Recent Advances in Robotic Navigation via Large Language Models

Authors: Haotian Pan, Shibo Huang, Jian Yang, Jinpeng Mi, Ke Li, Xiong You, Xuan Tang, Peidong Liang, Jinbo Yang, Yingjie Liu, Jianfeng Zhang, Muyu Wang, Jie Yang, Xinyu Zhang, Lijun Zhao, Mingsong Chen, Jie Zhou, Fiber, Jian Yang, Jinpeng Mi, Xuan Tang, Jie Zhou, Xian Wei

Date: 2025-03-06T00:00:00+00:00

Abstract

Recently, with the advances in Large Language Models (LLMs), robot navigation models have demonstrated superior generalization capabilities, including environment perception, decision-making, reasoning, planning, instruction understanding, and human-robot interaction.In this paper, we systematically review recent LLM-based robot navigation research papers, categorizing existing studies into a novel taxonomy comprising perception, planning, control, interaction, and coordination.We also present an overview of the principal datasets and metrics used in robot navigation, analyzing the distinctive characteristics of the datasets and the performance of the main LLMs-based methods.Furthermore, we discuss the challenges hindering the integration of LLMs into robot navigation and provide opportunities and potential directions for future development.

Full Text

Preamble

Recent Advances in Robotic Navigation via Large Language Models
Haotian Pan[1], Shibo Huang[1], Jian Yang[2], Jinpeng Mi[3], Ke Li[2], Xiong You[2], Xuan Tang[1], Peidong Liang[4], Jinbo Yang[3], Yingjie Liu[1], Jianfeng Zhang[1], Muyu Wang[1], Jie Yang[1], Xinyu Zhang[1], Lijun Zhao[5], Mingsong Chen[1], Jie Zhou[6], Xian Wei[1] ∗†
March 6, 2025

Abstract

Recent advances in Large Language Models (LLMs) have demonstrated superior generalization capabilities in robot navigation, encompassing environment per-

ception, decision-making, reasoning, planning, instruction understanding, and human-robot interaction. In this paper, we systematically review recent LLM-based robot navigation research, categorizing existing studies into a novel taxonomy comprising perception, planning, control, interaction, and coordination. We present an overview of the principal datasets and metrics used in robot navigation, analyzing the distinctive characteristics of the datasets and the performance of the main LLM-based methods. Furthermore, we discuss the challenges hindering the integration of LLMs into robot navigation and provide opportunities and potential directions for future development.

Keywords: Large Language Models (LLMs), robot navigation, environment perception, decision-making, reasoning, planning, instruction understanding

Introduction

Robot navigation refers to a robot' s capability to identify its location within its environment and plan a path to reach a target destination. It is a multidisciplinary field encompassing artificial intelligence, machine learning, computer vision, sensor technology, and robotics. Traditionally, robot navigation has been regarded as a geometric mapping and planning problem [**?**], requiring robots to parameterize geometric problems to identify and plan paths from a starting point to a destination in known or unknown environments.

From early model-based approaches to recent advancements in deep learning and reinforcement learning, significant progress has been made in robot navigation technology [**?**, **?**, **?**, **?**, **?**]. For example, Leonard et al. [**?**] utilized the Extended Kalman Filter for mobile robot navigation in known environments, while Hu et al. [**?**] employed landmark identification and recognition of dynamically extracted environmental features for navigation. As technology advanced, researchers began incorporating machine learning to transcend simple geometric abstractions, enabling systems to make decisions based on real-world experiences, consider the physical consequences of actions, and leverage patterns in the environment [**?**]. However, these methods are data-intensive and lack interpretability, making debugging and improvement challenging. Consequently, many machine learning-based methods remain predominantly in simulation, with only occasional application in simple real-world environments as "concept validation" systems for learning navigation [**?**].

Recently, LLM-based robot navigation methods have garnered significant attention. LLMs such as GPT-3 [**?**] and BERT [**?**], pre-trained on vast textual data, learn rich language patterns that enable them to perform various language tasks with minimal examples. These models enhance embodied intelligent systems with environmental awareness and decision support. Leveraging their powerful language and image processing capabilities, LLMs can effectively plan and make decisions for new tasks with minimal or even zero sample data. LLMs also enhance human-machine interaction; for instance, the LIM2N [**?**] framework enables language and hand-drawn inputs to serve as navigation con-

straints and control objectives. Furthermore, the concept of generative agents—computational software agents simulating human behavior using LLMs—offers new perspectives for improving human-machine interaction [**?**].

In summary, applying LLMs to robot navigation is a promising research direction, yet the field faces numerous challenges: how to effectively encode environmental information into text, enable robots to understand and process complex environmental information, facilitate rational decision-making, improve human-robot interaction, and achieve autonomous decision-making and reasoning.

To comprehensively understand LLM-based navigation technology and advance further research, this paper summarizes the latest advancements and discusses future directions. Notably, recent surveys [**?**, **?**] have reviewed LLM-based navigation research. Compared to these works, this paper differs in three key aspects:

- This study focuses specifically on LLM-based navigation, which plays a pivotal role in advancing this technology.
- This paper examines the role of LLMs across various navigation aspects: Perception, Planning, Control, Interaction, and Coordination.
- LLM-based navigation methods are classified based on the physical environments where robot navigation tasks are applied: indoor, on-road, and off-road environments.

This article presents a comprehensive survey of LLMs applied to navigation, encompassing theoretical foundations and practical applications. Figure 1 [Figure 1: see original paper] illustrates the structure of this work.

Following the introduction, this paper is organized into six sections. Section 2 briefly introduces the development of LLMs and robot navigation. Section 3 discusses the application of LLMs in various navigation stages, examining different motivations and technologies. Section 4 provides a comparative analysis of relevant datasets. Section 5 introduces evaluation metrics and compares model performance. Section 6 discusses unresolved issues, potential research directions, and future challenges. Finally, Section 7 concludes the paper.

2 Background

This section briefly reviews progress in LLMs and robot navigation.

2.1 Large Language Models

LLMs represent a class of Transformer-based language models renowned for their expansive parameter counts, often reaching hundreds of billions. These models undergo training using vast quantities of internet data, endowing them with broad language capabilities, primarily manifested through text generation. Prominent examples include GPT-3 [**?**], PaLM [**?**], LLaMA [**?**], and GPT-4 [**?**]. A salient attribute exhibited by LLMs is their emergent proficiencies, such as

in-context learning (ICL) [?], instruction comprehension, and chain-of-thought reasoning (CoT) [?].

In contrast to traditional machine learning models, the prowess of large language models is evidenced by their deep bidirectional representations, potent context comprehension, and efficient handling of complex tasks. Conventional models such as Long Short-Term Memory Networks (LSTM) typically rely on specific data structures and algorithms. In contrast, substantial language models like GPT-4 and Sora [?], founded on the Transformer architecture, rely solely on attention mechanisms for information processing.

Leveraging language models to develop embodied models that aid intelligent agents in achieving human-like capabilities represents a burgeoning direction. The language competence and knowledge inherent in LLMs can assist agents in reasoning about semantic information and planning executable actions [?, ?, ?, ?, ?, ?]. Through embodied language models [?], sensor data can be directly converted into interactive code, facilitating a direct transformation from perception to words. The gap between perceptual information and text disappears as real-world sensor data seamlessly integrates with language models. Voyager [?] introduces lifelong learning through three primary components: an autonomous curriculum that encourages exploration, a skill repository to store and retrieve intricate behaviors, and an iterative prompting mechanism to generate code for low-level control. Voxposer leverages LLMs to generate robot trajectories for diverse manipulation tasks, guided by open-ended instructions and object cues [?].

