A Glimpse into LLMs & MLLMs Based Navigation: A Comprehensive Survey

Tianqi Shao 2200013082

Abstract

With the rapid advancement of large language models (LLMs) and multimodal large language models (MLLMs), especially the release of Generative Pre-trained Transformer(ChatGPT) and its subsequent versions like GPT-4o, o1 and so on, these novel models have made groundbreaking advances in numerous fields and tasks such as navigation, locomotion and some other essential applications on robots. Although there exist abundant amazing attempts and papers on tackling navigation tasks with complex and diverse environments and settings, a systematic overview focusing on this typical field is still relatively rare. To help researchers gain comprehensive understandings of relative tasks in this field and actionable insights to integrate LLMs & MLLMs into their robotic systems, this survey paper collects and categorizes the most representative state of the art works in this field, and provides a deep analysis of their advantages and limitations. Finally, the article discusses the open challenging problems and forecasts the future trends of LLMs & MLLMs based navigation.

1. Introduction

In recent years, the continuous development of large language models (LLMs) and multimodal large language models (MLLMs) has attracted increasing attention due to their potential in a variety of practical applications such as natural language processing or combining with other fields like robotics, enabling machines to understand, generate and even interact with human language in previously unimaginable ways. Nowadays, these LLMs, characterized by their vast parameter sizes and training on internet-scale datasets, have already achieved remarkable success in showing few-shot [39] or zero-shot [6, 20] learning capabilities, enabling time-effective planning and decision-making for new tasks with only minimal or even no sample data. As LLMs & MLLMs emerge one after another and they become more and more powerful, people have gradually realized that many difficult problems and tasks can be better solved by utilizing these expressive models.

It is well known that the physical world we live in is inherently three-dimensional, thus understanding spatial 3D environments is crucial for embodied agents to make many real-world applications to fulfill tasks involving perception, navigation, and interaction within these 3D spaces. With the integration of large language models, embodied intelligence has been undergoing a rapid development and poses as a significant area of focus. Navigation, manipulation, locomotion and lots of other representative tasks of robotics have made tremendous advancements with the help of them. Among the myriad applications of LLMs, navigation tasks are particularly noteworthy, which demand deep understanding of the 3D environment and quick, accurate decision-making. LLMs can augment embodied intelligence systems with sophisticated environmental perception and decision-making support, leveraging their robust language and image-processing abilities. Many excellent works will be shown in the following sections, demonstrating that these LLM-based agents are valuable tools for navigation.

However, alongside these achievements, numerous technical and theoretical challenges persist. How to reduce the latency for real-time applications with the integration of text, images, and other sensor data simultaneously and how to enhance the training efficiency without sacrificing performance are still unresolved issues.

This survey will be constructed as follows:

1. In Section 2, I will provide a brief overview of the background and related works in the field of Embodied Intelligence and LLMs & MLLMs.
2. In Section 3, I will define the navigation tasks and introduce the basic problems to be resolved in this field.
3. In Section 4, I will present the state-of-the-art methods and models for LLMs & MLLMs based navigation.
4. In Section 5, I will discuss the challenges and open issues existing in this field.
5. In Section 6, I will conclude the survey and provide some insights into the future trends of LLMs & MLLMs based navigation.

2. Background

2.1. Embodied Intelligence

Robots have played an important role in various scenarios and industries such as manufacturing, healthcare, and entertainment. With the increasing task complexity and working environment variability, the demands for robot tasks have shifted from fixed automation (e.g., rule-based robots) to general artificial intelligence, where robot learning will be the core enabling techniques of the autonomous systems. Embodied Intelligence, an emerging research field, mainly focuses on understanding and developing intelligent agents that closely interact with their environment [21], with various downstream tasks listed below:

2.1.1. Manipulation

Tasks requiring extensive physical interaction with objects in 3D environment, such as cutting vegetables, washing dishes, and packing clothes and some other routines in daily life, are still not-fully-addressed challenges in dealing with robot manipulation. It can be formulated as a policy to connect a known start state to a desired goal state with a sequence of actions, during which the manipulation task is represented by a set of points denoting the starting and goal states, along with the constraints, such as the physical laws, imposed on the middle states. Some widely-used methods for tackling manipulation tasks include:

