Although lightweight navigation models show competitive performance on certain instruction navigation tasks, they poorly generalize to NavSpace. Case analysis shows they tend to latch onto objects and actions mentioned in the NavSpace instructions and perform only shallow semantic-to-action inference. Perhaps this semantic-to-action mapping can work for some VLN tasks, but it fails to succeed on NavSpace. We also found that, although lightweight navigation models like BEVBert and ETPNav outperform NaVid and StreamVLN on VLN tasks, their success rates on NavSpace are far lower than theirs, which further indicates that lightweight navigation models do not truly understand spatial relations during navigation.

How to improve the spatial intelligence of navigation models? Benefiting from pretraining on internet-scale image–text corpora and open-source instruction navigation data, navigation large models can already satisfy a nontrivial fraction of the benchmark instructions. However, under a Q&A–based evaluation conducted along ground-truth trajectories, we observe that NaVid and StreamVLN perform poorly on questions about spatial scale, floor-level relations, and spatial structure. One example is shown in Figure 7 (Right). We hypothesize that these models primarily rely on general multimodal understanding and instruction-following capabilities and only incidentally succeed on a subset of spatial navigation instructions. This hypothesis is supported by their markedly worse performance on the Precise Movement and Space Structure instructions that depend less on visual semantic cues and more on spatial reasoning. Accordingly, future work should pursue, in parallel, (1) substantial improvements in spatial perception and (2) enhanced inferential mechanisms that translate spatial perception into action decisions.

VII. Acknowlegment

This work was supported by the National Natural Science Foundation of China (62136001) and the National Youth Talent Support Program (8200800081).

REFERENCES

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] V. Jain, G. Magalhaes, A. Ku, A. Vaswani, E. Ie, and J. Baldridge, "Stay on the path: Instruction fidelity in vision-and-language navigation," *arXiv preprint arXiv:1905.12255*, 2019.

[3] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding," in *EMNLP*, 2020.

[4] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *NeurIPS*, 2020.

[5] H. Wang, A. G. H. Chen, X. Li, M. Wu, and H. Dong, "Find what you want: Learning demand-conditioned object attribute space for demand-driven navigation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 16 353–16 366, 2023.

[6] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *Conference on Robot Learning*. PMLR, 2020, pp. 394–406.

[7] H. Su, F. Song, C. Ma, W. Wu, and J. Yan, "Robosense: Large-scale dataset and benchmark for egocentric robot perception and navigation in crowded and unstructured environments," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 446–27 455.

[8] C. Gao, L. Jin, X. Peng, J. Zhang, Y. Deng, A. Li, H. Wang, and S. Liu, "Octonav: Towards generalist embodied navigation," *arXiv preprint arXiv:2506.09839*, 2025.

[9] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[10] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," 2024. [Online]. Available: https://arxiv.org/abs/2410.02713

[11] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and W. He, "Navid: Video-based vlm plans the next step for vision-and-language navigation," in *RSS*, 2024.

[12] A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," *arXiv preprint arXiv:2412.04453*, 2024.

[13] Z. Yu, Y. Long, Z. Yang, C. Zeng, H. Fan, J. Zhang, and H. Dong, "Correctnav: Self-correction flywheel empowers vision-language-action navigation model," 2025. [Online]. Available: https://arxiv.org/abs/2508.10416

[14] M. Wei, C. Wan, X. Yu, T. Wang, Y. Yang, X. Mao, C. Zhu, W. Cai, H. Wang, Y. Chen, X. Liu, and J. Pang, "Streamvln: Streaming vision-and-language navigation via slowfast context modeling," 2025. [Online]. Available: https://arxiv.org/abs/2507.05240

[15] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," 2025. [Online]. Available: https://arxiv.org/abs/2412.06224

[16] J. Krantz, E. Wijmans, A. Majundar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision and language navigation in continuous environments," in *ECCV*, 2020.

[17] H. Gardner, *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, 2011. [Online]. Available: https://books.google.com.sg/books?id=wxj6npSaykgC

[18] X. Puig, E. Undersander, A. Szot, M. D. Cote, R. Partsey, J. Yang, R. Desai, A. W. Clegg, M. Hlavac, T. Min, T. Gervet, V. Vondruš, V.-P. Berges, J. Turner, O. Maksymets, Z. Kira, M. Kalakrishnan, J. Malik, D. S. Chaplot, U. Jain, D. Batra, A. Rai, and R. Mottaghi, "Habitat 3.0: A co-habitat for humans, avatars and robots," 2023.

[19] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra, "Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. [Online]. Available: https://arxiv.org/abs/2109.08238

[20] B. Yin, Q. Wang, P. Zhang, J. Zhang, K. Wang, Z. Wang, J. Zhang, K. Chandrasegaran, H. Liu, R. Krishna *et al.*, "Spatial mental modeling from limited views," *arXiv preprint arXiv:2506.21458*, 2025.

[21] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *ICCV*, 2023.

[22] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024.

[23] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan, "Qwen2 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2407.10671

[24] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[25] V. Team, W. Hong, W. Yu, X. Gu, G. Wang, G. Gan, H. Tang, J. Cheng, J. Qi, J. Ji, L. Pan, S. Duan, W. Wang, Y. Wang, Y. Cheng, Z. He, Z. Su, Z. Yang, Z. Pan, A. Zeng, B. Wang, B. Chen, B. Shi, C. Pang, C. Zhang, D. Yin, F. Yang, G. Chen, J. Xu, J. Zhu, J. Chen, J. Chen, J. Chen, J. Lin, J. Wang, J. Chen, L. Lei, L. Gong, L. Pan, M. Liu, M. Xu, M. Zhang, Q. Zheng, S. Yang, S. Zhong, S. Huang, S. Zhao, S. Xue, S. Tu, S. Meng, T. Zhang, T. Luo, T. Hao, T. Tong, W. Li, W. Jia, X. Liu, X. Zhang, X. Lyu, X. Fan, X. Huang, Y. Wang, Y. Xue, Y. Wang, Y. Wang, Y. An, Y. Du, Y. Shi, Y. Huang, Y. Niu, Y. Wang, Y. Yue, Y. Li, Y. Zhang, Y. Wang, Y. Wang, Y. Zhang, Z. Xue, Z. Hou, Z. Du, Z. Wang, P. Zhang, D. Liu, B. Xu, J. Li, M. Huang, Y. Dong, and J. Tang, "Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2507.01006

[26] Y. Hong, Z. Wang, Q. Wu, and S. Gould, "Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation," in *CVPR*, 2022.

[27] J. Krantz, A. Gokaslan, D. Batra, S. Lee, and O. Maksymets, "Waypoint models for instruction-guided navigation in continuous environments," in *CVPR*, 2021.

[28] J. Krantz and S. Lee, "Sim-2-sim transfer for vision-and-language navigation in continuous environments," in *ECCV*, 2022.

[29] D. An, H. Wang, W. Wang, Z. Wang, Y. Huang, K. He, and L. Wang, "Etpnav: Evolving topological planning for vision-language navigation in continuous environments," *IEEE TPAMI*, 2024.

[30] D. An, Y. Qi, Y. Li, Y. Huang, L. Wang, T. Tan, and J. Shao, "Bevbert: Multimodal map pre-training for language-guided navigation," in *ICCV*, 2023.

[31] J. Yang, S. Yang, A. W. Gupta, R. Han, L. Fei-Fei, and S. Xie, "Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces," *arXiv preprint arXiv:2412.14171*, 2024.

[32] W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9490–9498.