2.2 Robot Navigation

Mobile robot navigation technology [?, ?] has garnered widespread attention due to its comprehensiveness and practicality [?]. Over the years, research has been fruitful, integrating algorithms ranging from classical control to machine learning [?]. It involves a multi-layered architecture encompassing perception, planning, and control. Resolving these three core questions involves key technologies including environmental perception, autonomous localization, and motion planning.

Environmental perception technology involves using onboard sensors to perceive the surrounding environment and processing acquired data to obtain specific information about surroundings (including feature and positional information) [?]. In scenarios where the environmental map is unknown and the initial position is uncertain, mobile robots must first rely on sensors to perceive external information before proceeding with localization, map construction, and path planning. Therefore, environmental perception forms the foundation of autonomous navigation [?].

The localization issue is fundamental to achieving autonomous mobility. Depending on task requirements, localization can be divided into pose tracking, global localization, and kidnapping scenarios. Autonomous localization entails

the robot utilizing prior environmental map information, current pose estimates, and sensor observations as input data to generate more accurate pose estimations [?].

Path planning involves the robot integrating prior map information and real-time sensory input of the surrounding dynamic environment to search for a trajectory connecting the starting point and goal point. This trajectory, under specific criteria, is optimal and ensures the robot can navigate past dynamic obstacles in real-time, ultimately reaching the target smoothly [?, ?, ?]. Achieving autonomous robot navigation necessitates addressing three core challenges, categorized into Map-Based Navigation [?] and Mapless Navigation [?].

Having an environmental map aids efficient path planning. However, in many scenarios, the environmental map is initially unknown, leading to the development of mapless navigation and Simultaneous Localization and Mapping (SLAM). Moreover, the quality of map-based navigation is directly influenced by the representation and accuracy of the environmental map. Maps can be metric or topological. In metric maps, detected obstacles, landmarks, and the robot' s position are represented relative to a specific reference frame.

In recent years, algorithms based on traditional search and sampling methods continue to emerge [?, ?, ?]. Furthermore, learning-based planning methods have attracted attention, including approaches that integrate both conventional and learning paradigms [?]. Traditional search methods encompass visual graph search algorithms and grid map-based search algorithms, while sampling-based algorithms include probabilistic roadmaps and methods based on random search trees. Learning-based methods consist of socially aware motion planning based on reinforcement learning. Hybrid methods that combine traditional and learning approaches involve learning methods for predicting self-vehicle posture in sampling-based incomplete motion planning tasks [?].

3 LLMs in Robot Navigation

The development of robot navigation has been swift, yet it remains riddled with challenges. The primary technical hurdle lies in path planning. Effective navigation demands consideration of dynamic environmental variations and uncertainties, alongside the art of reaching intended destinations while evading obstacles [?]. As robots navigate unfamiliar terrains, environmental uncertainties and intricacies pose additional challenges. Human activities, for instance, exert noteworthy influence on robot navigation. Robots must adeptly cohabit spaces with humans and avert conflicts [?]. Furthermore, the selection of sensors and the fusion of their data stand out as pivotal challenges, as each sensor boasts distinct levels of precision and reliability. Harmonizing data from diverse sensors to enhance positioning and navigation accuracy remains a critical research avenue [?, ?].

To confront these challenges, amalgamating LLMs with robot navigation emerges as a viable convergence. The fusion of robots with LLMs traces back

to 2017, which witnessed the inception of the Transformer model founded on attention mechanisms [**?**]. Subsequently, in 2019, the introduction of the BERT model propelled deep bidirectional learning representations [**?**]. These advancements laid a robust foundation for subsequent fusion. Moreover, commencing in 2020, the method of pre-training massive language models and fine-tuning them for specific tasks has led to significant performance enhancements across various NLP tasks [**?**]. The effective collaboration between robots and LLMs encapsulates a broad spectrum of outcomes, enhancing intelligence and human-machine interactions, fortifying autonomy, decision-making prowess, perceptual and control capabilities, and extending to fostering learning, adaptability, social interactions, and emotional exchanges [**?**, **?**, **?**, **?**]. These breakthroughs showcase the extensive application potential of LLMs within robotics and furnish novel directions for future research.

In the ensuing sections, we delve into the impact of LLMs across various stages of robot navigation and the associated models (Figure 2 [Figure 2: see original paper]).

3.1 Robot' s Environment Perception and Semantic Understanding

Humans primarily rely on their five senses for perception while moving, without prior prompts. To endow LLMs with similar capability, it is necessary to provide them with comprehensive multimodal descriptions of the environment. To achieve this, a Converter with multimodal perception capabilities is indispensable, as it conveys detailed descriptions of environmental features to LLMs. Currently, three main forms of Converters exist:

- **Capturing correlation between vision and language**, as exemplified in LM-Nav [**?**], involves utilizing Vision Language Models (VLM) [**?**] like CLIP as grounding modules for LLMs. LLMs leverage semantic understanding capabilities to parse free-form textual instructions and match landmark descriptions with images observed by the robot. VLMs such as CLIP may not be flawless; for instance, they might fail to detect specific landmarks like fire hydrants or cement mixers. If a landmark exists and can be recognized by the VLM [**?**, **?**, **?**], the robot can effectively localize and follow it. However, if the VLM fails to detect a landmark correctly, the robot may choose an incorrect path.
- **Describing multimodal information in natural language**, as exemplified by Matcha [**?**], involves utilizing Multimodal Perception Modules to convert results into natural language form for improved comprehension and unified processing by language models. For instance, the visual perception module uses a pre-trained ViLD [**?**] model to detect object categories and positions, subsequently generating scene descriptions. The impact sound perception module categorizes impact sounds and translates them into natural language descriptions. Similarly, weight measurement information is directly transformed into expressions such as "it is light"

or "it weighs 30 grams," facilitating seamless integration with language processing frameworks.

- **Encoding original images into Visual Tokens**, akin to the Image-oriented Tokenizer in [**?**], involves techniques like Vector Quantized Generative Adversarial Network (VQ-GAN). VQ-GAN decomposes images into discrete visual units mapped to a visual dictionary, transforming image data into a format processable by language models. Subsequently, a lightweight projection module utilizes a trainable projection matrix to map these visual tokens into the same vector space as text embeddings, enabling LLMs to integrate visual and language information for multimodal response generation.

Upon receiving multimodal information from the Converter, LLMs engage in semantic reasoning by understanding and parsing environmental information, language instructions, and historical records. For instance, NavGPT [**?**] utilizes Visual Foundation Models (VFM) to convert visual information into natural language descriptions, then combines navigation system principles and history to infer the robot' s current location. If LLMs receive visual information describing "refrigerator" and "sink," and previous history indicates passing through the "kitchen,"they may infer the current location is likely in the kitchen. LLMs make decisions on subsequent actions based on this information, such as selecting actions corresponding to identified viewpoint IDs, but do not directly "recognize" rooms; instead, they make judgments based on contextual information.

Beyond localization, LLMs can infer interactive information in the environment through perception. For instance, to facilitate effective object interaction, robots can utilize LLMs to understand object affordances and determine which objects require avoidance based on language instructions [**?**]. Within the VoxPoser [**?**] framework, affordances represent how objects can be used (e.g., a drawer being opened), while constraints limit action possibilities (e.g., avoiding contact with a vase). By mapping affordances and constraints onto a 3D value map, robots can synthesize motion trajectories to execute complex tasks, considering obstacles and objectives. This approach enables robots to comprehend language instructions, generalize to unseen instructions with zero-shot learning, and operate without extensive robot-specific data.