1. **Learning of Robotic Skill Primitives**: This method breaks down the whole complex task into a set of more fine-grained and manageable skill or movement primitives, which are learned from demonstrations or reinforcement learning. The skill primitives learnt can generalize to novel tasks and unseen environments.
2. **Learning of Complex Manipulation Tasks for Robots**: This method consists of acquiring knowledge, accumulating experience, and continuously updating and expanding manipulation skills, which empower robots with the capability to autonomously learn from their environmnet and independently accomplish complex manipulation tasks. Notable methods include RoboGen [44] for generating simulated data, Eureka [26] for human-like reward design, and CLIPORT [38] for language-conditioned imitation learning, among others.
3. **Robot Manipulation Learning with Multimodal Fusion**: This approach handles the challenges by introducing multiple modalities such as vision, touch, and audio. With this integration, robots are able to understand and process information in various forms, greatly improving their robustness and adaptability in complex tasks. Some state-of-the-art methods include MOMA-Force [47] for visual binding force, VIMA [13] for processing multimodal prompts, and RT-2 [3] for transferring network knowledge to robot control.

2.1.2. Task Planning

In this context, "planning" is a broad concept encompassing not only for single-robot, decomposing a complex task down into a sequence of sub-tasks, but also for multi-agent systems (MAS), considering the relationship between tasks, robot capabilities, cooperation and other challenges.

1. **Single-Robot Task Decomposition**: This method focuses on single robot task planning, by breaking down the given target into simpler sub-tasks, to enhance the effectiveness of the whole task. With the help of LLMs, robots deal with the task sequence more stable and efficient and gain a powerful capability that they can convert natural language instructions into practical actions. LLM+P [22] incorporate the strengths of the classical planner into LLM firstly. Wake et al. use ChatGPT to convert natural language instructions into executable robotic actions in long-step scenarios [41].
2. **Multi-Robot Task Planning**: The main challenge under this scenario is, when facing a multi-agent system, the manager needs to allocate the fine-grained tasks to each agent and achieve optimal results in general. A novel framework, named as RoCo [27], is proposed to solve this difficulty, leveraging pre-trained LLMs for high-level communication with other agents or humans and low-level path planning.

2.1.3. Reasoning

In this field, robots are demanded to possess a detailed understanding and an explicit representation of their environment to reason based on logic, common sense, affordance, personalized customization and so on. Below are some workable methods:

1. **Logical Reasoning**: PaLM-E plugs real-world continuous sensor modalities into a LLM to establish a link between words and perception. [8]
2. **Common Sense Reasoning**: Kim et al. proposed KG-GPT, a general framework that utilizes LLM for knowledge graph reasoning. [17]
3. **Affordance Reasoning**: Yoneda et al. published the Statler framework, which enables LLMs to have representations of world states that change over time while maintaining memory. [48]
4. **Personalized Reasoning**: Wu et al. used TidyBot, a robot to learn personal preferences to personalize the cleaning of a room. [45]

2.1.4. Navigation

To be brief, navigation is the task of guiding a robot or agent through a physical environment to reach a specified goal while avoiding obstacles and adapting to dynamic conditions. Since it is the main topic of this survey, I will discuss it more detailed than other fields of embodied intelligence in Section 3

2.2. Language Models

Prior to the emergence of language models, natural language processing (NLP) relies on a variety of traditional techniques that are more rule-based and structured. They are listed below:

1. **Rule-Based Systems**: These systems use handcrafted rules to parse and understand text. They often involve extensive sets of linguistic rules and patterns to identify parts of syntactic structures and semantic relationships.
2. **Statistical Methods**: They play a significant role before the deep learning revolution. Techniques such as n-gram models are used for language modeling, while Hidden Markov Models (HMMs) are designed for sequence tagging tasks.
3. **Machine Learning Algorithms**: Supervised learning algorithms like Support Vector Machines (SVMs) and Naive Bayes classifiers are commonly used for text classification tasks. These algorithms require labeled training data to learn the patterns and make predictions on new, unseen data. Feature engineering was a critical step in this process, where relevant features such as word frequencies, n-grams, and syntactic features were extracted from the text to improve model performance.

However, these traditional methods have several limitations. They are often labor-intensive, requiring domain expertise to design rules and features. They also struggled to capture the complexity and nuances of natural language, leading to suboptimal performance on tasks like language understanding, generation, and translation. The rise of deep learning and neural networks revolutionized the field of NLP by enabling the development of large-scale language models that could learn from vast amounts of text data and generate human-like text. Here are some influential types of language models that have been developed over the years:

2.2.1. Large Language Models (LLMs)

LLMs refer to Transformer-based neural language models that contain hundreds of billions (or even more) of parameters. The development of LLMs represents a significant milestone in the domains of Natural Language Processing (NLP) [5] and Machine Learning, empowering the emotionless machines to perform sophisticated tasks such as creative writing, reasoning, and decision-making, arguably comparable to human level [50]. They are typically trained on extensive text data, often consisting of books, articles, and web content, to capture the grammatical structures, inter-word relationships, and contextual nuances of language, benefiting from the self-attention mechanism, which excels at capturing long-range dependencies in text, and massive parameters of its deep neural network architecture. The inherent architecture of language models has undergone multiple iterations, continuously exhibiting strong capacities to understand natural language and solve complex tasks via text generation.

Traditional language models, such as RNNs integrated with word embeddings, made remarkable effectiveness in solving various NLP tasks. Nevertheless, during their training, RNNs are plagued by the vanishing and exploding gradient problems, which hinder their capacity to capture long-term dependencies and contribute to training instability. The advent of the Transformer architecture, an end-to-end learning architecture inspired by word embeddings and sequence models, proposed by Vaswani et al. in 2017 [40], save the day by introducing the self-attention mechanism. To be more specific, the Transformer architecture comprises multiple layers, with each layer containing several heads (called Attention Heads) that process different types of information in parallel, enhancing the model's capacity to handle long-sequence information. Categorized by the components, here are some kinds of Transformer list below:

1. **encoder & decoder**: T5 [32], BART [19]
2. **encoder-only**: BERT [16], RoBERTa [23], ALBERT [18]
3. **decoder-only**: GPT-x [1, 4]

With countless explorations, researchers found that the Transformer architecture, where multi-head attention layers are stacked in a very deep neural network, is highly scalable, not only for the scalability of model size & dataset size, but also about the amount of compute used for training, and all these can lead to substantial improvements of the model capabilities. Generally speaking, we regard the language model with more than one billion parameters as a LLM, and when it scales up properly, new abilities of the model will emerge, according to *KM scaling law* [14] and *Chinchilla scaling law* [11], which give birth to following LLMs and other Foundation Models.

2.2.2. Multimodal Large Language Models (MLLMs)

Although LLMs demonstrate remarkable capabilities in text and language related tasks, these models taking textual input only are not powerful enough to be constituent part of agent systems such as embodied AI. To enhance the efficiency and generalization ability of these intelligence systems, lots of researchers are no longer limiting themselves to using text as the only format of input. Instead, they are integrating images, videos, point clouds [42, 43], and even voice prompts into their attempts.

Multimodality often refers to the combination and utilization of multiple sensory modalities in proprioception. By leveraging this expressive information, such as visual, auditory, and tactile cues, embodied agents can gather a more comprehensive understanding of their environment.

This multimodal perception enables them to make sense of complex spatial information, recognize objects and events, and finish tasks like navigate, locomotion and so on effectively in real-world scenarios [37].

Different from LLM, the datasets for MLLMs training are formatted with pair-wise, including non-textual data (e.g., images, videos, point clouds) and corresponding textual descriptions. Due to the variety of data modalities, an extra encoder is needed for MLLMs to transform the input into a unified representation for the subsequent parts of the system to process the data. An extra decoder is also required when the models, such as DALL-E [34], are demanded to make generations beyond text.

3. Probelm Definition

As mentioned in Section 2, there exist various types of tasks in the field of embodied intelligence. Among myriad research directions, navigation has its unique role and is widely studied. So what is the accurate definition of navigation tasks?