3.2 High-level Planning

Robot navigation planning faces challenges across multiple levels, from agile low-level control to high-level planning and reasoning, requiring extensive domain knowledge amidst vast search spaces [**?**, **?**]. The goal of navigation planning is to convert natural language instructions, including spatial and temporal constraints, into coordinates or encodings of time path points, such as (x_i, y_i, z_i) [**?**]. Through vast pre-training data and self-supervised learning, LLMs can provide navigation planning for executing complex long-horizon tasks [**?**]. A prompting approach introduced in [**?**] utilizes LLMs to directly generate action

sequences without additional domain knowledge. LLMs empower navigation planning through decomposing instructions into subtasks, tracking navigation progress, and adapting to exceptions in plan adjustments [**?**] (Figure 3 [Figure 3: see original paper]).

Sub-task decomposition involves generating a sequence of subtasks from language instructions, where each subtask is an independent unit contributing to the overall task. For instance, a "move from point A to point B and then turn back" task may include "navigate to point B" and "turn back" subtasks. Decomposed subtasks are handed to low-level controllers responsible for executing simple actions and basic environmental interactions [**?**], translating each subtask into executable control commands. For example, to move to a table, the controller takes the agent' s position and orientation as input and generates forward, left turn, and right turn commands.

Existing subtask decomposition methods fall into three categories: **Zero-shot Planning**, **Recursive Planning**, and **Task Planning + Feedback** [**?**]:

- **Zero-shot Planning**, as proposed in [**?**], involves LLMs generating the entire subtask sequence at once by integrating geospatial data and natural language instructions, without verifying executability.
- **Recursive Planning** involves iteratively prompting LLMs to generate each subsequent subtask based on preceding subtasks. In SayCan [**?**], the value function module generates a value function space based on the current scene and previous task results, representing the likelihood of different task executions, then selects the most probable next subtask from candidates.
- **Task Planning + Feedback** combines subtask sequence generation with navigation progress tracking and feasibility checking. This approach finds a subtask sequence that satisfies the entire task and verifies feasibility before execution, enabling LLMs to maintain comprehensive understanding of navigation history and track progress. For infeasible subtasks, LLMs can prompt feedback regarding the infeasible actions to generate a new subtask sequence, similar to the hierarchical approach in [**?**].

When navigation planning involves complex environments and temporal constraints, simultaneously performing task inference and motion planning can render subtask decomposition infeasible. To address this challenge, [**?**] proposed a method that does not directly plan subtasks using LLMs. Instead, it executes a few-shot transformation from natural language task descriptions to intermediate task representations, then utilizes the TAMP algorithm to jointly solve task and motion planning. Additionally, by introducing self-regressive prompting techniques, it detects and corrects potential synthetic or semantic errors, ensuring robots complete tasks as intended.

Furthermore, navigating in unknown environments makes it infeasible to generate multi-step long-range plans initially, potentially leaving agents unable to

modify plans during execution. During exploration, effective new location navigation search plans need incremental generation and updating. For this purpose, robots can use LLMs as memory to store previously visited areas [**?**]. As described in SayNav [**?**], SayNav utilizes a 3D scene graph to support memory for future planning, automatically annotating nodes of explored rooms so agents do not replan for revisited rooms. Additionally, LLMs can track their plans through the conversation chain module in the LangChain framework, which avoids reaching maximum prompt limits by not relying on entire conversation history.

Traditional robot memory implementations rely on advanced storage and processing technologies simulating human or animal brain mechanisms, ranging from sparsely distributed memory [**?**] to lifelong learning frameworks [**?**], experience-based predictive mapping [**?**], and biologically inspired cognitive models based on hippocampal mechanisms [**?**]. These methods enable robots to effectively learn and adapt in different environments. LLM-based tracking may generalize better to other tasks by eliminating the need for modules that track planning history, thus exhibiting strong generalization capabilities for diverse task implementations.

3.3 Low-level Control

Traditional robot control systems typically employ predefined functions, behaviors, and algorithms for navigation, using PID controllers to adjust speed and direction to ensure safe, effective movement to target locations. Hardware limitations and safety constraints are often hard-coded to prevent unexpected behaviors. To enable LLMs to control robots, one can teach them the mapping between natural language and program code, issuing commands in natural language that are subsequently converted into machine-executable program code. For instance, Code-as-Policies [**?**] utilizes programming-oriented LLMs to generate robot policy code based on natural language instructions.

Using LLMs as robot controllers offers several advantages. Firstly, LLMs can generate highly flexible robot strategies, aiding adaptation to various environments. Secondly, there is no need to collect training data or undergo training; existing visual perception models can be directly utilized, so improvements in underlying models directly translate to enhanced operational accuracy without additional costs. Lastly, LLM strategy codes exhibit high interpretability. Typical approaches [**?**, **?**] involve designing guiding prompts for LLM generation, including Application Programming Interfaces (APIs) and contextual examples, with natural language instructions provided to the LLMs. The final input consists of code APIs, usage examples, and task instructions.

To break the constraint of being limited to fixed skill sets, [**?**] introduces a framework utilizing an LLM-based planner to query for new skills, instantiating skills from an imitation learning agent and gathering human demonstrations when necessary to effectively learn new skills. Additionally, [**?**] mentions that com-

bining LLMs with lifelong robot learning allows LLMs to request new skills to accomplish given tasks. When a robot cannot accomplish a task using existing skills, LLMs request to learn a new skill by observing human demonstrations, acquiring and incorporating it into the skill library for future reuse, demonstrating the potential of open-world and lifelong learning [**?**].

3.4 Human-computer Interaction

Human-robot interaction is crucial in robot navigation, enabling seamless collaboration and communication between users and robots. The capability of linking vision with language in VLMs can facilitate this interaction. However, VLM' s fuzzy semantic understanding makes human-robot interaction challenging. For example, VLM features may identify a grass field as green and belonging to the grass category but may not infer that it could be the soccer field indicated in language instructions [**?**]. Therefore, applying LLMs in the human-robot interaction module for navigation is essential. LLMs can overcome VLM limitations by enhancing object understanding for more comprehensive, intelligent environmental processing. Additionally, converting sound into language through voice sensors enables conversational interactions, allowing robots to update semantic maps [**?**], eliminate semantic ambiguities, detect ground types [**?**], and perform functions unattainable through visual information alone.

Multimodal sensors further assist LLMs in effectively guiding robot behavior in real-world scenarios. Multimodal input enables robots to have more comprehensive understanding of human commands. For example, LIM2N [**?**] combines language and hand-drawn sketches for intuitive, user-centered interactions, allowing users to guide robots through language instructions or drawn paths. Multimodal inputs supplement information difficult to describe in words. Zhong et al. [**?**] proposed an intuitive remote operation system using VR devices, demonstrating VR-based control feasibility. HRC [**?**] integrates VR thoroughly into robot navigation, utilizing a VR-based remote operation system as a bridge for human-robot collaboration, enhancing transparency and controllability of remote operations.

Beyond conventional Visual-Linguistic Navigation (VLN) and Socially Aware Navigation (SAN) tasks, navigating in human-populated spaces is crucial for supporting advanced services such as collaborating with users or walking alongside them. SAN tasks face two main challenges: real-time highly volatile user requests or perceptions, and managing constraints on socially compatible navigation behaviors in dynamic environments. Current approaches have addressed challenges of realizing social robots in public environments [**?**], drawing inspiration from machine learning [**?**], sociology [**?**], analytical mechanics [**?**], algebra, geometry [**?**, **?**], and other fields.