In the context of embodied intelligence, the navigation task refers to the ability of an agent—whether a robot, virtual entity, or biological organism—to plan, decide, and execute movements to reach a specific destination or perform a goal-oriented activity within an physical environment. Unlike traditional tasks where agents work with abstract data, navigation requires agents to interact with their surroundings in real-time, considering factors like obstacles, terrain, and spatial constraints. Nowadays, navigation is a relatively mature robot skill but play a fundamental role in the field of embodied intelligence, as it has strong capabilities to explore the circumstances and reach the desired or task-required places, being a prerequisite for many other downstream tasks, such as object manipulation.

A pivotal aspect of navigation tasks is the integration of multimodal sensory information, involving visual information, auditory cues, or tactile feedback, all of which help the agent build a map or representation of its environment. The agent must then use this information to calculate the best path to its goal, often in dynamic or unpredictable settings, and determine to go forward, turn or stop, given the past observations at each time step.

For instance, autonomous robots need to navigate through rooms or outdoor environments, avoiding obstacles and efficiently reaching destinations. Similarly, in virtual environments, navigation tasks can be used to train AI agents to move through 3D spaces, such as in video games or simulators.

Moreover, navigation can be take part in several decision-making processes like pathfinding (utilizing some search algorithms to find the shortest path on the map), localization (determining the agent's position within the environment), and motion planning (choosing the optimal movement strategy). It requires algorithms like A* or reinforcement learning, which help the agent learn to make decisions based on past experiences or by optimizing for specific objectives.

Nowadays, navigation has evolved into a variety of solutions, as follows:

3.1. Map Based Navigation

Map-based navigation is a traditional navigation method that relies on a pre-constructed map of the environment, which can be represented as grid maps or topological maps in practice. Grid maps divide the environment into a grid of cells while topological maps represent the environment as a graph of nodes and edges. Given a precise map, path planning algorithms such as A* and Dijkstra are used to find the shortest path from the start to the goal.

3.2. Sensor Based Navigation

With the improvement of electronic component manufacturing processes, the precision of various sensors has been enhanced greatly, which makes it possible to integrate sensors to finish navigation tasks. Classical navigation systems are usually built with simultaneous localization and mapping (SLAM), making use of various kinds of sensors such as LIDAR, cameras, and IMUs to perceive the environment.

3.3. Visual Language Navigation (VLN)

Based on natural language instructions, VLN is a helpful tool for robots to navigate to a desired location. The model is commonly trained with large-scale datasets such as Room-to-Room (R2R) [2] dataset and so on, which contain pairs of environments and corresponding instructions. After the sequence model is trained, the robot can understand the instructions in natural language and establish an association between the instructions and image observations, which is beneficial for the robot to make sequential actions and navigate to the target location successfully.

3.4. LLMs & MLLMs Based Navigation

In recent years, with the demand for real applications increasing and the surroundings becoming more complex with multimodal information, traditional methods are not competent enough. Thanks to the powerful capabilities brought by advancement and iterations of LLMs and MLLMs, the demands of navigation tasks, namely a deep understanding of the environment and quick, accurate decision-makingm, are satisfied. Robust language and image processing abilities are exactly what LLMs and MLLMs are skilled in.

In order to gain a more vivid understanding of navigation tasks, we can divide the whole task into three basic problems to resolve, that is, "Where am I?", "What is my surroundings?" and "What should I do next?". [46] The first problem is about localization, which is to determine the agent's position in the environment. The second problem is about perception, which is to understand the surroundings of the agents. The third problem is about decision-making, which is to decide the next action based on the agent's current state, the target as well as the history information. Therefore, a series of core technologies such as environmental perception, map creation, autonomous positioning, and motion planning are required, which can be perfectly provided by foundation models like LLMs and MLLMs, given that LLMs show great potential in reasoning and decision-making while MLLMs are designed for encode information in rich modalities into the same vector space, which is conducive to cross-modal information processing.

In this survey, I mainly focus on state-of-the-art navigation methods of the last type, the LLMs & MLLMs based one, which have shown great potential in the field of embodied intelligence and promising applications for the society.

4. State of the Art Methods

In the field of embodied navigation, state-of-the-art models enhance their performance and practicality through various methods such as

1. **Multimodal integration**: The combination of visual, linguistic, and auditory data, hugely improves the agent's understanding of the environment and the decision-making process.
2. **Deep learning algorithms involvement**: The powerful comprehension and reasoning capabilities of LLMs make lots of contributions in handling the complex navigation tasks.
3. **Long-term memory mechanisms**: This special mechanism enables the agent to remember past navigation experiences and features of surroundings, having a positive effect on path planning in changing environments.
4. **Reinforcement learning**: The adoption of reinforcement learning methods allows models to self-optimize and learn through continuous interaction with the environment, providing the model with the ability to adapt to new environmental changes as well.

In this section, I will introduce the most representative state-of-the-art methods in the field of LLMs & MLLMs based navigation, categorized by the core problems they aim to solve, including planning, semantic understanding, automatic localization.

- **Planning**: In this context, the term "planning" is not as broad as that mentioned in Section 2.1.2, but refers to the process by which a robot intelligently selects an optimal sequence of actions to move towards a target location based on its current position within a reference frame, directly generating actions and leveraging exploration policies to guide agents. Below are some representative works in this field:

1. **CLIP-Nav** [7] proposes a groundbreaking "zero-shot" navigation scheme, in order to solve coarse-grained instruction. The whole architecture is organized as follows: Firstly, breaking down the guidance in natural language into a set of keyphrases. Secondly, visually grounding them within the environment so that we can acquire the grounding scores. Finally, take advantage of these resulting scores to direct **CLIP-Nav**. Following this pipeline, the sequence-to-sequence model in CLIP-Nav predicts a subsequence of actions for the agent to take, under the guidance of current state of the agent and the grounding scores.

   Furthermore, the introduction of *backtracking* mechanism drastically improves model's performance, which allows the agent to retrace its steps, facilitating revisions to prior decisions, showcasing the necessity of backtracking in solving such tasks.

   Due to its innovation on "zero-shot" framework, newly established benchmarks and evaluation metrics are needed. This paper proposes a zero-shot baseline on the task of REVERIE [29], and a metric named Relative Change in Success (RCS) to measure generalizability of the model.
2. **NavGPT** [51] introduces a novel instruction-following agent based on large language model to deal with visual navigation with a supportive system to interact with the environment and track navigation history. This paper mainly focuses on investigating the strengths and limitations of LLMs' reasoning abilities for making navigation decisions, under the intricate and embodied contexts.

   Lots of experiments show that NavGPT excels in dissection of instructions into sub-goals, landmark identification in observed scenes, navigation progress tracking, and adjustments to plans based on unexpected developments. So as to make a better performance in zero-shot Room-to-Room (R2R) tasks, the researchers advocate the combination of multimodal inputs and the application of LLMs' explicit reasoning to train learning-based models.
3. **SayNav** [33] is pioneering framework of LLM-based high-level planner specifically for navigation tasks in large-scale expansive and unknown photo-realistic environments. The LLM-based planner incrementally generates step-by-step instructions, which are consistent and non-redundant, in a dynamic manner during the navigation process.

   What's more, this paper employs a unique grounding mechanism to LLMs, who incrementally expands and builds a 3D accurate map of explored terrains and

Category	Works	Design Structure	Zero-shot	Multimodal
Planning	**CLIP-Nav** [7]	CLIP, GPT-3	✓	✓
	NavGPT [51]	GPT-4	✓	✓
	SayNav [33]	GPT-3.5-turbo, GPT-4	✗	✓
	VELMA [35]	CLIP, GPT	✗	✓
	Mic [30]	GPT-3	✓	✓
Semantic Understanding	**LM-Nav** [36]	ViNG, CLIP, GPT-3	✗	✓
	ESC [52]	Deberta	✓	✓
	L3MVN [49]	RoBERTa	✗	✓
	SQA3D [25]	CLIP, GPT-3	✓	✓
Automatic Localization	**CoW** [9]	CLIP	✓	✓
	AnyLoc [15]	DINO, DINOv2	✗	✓
	WAY [10]	GPT-4	✗	✓

Table 1. Comparison of State-of-the-Art Navigation Methods

obstacles, which will be also fed back to the LLMs to refined and updated the model itself subsequently.

As for the low-level planner, it is responsible for generating brief motion control commands. It takes RGBD images as well as its position data as inputs and give out basic navigational commands in alignment with standard PointNav configurations. To unify two parts of planners, SayNav treats each pseudocode instruction from the LLM as a short-distance point-goal navigation sub-task.