3.5 Multi-robot Coordination

Studying multi-robot collaboration across domains can alleviate limitations of single-agent systems, including active mapping [**?**], exploration [**?**], and search [**?**]. In exploration tasks, Multi-Agent Reinforcement Learning (MARL) [**?**] and Graph-based Learning [**?**] facilitate transitioning planners from single-agent configurations [**?**], emphasizing Multi-Agent Visual Semantic Navigation that leverages scene prior knowledge to localize objects and formulate navigation strategies through reinforcement learning. While planning-based and learning-based techniques have been successful, they often require common-sense learning for robot missions [**?**]. In contrast, LLMs encompass broader, richer prior knowledge, enabling application in multi-robot navigation tasks. Recent works [**?**, **?**, **?**] have demonstrated feasibility of converting environmental observations into linguistic inputs for LLMs to facilitate communication and decision-making within Multi-Agent Systems (MAS). Most works employ hierarchical structures to ensure smooth MAS operation. Mainstream LLM-based multi-agent planning frameworks fall into two branches: centralized [**?**, **?**, **?**, **?**] and decentralized [**?**, **?**, **?**, **?**].

- **Centralized systems** employ LLMs to comprehend observations, history, and task progress of multiple agents, collaboratively allocating tasks to robot groups [**?**] or individuals [**?**]. For example, [**?**] implements a centralized multi-agent navigation framework extracting boundary and semantic information from maps, utilizing LLMs to allocate exploration areas to each robot. This demonstrates good coordination in small-scale teams, but as team size grows, increasing communication and information processing burdens pose challenges to rational, timely planning [**?**].
- **Decentralized systems** treat each robot as an autonomous entity that exchanges historical observation results through human-like language communication and makes adaptive decisions [**?**]. CoELA [**?**] provides a systematic template for decentralized communication and collaboration, dividing each agent' s execution into five modules: observation, belief, communication, reasoning, and planning, where LLMs facilitate communication and reasoning among agents.

3.6 Summary

Robot navigation technology has evolved from simple geometric feature tracking [**?**] to complex path planning in dynamic environments. These advancements enhance robot autonomy and efficiency in known or unknown environments while strengthening adaptability in complex, dynamic settings. Different environments present various challenges and requirements. LLM application in robot navigation leads to significant differences in strategies across three distinct environments: indoor, on-road, and off-road, manifested in environmental perception, path-planning algorithm selection, and localization technique application.

- **Indoor environments**: Target objects are more likely to appear near specific rooms and objects, aiding agents in searching. Upon detecting room and object information, agents can leverage pre-trained LLMs to perform commonsense reasoning on target objects and semantic scene information through text prompts [**?**].
- **On-road environments**: The focus lies in environmental complexity and navigation task scale. Dense city street networks require handling intricate intersections and directional changes (3-way, 4-way, 5-way), posing higher demands on landmark detection and directional understanding. Urban navigation paths typically average 40 steps, far exceeding the 6-step average for indoor tasks, necessitating continuous visual and language comprehension throughout the process [**?**]. LLMs' semantic understanding assists agents in better identifying landmarks, understanding instructions, and learning trajectory memories. Road navigation must also consider human behavior, especially in densely populated urban environments. LLMs' ability to track user language feedback and adjust robot behavior inference is crucial, capturing human intent and spatiotemporal dependencies to comprehend and adapt socially acceptable navigation behaviors [**?**].
- **Off-road environments**: Terrain variability significantly impacts robot movement, making it challenging to devise universally effective navigation policies across all scenarios [**?**, **?**]. LLM-based approaches combine human insights with technical solutions. Humans intuitively grasp environmental contexts, such as associating wet grass with high impedance—concepts current algorithms struggle to comprehend but can be elucidated using LLMs to interpret environmental backgrounds. When robots receive contextual information from human observers, such as "you are entering a grassy area after rain" or "you are walking on dry rocky paths under the sun," an LLM-based translator can extract embeddings representing contextual information from human explanations, enhancing navigation decision-making and sensory observations [**?**].

In robot navigation and autonomous driving, numerous models constantly emerge, each with unique advantages and applicable scenarios. However, achieving more efficient and accurate performance requires comparative analysis. By researching different models' capabilities in handling complex environments and completing specific tasks, we can better understand their characteristics, providing targeted choices and optimization solutions for practical applications. As shown in Table 1 , our comparison reveals strengths and weaknesses of typical models, providing important references for future research and development. We should leverage these advantages and carry out innovations combined with actual needs to promote greater breakthroughs. Through continuous exploration, these models will play more significant roles in their respective domains, bringing more convenience and progress to society.

4 Datasets

LLMs play a crucial role in assisting robot navigation by accurately understanding and parsing instruction information, providing accurate path planning and decision support that greatly enhances autonomous navigation capabilities in complex environments. Different datasets also have unique roles in visual navigation, covering rich information including semantics, visual features, multimodal fusion, and environmental interaction, providing valuable resources for robot learning and training.

Table 2 presents a multi-dimensional framework that systematically organizes key datasets in robot language visual navigation, encompassing scale size, applicable environment, types, and exploration methods, offering a clear, intuitive, and highly valuable reference for precise comparison and in-depth analysis.

Semantic information datasets are crucial for improving robot understanding and task execution ability, providing precise semantic knowledge and contextual information that enables robots to accurately understand human instructions and intentions for better planning and execution. By learning these datasets, robots can accurately identify objects, scenes, and concepts, associating them with corresponding actions and decisions to improve problem-solving efficiency and task execution. This also enhances adaptability to different tasks and environments, increasing flexibility. The R2R dataset [**?**] studies vision and language navigation, containing 21,567 natural language instructions for agent navigation in the Matterport3D simulator. The REVERIE dataset [**?**] contains 21,702 instructions involving navigation and referential expression information, helping agents navigate real indoor environments and identify remote target objects. Just Ask [**?**] introduced human-computer interaction in visual and language navigation based on R2R, including MC and ASA models and data enhancement strategies to improve navigation success rates and adapt to noise in human responses. The FAO dataset [**?**], based on Matterport3D, provides instructions containing object attributes and relationships so agents can find target objects from any position. The VNLA dataset [**?**] contains rich indoor environment information and target object instructions, where agents learn to find targets through advisor interaction and independent exploration. The HANNA dataset [**?**], also based on Matterport3D, simulates agents searching for objects in indoor environments, allowing agents to request help and training their ability to interpret language and visual instructions. The XL-R2R dataset [**?**] is the first cross-language VLN dataset, extending R2R with Chinese instructions. The ALFRED dataset [**?**] contains 25,743 natural language instructions for translating instructions and egocentric vision into action sequences for domestic chores. DialFRED [**?**] is a conversational embodied instruction-following benchmark extended from ALFRED, allowing robots to actively ask questions and use answers to complete tasks. The IQA dataset [**?**], based on AI2-THOR, contains over 75,000 multiple-choice questions requiring agents to navigate scenes, interact with objects, and plan actions to answer questions. The TEACh dataset [**?**] contains 3,047 game sessions for completing

household tasks in AI2-THOR. The RoomNav dataset [**?**], based on House3D, allows agents to navigate from random positions to target rooms according to high-level semantic concepts. The EQA dataset [**?**] represents questions through executable functional programs to test various agent abilities.