An excellent benchmark dataset for MultiON task across different houses is established in this paper, which not only evaluates the performance of SayNav, but also make this field more standardized and comparable.

4. **VELMA** [35] is an LLM-based embodied intelligence agent designed for urban VLN in Street View settings. The agent navigates based on human-generated instructions, which include landmark references and directional cues. The system identifies landmarks from human-authored navigation instructions and employs CLIP to assess their visibility in the current panoramic view, achieving a linguistic representation of visual information.

   The paper introduces a landmark scorer to assess landmark visibility within panoramic images. This scorer calcu- lates similarity metrics between textual descriptions of landmarks and their visual representations, using CLIP models. Each landmark receives a normalized score based on visual similarity. If this score exceeds a predefined threshold, the landmark is deemed visible.
5. **MiC** [30] (March in Chat) is an environment-aware instruction planner that employs an LLM for dynamic dialogues, specifically designed for the REVERIE dataset. The architecture of MiC is trifurcated into three mod- ules: Generalized Object-and-Scene-Oriented Dynamic Planning (GOSP), Step-by-step Object-and-Scene-Oriented Dynamic Planning (SODP), and Ro-om-and-Object Aware Scene Perceiver (ROASP).

   Initiating with GOSP, it queries the LLM to ascertain the target object and its probable locations, subsequently generating a rudimentary task plan. The prompt for SODP is tripartite: the first part utilizes ROASP for scene percep- tion, acquiring room types and visible objects, and translates this information into a natural language description. The second part involves the generation of fine-grained step-by-step instructions based on the selected strategy. The final part includes previously generated instructions. These elements are concatenated and input into the LLM, which then produces detailed planning instructions for the ensuing step.

- **Semantic Understanding**: Semantic understanding is a crucial part of navigation tasks and in this context, researchers usually employs LLMs to scrutinize incoming visual or textual data to isolate goal-relevant information, based on which exploration policies subsequently produce suitable actions for agent navigation. Here are some works making great advancements:

  1. **LM-Nav** [36] proposes an embodied instruction following navigation system, called Large Model Navigation (LM-Nav), which consists of three large independently pre-trained models: a robotic control model that utilizes visual observations and physical actions (VNM), a vision-language model that grounds images in text but has no context of embodiment (VLM), and a large language model that can parse and translate text but has no sense of visual grounding or embodiment (LLM). Remarkably, the system eliminates the need of costly supervision and fine-tuning, relying solely on pre-existing models for navigation, image-language

correlation, and language modeling.

Numerous experiments demonstrate that LM-Nav, implemented on a real-world mobile robot, has the ability to accomplish long-horizon navigation in intricate, outdoor settings solely based on natural language directives. However, due to the absence of fine-tuning, all three component models must be trained on extensive datasets to achieve the desired performance.

2. **ESC** [52] stands for the first letters of "Exploration with Soft Commonsense constraints". This paper introduces a outstanding approach to zero-shot object navigation by leveraging pre-trained models for open-world understanding and navigation, as well as object-level and room-level commonsense knowledge reasoning. The prominent feature of ESC is the integration of Frontier-based Exploration and Probabilistic Soft Logic, which are training-free, to model the soft commonsense constraints.

   Unlike other models, ESC still perform well in commonsense reasoning with novel objects or settings, benefiting from the utilization of pre-trained visual and language models for open-world.

3. **L3MVN** [49] introduces a pioneering module framework that capitalizes on large language models to enhance visual target navigation. Thanks to the mighty understanding and reasoning capabilities of LLMs, L3MVN is able to construct the environment map and select the long-term goal based on the frontiers given by the LLMs, to achieve efficient exploration and searching.

   The architecture comprises two principal modules: a language module and a navigation module. The former handles natural language instructions, generating a semantic map embedded with general physical world knowledge. The latter employs this semantic map to guide robotic exploration, deducing the semantic pertinence of visible frontiers and opting for the most cost-efficient maneuvers.