Visual feature datasets are critical for improving visual perception ability, providing rich feature information (shape, color, texture) that enables robots to accurately perceive and recognize surroundings. By learning these datasets, robots can optimize visual perception models and improve information processing and understanding, leading to better navigation, obstacle avoidance, and adaptation to different lighting and environmental changes. Matterport3D [**?**] is a large-scale RGB-D dataset containing 10,800 panoramic views of 90 building-scale scenes, composed of 194,400 RGB-D images with annotations for surface reconstruction, camera poses, and 2D/3D semantic segmentation. The TOUCHDOWN dataset [**?**] builds an outdoor visual street environment based on Google Street View, containing 29,641 panoramas and 61,319 edges covering New York City for studying natural language navigation and spatial reasoning. The RxR dataset [**?**] is a new visual and language navigation dataset with path designs meeting various expected characteristics including path length variance, indirect approaches, naturalness, and uniform environmental viewpoint coverage.

Multimodal fusion datasets enhance the ability to process complex information by integrating multiple modal data (text, image, audio), providing comprehensive information input. By fusing and analyzing these data, robots can better understand correlations and complementarity between modalities, more accurately process complex situational information, and flexibly respond to challenges. AI2-THOR [**?**] is an interactive 3D environment framework for visual AI research containing various scene datasets, agents, supported actions, image modalities, and environmental metadata, interacting with the Unity 3D game engine through Python API.

Datasets involving environmental interaction provide fundamental training for effective navigation and task execution, containing rich environmental information that simulates real-world scenarios. By learning these datasets, robots master environmental interaction skills, improve navigation accuracy and efficiency, and better understand task requirements to ensure smooth execution. The RobotSlang dataset [**?**] is collected through cooperative experiments where human drivers control physical robots and human commanders provide navigation guidance, studying language-guided simultaneous localization and mapping. The CVDN dataset [**?**] contains 2,050 human navigation dialogues in simulated home environments, studying robot navigation through dialogue. The Gibson dataset [**?**] is a virtual environment for training and testing real-world perception agents, built from real space scanning with neural network view synthesis and physical engine integration. The PROCTHOR dataset [**?**] contains 10,000 fully interactive houses with different room layouts, furniture placements, lighting, and materials for training agents in navigation and manipulation tasks. The HM3D dataset [**?**] is a large-scale indoor 3D environment dataset composed of

1,000 building-scale reconstructions from real-world locations. The HM3DSEM dataset [**?**] is the largest 3D real-world spatial dataset with dense semantic annotations, adding a dense semantic layer to HM3D. The Open X-Embodiment dataset [**?**] is a large-scale robot learning dataset collected by 21 institutions, containing over 1 million real robot trajectories from 22 robot types. The RoboVLN dataset [**?**] forms continuous control by mapping human-annotated R2R instructions to sparse waypoints in 3D reconstructed environments. The VLN-CE dataset [**?**] converts R2R' s panoramic image trajectories into fine paths in continuous Matterport3D environments. The AerialVLN dataset [**?**] designs tasks for drone navigation in urban outdoor environments, generating flight paths from experienced operators and annotations from AMT workers. The Talk2Nav dataset [**?**] realizes automatic pathfinding based on language in outdoor Google Street View environments. The StreetNav dataset [**?**] provides human-like tasks, following navigation instructions to reach destinations by randomly sampling start and target positions. The Retouchdown dataset [**?**] provides human-annotated instructions and spatial description parsing tasks for navigating New York City streets. StreetLearn [**?**] defines navigation tasks such as courier tasks where agents learn navigation strategies based on visual observation and target location encoding.

This research focuses on datasets in robot language visual navigation, clarifying the importance of LLMs and the unique roles of different datasets in semantic information, visual features, multimodal fusion, and environmental interaction. Through classification and analysis of multiple datasets, we provide a clear framework for related research, helping to improve robot navigation ability and expected to bring innovative breakthroughs across fields.

5 Evaluation Metrics and Analysis

LLM-based navigation techniques have been widely applied across society, making performance optimization and improvement vitally important. To abstract effectiveness into mathematical terms, this paper presents mainstream evaluation metrics providing insights into accuracy and adaptability, encompassing navigation accuracy, efficiency, and instruction adherence [**?**].

Key metrics include:

- **Success Rate (SR)** [**?**]: The frequency of task completion within defined proximity to the target.
- **Path Length (PL)** [**?**]: The overall navigation route length.
- **Shortest Path Distance (SPD)**: Evaluates the final position relative to the predetermined destination.
- **Trajectory Length (TL)** [**?**]: The average distance traveled. Longer paths reduce efficiency due to increased wear and resource utilization.
- **Success weighted by Path Length (SPL)** [**?**]: Balances SR and path efficiency.
- **Oracle Success Rate (OSR)**: Assesses whether any path portion is close

to the predefined target location.

- **Remote Grounding Success Rate weighted Path Length (RGSPL)**: Reconciles Remote Grounding Success (RGS) failure rate with navigation path efficiency.
- **Key Point Accuracy (KPA)**: Gauges correct decision rate at key points.
- **Distance to Goal (DTG)**: The minimum distance between robot and target when the episode concludes.

Success Rate (SR) delineates the proportion of accurately completed tasks, signifying goal achievement efficiency:

$$SR = \frac{\text{Number of tasks accurately completed}}{\text{Total tasks}}$$

Path Length (PL) mirrors the total distance covered throughout task completion, where shorter paths signify greater efficiency:

$$PL = \sum_{i=1}^{N} P_i$$

Success weighted by Path Length (SPL) combines SR and PL to evaluate task completion efficiency:

$$SPL = \sum_{i=1}^{N} \frac{S_i \cdot \min(P_i, L_i)}{N}$$

where S_i indicates task i success (1 if successful, 0 otherwise), P_i denotes navigated path length, L_i represents the shortest path, and N is the task count.

Oracle Success Rate (OSR) indicates whether the distance from any path point to the target is within a pre-defined threshold:

$$OSR = \sum_{i=1}^{N} I\left(\min_{n \in P_i} \text{dist}(n, g_i) \leq \text{Threshold}\right)$$

The indicator function $I(\min_{n \in P_i} \text{dist}(n, g_i) \leq \text{Threshold})$ returns 1 if the minimum distance from any node within the path to the target is less than or equal to the threshold, otherwise 0.

Under the **Remote Grounding Success Rate (RGS)** standard, success is only attained when the agent locates an instance corresponding to the target semantic label. This metric is widely used in goal-oriented tasks. The **Remote Grounding Success Rate Weighted by Navigation Path Length**

(RGSPL) correlates robot efficiency in achieving RGS with traversed path length:

$$RGSPL = \sum_{i=1}^{N} \frac{SRGS_i \cdot \max(P_i, L_i)}{N}$$

Similar to OSR, $SRGS_i$ is an indicator where $SRGS = 1$ when task i is successful, otherwise 0.

Subsequent analysis attempts to evaluate LLM-based navigation methods using these metrics. Unfortunately, evaluation criteria remain non-standardized, with each method having unique characteristics. Therefore, Table 3 integrates performance data of various LLM-related navigation models across different environments using SR and SPL. By comparing data from models in environments such as HM3D, R2R, and MP3D, we can clearly understand each model' s strengths and weaknesses.

In summary, in-depth analysis of evaluation metrics and navigation model performance enables us to distinctly perceive current research achievements and deficiencies. These findings provide a solid foundation for optimizing existing models and indicate clear directions for subsequent research.

6 Challenges and Limitations

Applying LLMs to robot navigation shows significant potential. However, to ensure comprehensive and rigorous research, we must consider the following obstacles:

- **Limitations in spatial reasoning abilities**: Despite excelling in sequence modeling and pattern recognition, LLMs exhibit shortcomings in complex spatial reasoning tasks. For instance, ChatGPT-3.5 struggles with processing 3D robot trajectory data and may require specific prompting mechanisms to enhance performance [**?**].
- **Adaptation issues in the physical world**: A key challenge lies in effectively integrating LLMs with perception and action control. While some research shows LLMs can generate low-level control commands with minimal physical environment cues, this remains complex, especially in scenarios requiring dynamic behavior adjustments [**?**].
- **Compatibility issues in visual perception**: For tasks reliant on intricate environmental interactions, such as robotic visual navigation, textual instructions alone may not suffice. While multimodal LLMs (e.g., GPT-4V) have started enhancing robot motion planning [**?**], this still constrains LLM application in tasks requiring high-level visual perception.
- **Safety considerations as embodied platforms**: Robots' actions can have lasting environmental impacts, necessitating safe learning and inter-

action to prevent catastrophic events. This requires physical navigation systems to possess adequate self-monitoring and risk assessment capabilities.

- **Demands for real-time and reliability**: Practical applications require robot navigation systems to respond swiftly and make accurate decisions in uncertain environments. This necessitates LLMs to possess efficient processing capabilities while reducing time delays and ensuring correctness and reliability of plan execution [**?**].
- **Diversity in visual navigation**: Visual navigation encompasses various tasks such as visual-language navigation, grounded question answering, and scene navigation [**?**], requiring embodied navigation systems to exhibit sufficient flexibility and adaptability.

These limitations present an alternative perspective on current research progress. The primary challenge lies in LLM grounding mechanisms. Given sufficient training data, LLMs can effectively address navigation tasks in large-scale environments, contingent upon perfect visual perception and excellent robotic control strategies. In current research, LLMs often exist as processors or components within navigation models, particularly in outdoor navigation tasks, limiting their efficiency and success rates.

To fully leverage LLM capabilities, vast amounts of data are required. Current datasets typically focus on indoor environments or specific settings, making it challenging to acquire extensive datasets encompassing complex environments. Accelerating progress necessitates researchers to deeply address these issues.

7 Conclusions

This paper thoroughly examined various methods within LLM-based Navigation, describing their differences and commonalities, and provided an in-depth analysis of performance across tasks and environments, revealing fundamental design principles. Experimental results underscore the crucial role of LLMs in zero-shot navigation tasks. The paper also outlined inherent challenges and limitations, including the lack of direct connection between LLMs and the physical world, the need for large training datasets, and the complexity of understanding and generating natural language across environments.

Despite these obstacles, promising research trajectories can drive progress, including developing more robust and adaptive language models, investigating innovative training methods and architectures, establishing standardized benchmarks and evaluation metrics, and fostering urgent interdisciplinary collaboration among artificial intelligence, robotics, and social science researchers.

References

[1] S Haider Abdulredah and D Jasim Kadhim. Developing a real time naviga-

tion for the mobile robots at unknown environments. *Indones. J. Electr. Eng. Comput. Sci.(IJEECS)*, 20:500–509, 2020.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Al tenschmidt, Sam Altman, Shyamal Anadkat, et al. gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Saaket Agashe, Yue Fan, and Xin Eric Wang. Evaluating multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*, 2023.

[4] Swati Aggarwal, Kushagra Sharma, and Manisha Priyadarshini. Robot navigation: Review of techniques and research challenges. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 3660–3665. IEEE, 2016.

[5] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

[8] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton 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*, pages 3674–3683, 2018.

[9] Paola Ardón, Éric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020.

[10] Ioannis Arvanitakis, Anthony Tzes, and Konstantinos Giannousakis. Synergistic exploration and navigation of mobile robots under pose uncertainty in unknown environments. *International Journal of Advanced Robotic Systems*, 15(1):1729881417750785, 2018.

[11] Ashish Vaswani et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] Ben Beklisi Kwame Ayawli, Ryad Chellali, Albert Yaw Appiah, and Frimpong Kyeremeh. An overview of nature-inspired, conventional, and hybrid methods of autonomous vehicle path planning. *Journal of Advanced Transportation*, 2018(1):8269698, 2018.

[13] Shurjo Banerjee, Jesse Thomason, and Jason Corso. The robotslang benchmark: Dialog-guided robot localization and navigation. *Conference on Robot Learning*, pages 1384–1393. PMLR, 2021.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[15] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[16] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

[17] Yongchao Chen, Jacob Arkin, Charles Dawson, Yang Zhang, Nicholas Roy, and Chuchu Fan. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6695–6702. IEEE, 2024.

[18] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4311–4317, 2024.

[19] Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-Tur. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2459–2466, 2020.

[20] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[21] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh,

and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.

[22] Thomas Dean. High-level planning and low-level control. In *Intelligent Robots and Computer Vision VI*, volume 848, pages 496–501. SPIE, 1988.

[23] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Proc-thor: Large-scale embodied ai using procedural generation. *ArXiv*, abs/2206.06994, 2022.

[24] Yinan Deng, Jiahui Wang, Jingyu Zhao, Xinyu Tian, Guangyan Chen, Yi Yang, and Yufeng Yue. Opengraph: Open-vocabulary hierarchical 3d graph representation in large-scale outdoor environments. *arXiv preprint arXiv:2403.09412*, 2024.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[28] Stefan Edelkamp and Stefan Schrödl. *Heuristic search: theory and applications*. Elsevier, 2011.

[29] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.

[30] Felix Endres, Jürgen Hess, Nikolas Engelhard, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. An evaluation of the rgb-d slam system. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1691–1696. IEEE, 2012.

[31] Torvald Ersson and Xiaoming Hu. Path planning and navigation of mobile robots in unknown environments. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 858–864. IEEE, 2001.

[32] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *arXiv preprint arXiv:2312.07843*, 2023.

[33] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *Robotics and Automation Letters*, 7(4):10049–10056, 2022.

[34] Christophe Giovannangeli, Philippe Gaussier, and Gaël Désilles. Robust

mapless outdoor vision-based navigation. In *2006 IEEE/RSJ international conference on intelligent robots and systems*, pages 3293–3300. IEEE, 2006.

[35] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018.

[36] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[37] Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. Learning to follow directions in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11773–11781, 2020.

[38] HS Hewawasam, M Yousef Ibrahim, and Gayan Kahandawa Appuhamillage. Past, present and future of path-planning algorithms for mobile robot navigation in dynamic environments. *IEEE Open Journal of the Industrial Electronics Society*, 3:353–365, 2022.

[39] Huosheng Hu and Dongbing Gu. Landmark-based navigation of industrial mobile robots. *Industrial Robot: An International Journal*, 27(6):458–467, 2000.

[40] Junning Huang, Sirui Xie, Jiankai Sun, Qiurui Ma, Chunxiao Liu, Dahua Lin, and Bolei Zhou. Learning a decision module by imitating driver' s control behaviors. In *Conference on Robot Learning*, pages 1–10. PMLR, 2021.

[41] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. struct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023.

[42] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

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

[44] Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. Hierarchical cross-modal agent for robotics vision-and-language navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13238–13246. IEEE, 2021.

[45] Hadi Jahanshahi and Naeimeh Najafizadeh Sari. Robot path planning algorithms: a review of theory and experiment. *arXiv preprint arXiv:1805.08137*, 2018.