4. **SQA3D** [25] (Situated Question Answering in 3D Scenes) introduces a task formulated to assess the scene comprehension aptitude of embodied agents. The task necessitates that the agent garner an exhaustive understanding of its orientation within a 3D environment, guided by a text-based description, and subsequently generate precise answers to questions pertaining to that understanding.

   The principal objective of SQA3D is to gauge the capacity of embodied agents to engage in logical reasoning about their immediate environment and generate answers based on such reasoning. Unlike most existing tasks, which presume that observations are made from a third-person viewpoint, SQA3D uniquely demands that agents construct and reason from an egocentric perspective of the scene.

- **Automatic Localization**: Robot localization refers to a robot's ability to accurately determine its position and orientation within a predefined reference frame. The works listed below are representative in this field:

  1. **CoW** [9] is an innovative approach that adapts zero-shot visual models like CLIP to embodied AI tasks, particularly object navigation. The framework involves an agent identifying a target object in unfamiliar environments, defined through text. The key concept is dividing the task into zero-shot object localization and exploration. Although CLIP is adept at matching images and text, the integration of CLIP with the system still meets challenges. Firstly, CLIP struggles with precise spatial localization which is of vital importance for steering the agent. Secondly, CLIP, being static image-based, lacks mechanisms for processing dynamic scenario and guiding exploration. Finally, conventional fine-tuning of CLIP might reduce its robustness and generalizability.

     To address these puzzles, the CoW framework avoids fine-tuning and uses three techniques for object localization: Gradient-based (using CLIP gradients for saliency mapping), k-Patch-based (discretizing the image into sub-patches for individual CLIP model inference), and k-Language-based (matching the entire image with various captions for location information). For exploration, CoW leverages depth maps with two methods: Learning-based (incorporating a GRU, linear actor, and critic heads) and Frontier-based (a top-down map expansion approach).

  2. **AnyLoc** [15] emerges as a new baseline of Visual Place Recognition (VPR [24]) method that works universally across 12 datasets, with both structured and unstructured environments, exhibiting massive diversity along the axes of place, time, and perspective, without any fine-tuning or retraining.
     Self-supervised features (such as DINOv2 [28]) and unsupervised aggregation methods (like VLAD [12] and GeM [31]) are both crucial for strong VPR performance. AnyLoc employs these aggregation techniques on per-pixel features offers substantial performance gains over the direct use of per-image features from off-the-shelf models.

  3. **WAY** [10] focuses on the task of Localization From Embodied Dialog (LED), with the "Where Are You" (WAY) dialogue dataset as the training data proposed in this paper. Main contributions of WAY can be summarized as follows: Firstly, the WAY dataset consists of 6k dialogs in which two humans with asymmetric information complete a cooperative localization task in reconstructed 3D buildings. Secondly, it defines three challenging tasks: Localization from Embodied

Dialog (LED), Embodied Visual Dialog, and Cooperative Localization.

- **Comparision** The Table 1 exhibites a comparison of the state-of-the-art navigation methods in terms of their design structure, zero-shot capability, and multimodal integration, categorized by planning, semantic understanding, and automatic localization.

5. Challenges And Open Issues

Although LLMs & MLLMs based navigation has made remarkable process in the past few years, these expressive models and excellent methods still face many challenges and there still exist many unknown spaces to be explored. In this section, I will discuss the challenges and open issues at present.

1. **High obstacles**: Present SOTA methods are mainly focused on esay terrain with few obstacles. When it comes to legged robots, which is able to traverse uneven surfaces like stairs and obstacles with different shape, little research has specifically addressed the navigation tasks in such scenarios. This demand makes sense in real-world applications, which have countless complex and unknown environments.
2. **Requirment of high-quality data**: Lots of researchers found that the training of the LLM has a strong demand for high-quality data, and LLMs will perform poorly in navigation tasks when they get inaccurate or insufficient data. Although most of nowadays works take advantage of pre-trained models to construct the whole navigation system, these pre-trained models are not directly and task-specifically trained, leading to adverse effects on the final navigation results.
3. **Efficiency and resource consumption**: There exists a trade-off between the efficiency and the performance of the navigation system. Compared with non-LLM approaches, the LLMs based one can gain better understanding of the environment and make more accurate decisions. However, they are usually more computationally expensive and less efficient, making in-time navigation difficult to achieve. On the other hand, although the high parallelism of Transformer based LLMs, the large-scale dataset and models still require a large amount of computational resources.
4. **Dynamic environment and interactive navigation**: Most of the existing methods are based on the assumption that the environment is static. LLMs are not sufficiently adapted to interactive navigation and show limitations in navigation tasks in environments containing traversable obstacles, which are more than essential for auto driving and similar applications.
5. **Generalization and transferability**: The generalization and transferability of the LLMs based navigation system are still a big challenge. There is no doubt that we cannot train a model for every possible environment, whether the model can be generalized to unseen environments is a key point. Due to the noise of sensors and the lack of global map, this problem still remains unsolved.