[46] Sascha Jockel, Mateus Mendes, Jianwei Zhang, A Paulo Coimbra, and Manuel Crisóstomo. Robot navigation and manipulation based on a predictive associative memory. In *2009 IEEE 8th International Conference on Development and Learning*, pages 1–7. IEEE, 2009.

[47] Shirel Josef and Amir Degani. Deep reinforcement learning for safe local planning of a ground vehicle in unknown rough terrain. *IEEE Robotics and Automation Letters*, 5(4):6748–6755, 2020.

[48] Frank Joublin, Antonello Ceravola, Pavel Smirnov, Felix Ocker, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Stephan Hasler, Daniel Tanneberg, and Michael Gienger. Copal: Corrective planning of robot actions with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8664–8670, 2024.

[49] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[50] Jon M Kleinberg. The localization problem for mobile robots. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 521–531. IEEE, 1994.

[51] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[52] Yoram Koren, Johann Borenstein, et al. Potential field methods and their inherent limitations for mobile robot navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 2, pages 1398–1404, 1991.

[53] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020*, pages 104–120. Springer, 2020.

[54] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020.

[55] John J Leonard and Hugh F Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on robotics and Automation*, 7(3):376–382, 1991.

[56] Sergey Levine and Dhruv Shah. Learning robotic navigation from experience: principles, methods and recent results. *Philosophical Transactions of the Royal Society B*, 378(1869):20210447, 2023.

[57] Frank L Lewis and Shuzhi Sam Ge. *Autonomous mobile robots: sensing, control, decision making and applications*. CRC Press, 2006.

[58] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

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

[60] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8):1345–1365, 2023.

[61] Li Linjun. Research and implementation of key technologies for mobile robot navigation. *Master' s thesis*, University of Electronic Science and Technology of China, 2017.

[62] Bo Liu, Xuesu Xiao, and Peter Stone. A lifelong learning approach to mobile robot navigation. *IEEE Robotics and Automation Letters*, 6(2):1090–1096, 2021.

[63] Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Enhancing the llm-based robot manipulation through human-robot collaboration. *arXiv preprint arXiv:2406.14097*, 2024.

[64] Jijia Liu, Chao Yu, Jiaxuan Gao, Yuqing Xie, Qingmin Liao, Yi Wu, and Yu Wang. Llm-powered hierarchical language agent for real-time human-ai coordination. *arXiv preprint arXiv:2312.15224*, 2023.

[65] 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.

[66] Xinzhu Liu, Di Guo, Huaping Liu, and Fuchun Sun. Multi-agent embodied visual semantic navigation with scene prior knowledge. *IEEE Robotics and Automation Letters*, 7(2):3154–3161, 2022.

[67] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[68] David V Lu and William D Smart. Towards more efficient navigation for robots and humans. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1707–1713. IEEE, 2013.

[69] Jinjie Mai, Jun Chen, Bing chuan Li, Guocheng Qian, Mohamed Elhoseiny, and Bernard Ghanem. Llm as a robotic brain: Unifying egocentric memory and control. *ArXiv*, abs/2304.09349, 2023.

[70] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 286–299, 2024.

[71] Gil Manor, Joseph Z Ben-Asher, and Elon Rimon. Time optimal trajectories for a mobile robot under explicit acceleration constraints. *IEEE Transactions on Aerospace and Electronic Systems*, 54(5):2220–2232, 2018.

[72] Jerry W Mao. *A Framework for LLM-based Lifelong Learning in Robot Manipulation*. PhD thesis, Massachusetts Institute of Technology, 2023.

[73] Christoforos Mavrogiannis and Ross A Knepper. Hamiltonian coordination primitives for decentralized multiagent navigation. *The International Journal of Robotics Research*, 40(10-11):1234–1254, 2021.

[74] Christoforos I Mavrogiannis and Ross A Knepper. Multi-agent path topology in support of socially competent navigation planning. *The International Journal of Robotics Research*, 38(2-3):338–356, 2019.

[75] Harsh Mehta, Yoav Artzi, Jason Baldridge, Eugene Ie, and Piotr Mirowski. Retouchdown: Adding touchdown to streetlearn as a shareable resource for language grounding tasks in street view. *arXiv preprint arXiv:2001.03671*, 2020.

[76] Mateus Mendes, A Paulo Coimbra, Manuel M Crisostomo. Robot navigation based on view sequences stored in a sparse distributed memory. *Robotica*, 30(4):571–581, 2012.

[77] Jean-Arcady Meyer and David Filliat. Map-based navigation in mobile robots: Ii. a review of map-learning and path-planning strategies. *Cognitive Systems Research*, 4(4):283–317, 2003.

[78] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019.

[79] Jun Nakanishi, Shunki Itadera, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Towards the development of an intuitive teleoperation system for human support robot using a vr device. *Advanced Robotics*, 34(19):1239–1253, 2020.

[80] Khanh Nguyen and Hal Daumé. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. *ArXiv*, abs/1909.01871, 2019.

[81] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12537, 2019.

[82] Duy Nguyen-Tuong and Jan Peters. Model learning for robot control: a survey. *Cognitive processing*, 12:319–340, 2011.

[83] OpenAI, Matthias Plappert, Raul Sampedro, Tao Xu, Ilge Akkaya, Vineet Kosaraju, Peter Welinder, Ruben D' Sa, Arthur Petron, Henrique P d O Pinto, et al. Asymmetric self-play for automatic goal discovery in robotic manipulation. *arXiv preprint arXiv:2101.04882*, 2021.

[84] Kevin Osanlou, Christophe Guettier, Tristan Cazenave, and Eric Jacopin. Planning and learning: A review of methods involving path-planning for autonomous vehicles. *arXiv preprint arXiv:2207.13181*, 2022.

[85] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[86] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[87] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022.

[88] Meenal Parakh, Alisha Fong, Anthony Simeonov, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Lifelong robot learning with human assisted language planners. In *CoRL 2023 Workshop on Learning Effective Abstractions for Planning (LEAP)*, 2023.

[89] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.

[90] Anis Naema Atiyah Rafai, Noraziah Adzhar, and Nor Izzati Jaini. A review on path planning and obstacle avoidance algorithms for autonomous mobile robots. *Journal of Robotics*, 2022(1):2538220, 2022.

[91] Abhinav Rajvanshi, Karan Sikka, Xiao Lin, Bhoram Lee, Han-Pang Chiu, and Alvaro Velasquez. Saynav: Grounding large language models for dynamic planning to navigation in new environments. *Proceedings of the International Conference on Automated Planning and Scheduling*, 34(1):464–474, 2024.

[92] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[93] Francisco Rubio, Francisco Valero, and Carlos Llopis-Albert. A review of mobile robots: Concepts, methods, theoretical framework, and applications. *International Journal of Advanced Robotic Systems*, 16(2):1729881419839596, 2019.

[94] Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. 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, 2024.

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

[96] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.

[97] Manasi Sharma. Exploring and improving the large language spatial reasoning abilities of models. In *I Can' t Believe It' s Not Better Workshop: Failure Modes in the Age of Foundation Models*, 2023.

[98] Chak Lam Shek, Xiyang Wu, Dinesh Manocha, Pratap Tokekar, and Amrit Singh Bedi. Lancar: Leveraging language for context-aware robot locomotion in unstructured environments. *arXiv preprint arXiv:2310.00481*, 2023.

[99] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.

[100] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023.

[101] Jiankai Sun, Hao Sun, Tian Han, and Bolei Zhou. Neuro-symbolic program search for autonomous driving decision module design. *Conference on Robot Learning*, pages 21–30. PMLR, 2021.

[102] PE Teleweck and B Chandrasekaran. Path planning algorithms and their use in robotic navigation systems. In *Journal of Physics: Conference Series*, volume 1207, page 012018. IOP Publishing, 2019.

[103] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020.

[104] Sebastian Thrun and Arno Bücken. Integrating grid-based and topological maps for mobile robot navigation. In *Proceedings of the national conference on artificial intelligence*, pages 944–951, 1996.

[105] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. A probabilistic approach to concurrent mapping and localization for mobile robots. *Autonomous Robots*, 5:253–271, 1998.

[106] Edmond Tong, Anthony Opipari, Stanley Lewis, Zhen Zeng, and Odest Chadwicke Jenkins. Oval-prompt: Open-vocabulary affordance localization for robot manipulation through llm affordance-grounding. *arXiv preprint arXiv:2404.11000*, 2024.

[107] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[108] Pete Trautman, Jeremy Ma, Richard M Murray, and Andreas Krause. Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation. *The International Journal of Robotics Research*, 34(3):335–356, 2015.

[109] Spyros G Tzafestas. Mobile robot control and navigation: A global overview. *Journal of Intelligent & Robotic Systems*, 91:35–58, 2018.

[110] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129:246–266, 2021.

[111] Sai H Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. *IEEE Access*, 2024.

[112] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[113] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, et al. Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv preprint arXiv:2401.04334*, 2024.

[114] Jun Wang, Guocheng He, and Yiannis Kanaros. Safe task planning for language-instructed multi-robot systems using conformal prediction. *arXiv preprint arXiv:2402.15368*, 2024.

[115] Weizheng Wang, Le Mao, Ruiqi Wang, and Byung-Cheol Min. Srlm: Human-in-loop interactive social robot navigation with large language model and deep reinforcement learning. *arXiv preprint arXiv:2403.15648*, 2024.

[116] Weizheng Wang, Ruiqi Wang, Le Mao, and Byung-Cheol Min. Navistar: Socially aware robot navigation with hybrid spatio-temporal graph transformer and preference learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11348–11355. IEEE, 2023.

[117] Yen-Jen Wang, Bike Zhang, Jianyu Chen, and Koushil Sreenath. Prompt a robot to walk with large language models. *arXiv preprint arXiv:2309.09969*, 2023.

[118] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[119] Jiajing Wu, Jieli Liu, Yijing Zhao, and Zibin Zheng. Analysis of cryptocurrency transactions from a network perspective: An overview. *Journal of Network and Computer Applications*, 190:103139, 2021.

[120] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018.

[121] Yuchen Wu, Pengcheng Zhang, Meiying Gu, Jin Zheng, and Xiao Bai. Embodied navigation with multi-modal information: A survey from tasks to methodology. *Information Fusion*, page 102532, 2024.

[122] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.

[123] Xuesu Xiao, Bo Liu, Garrett Warnell, and Peter Stone. Motion planning and control for mobile robot navigation using machine learning: a survey. *Autonomous Robots*, 46(5):569–597, 2022.

[124] Xuesu Xiao, Zizhao Wang, Zifan Xu, Bo Liu, Garrett Warnell, Gauraang Dhamankar, Anirudh Nair, and Peter Stone. Appl: Adaptive planner parameter learning. *Robotics and Autonomous Systems*, 154:104132, 2022.

[125] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2023.

[126] An Yan, Xin Eric Wang, Jiangtao Feng, Lei Li, and William Yang Wang. Cross-lingual vision-language navigation. *arXiv preprint arXiv:1910.11301*, 2019.

[127] Shu Yang, Muhammad Asif Ali, Cheng-Long Wang, Lijie Hu, and Di Wang. Moral: Moe augmented lora for llms' lifelong learning. *ArXiv*, abs/2402.11260,

2024.

[128] Kai Ye, Siyan Dong, Qingnan Fan, He Wang, Li Yi, Fei Xia, Jue Wang, and Baoquan Chen. Multi-robot active mapping via neural bipartite graph matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14839–14848, 2022.

[129] Lance Ying, Kunal Jha, Shivam Aarya, Joshua B Tenenbaum, Antonio Torralba, and Tianmin Shu. Goma: Proactive embodied cooperative communication via goal-oriented mental alignment. *arXiv preprint arXiv:2403.11075*, 2024.

[130] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. Co-navgpt: Multi-robot cooperative visual semantic navigation using large language models. *arXiv preprint arXiv:2310.07937*, 2023.

[131] Chao Yu, Xinyi Yang, Jiaxuan Gao, Jiayu Chen, Yunfei Li, Jijia Liu, Yunfei Xiang, Ruixin Huang, Huazhong Yang, Yi Wu, et al. Asynchronous multi-agent reinforcement learning for efficient real-time multi-robot cooperative exploration. *arXiv preprint arXiv:2301.03398*, 2023.

[132] Chao Yu, Xinyi Yang, Jiaxuan Gao, Huazhong Yang, Yu Wang, and Yi Wu. Learning efficient multi-agent cooperative visual exploration. In *European Conference on Computer Vision*, pages 497–515. Springer, 2022.

[133] Fengpei Yuan, Marie Boltz, Dania Bilal, Ying-Ling Jao, Monica Crane, Joshua Duzan, Abdu rhman Bahour, and Xiaopeng Zhao. Cognitive exercise for persons with alzheimer' s disease and related dementia using a social robot. *IEEE Transactions on Robotics*, 39(4):3332–3346, 2023.

[134] Jinsheng Yuan, Wei Guo, Fusheng Zha, Pengfei Wang, Mantian Li, and Lining Sun. A bionic spatial cognition model and method for robots based on the hippocampus mechanism. *Frontiers in Neurorobotics*, 15:769829, 2022.

[135] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023.

[136] K.R. Zentner, Ryan Julian, Brian Ichter, and Gaurav S. Sukhatme. Conditionally combining robot skills using large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14046–14053, 2024.

[137] Bin Zhang, Hangyu Mao, Jingqing Ruan, Ying Wen, Yang Li, Shao Zhang, Zhiwei Xu, Dapeng Li, Ziyue Li, Rui Zhao, et al. Controlling large language model-based agents for large-scale decision-making: An actor-critic approach. *arXiv preprint arXiv:2311.13884*, 2023.

[138] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative

embodied agents modularly with large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12689–12699, 2023.

[139] Tianyao Zhang, Xiaoguang Hu, Jin Xiao, and Guofeng Zhang. A survey of visual navigation: From geometry to embodied ai. *Engineering Applications of Artificial Intelligence*, 114:105036, 2022.

[140] Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Chat with the environment: Interactive multimodal perception using large language models. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3590–3596. IEEE, 2023.

[141] Zhonghan Zhao, Kewei Chen, Dongxu Guo, Wenhao Chai, Tian Ye, Yanting Zhang, and Gaoang Wang. Hierarchical auto-organizing system for open-ended multi-agent navigation. *arXiv preprint arXiv:2403.08282*, 2024.

[142] Sipeng Zheng, Yicheng Feng, Zongqing Lu, et al. Steve-eye: Equipping llm-based embodied agents with visual perception in open worlds. In *The Twelfth International Conference on Learning Representations*, 2023.

[143] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649, 2024.

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

[145] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15580–15590, 2021.

[146] Weiqin Zu, Wenbin Song, Ruiqing Chen, Ze Guo, Fanglei Sun, Zheng Tian, Wei Pan, and Jun Wang. Language and sketching: An llm-driven interactive multimodal multitask robot navigation framework. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1019–1025, 2024.

Note: Figure translations are in progress. See original paper for figures.