With the above listed challenges we still face, here are some open issues that await for further research:

1. **Dynamic environment navigation**: As mentioned above, most current works focus on static scenarios. How to make the LLMs based in-real-time navigation integrated with dynamic environment and obstacles with high difficulties will be a promising direction.
2. **A convincing theoretical framework**: As known to all, the decision-making process of LLMs is still a black box, make researchers and developers unable to fully understand its internal mechanisms. With the lack of a convincing theoretical framework, the trust and acceptance of LLMs will be weakened. Therefore, to develop a self-explainable theoretical framework is an urgent task for theoretical researchers.
3. **Efficiency optimization**: As green AI is becoming a hot topic, the efficiency optimization of our navigation system is a major trend. An inspiring direction is to adjust the model architectures and using LLMs trained for a specific task, which is more likely to be helpful to reduce resource consumption and improve efficiency.
4. **Sophisticated automated driving**: With the development of the society, the demand for automated driving is increasing. The methods making improvements on navigation performance such as fusing multimodal information collected from GPS radar and all-around cameras will have a widely effect on the whole society.
5. **Data security and user privacy protection**: This is not only a technical issue, but also a ethical problem. As mentioned above, the mechanism of LLMs is still a black box, whether the data and privacy of users are safe is what we should pay close attention to. The research of stringent methods to ensure data security and protect user privacy from being misused, leaked or tampered with is of vital significance for the future.

6. Conclusion

This paper carries on a comprehensive survey of the state-of-the-art methods for navigation using LLMs and MLLMs. The author carefully selected 12 most representative works out of countless references and after thorough reading, analysing and finally composed this survey consisting of the background, the definition, the state-of-the-art methods, in which the author detailedly analyses the structure of the models and the outstanding contributions, and the challenges and the open issues, which provides promising research directions in this domain. The latter two parts are the hightlights of the whole paper. I hope this survey will be useful for researchers and practitioners working in the field of navigation using LLMs and MLLMs and help in the development of new methods.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] 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. 4

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 2

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[5] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020. 3

[6] Grace Colverd, Paul Darm, Leonard Silverberg, and Noah Kasmanoff. Floodbrain: Flood disaster reporting by web-based retrieval augmented generation with an llm. *arXiv preprint arXiv:2311.02597*, 2023. 1

[7] Vishnu Sashank Dorbala, Gunnar Sigurdsson, Robinson Piramuthu, Jesse Thomason, and Gaurav S Sukhatme. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*, 2022. 5, 6

[8] 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. 2

[9] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*, 3(4):7, 2022. 6, 7

[10] Meera Hahn, Jacob Krantz, Dhruv Batra, Devi Parikh, James M Rehg, Stefan Lee, and Peter Anderson. Where are you? localization from embodied dialog. *arXiv preprint arXiv:2011.08277*, 2020. 6, 7

[11] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 3

[12] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010. 7

[13] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: Robot manipulation with multimodal prompts. 2023. 2

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

[15] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023. 6, 7

[16] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2. Minneapolis, Minnesota, 2019. 3

[17] Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. *arXiv preprint arXiv:2310.11220*, 2023. 2

[18] Zhenzhong Lan. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 3

[19] Mike Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 3

[20] Rong Li, Lei Zhao, ZhiQiang Xie, Chunhou Ji, Jiamin Mo, Zhibing Yang, and Yuyun Feng. Mining and analyzing the evolution of public opinion in extreme disaster events from social media: A case study of the 2022 yingde flood in china. *Natural Hazards Review*, 26(1):05024015, 2025. 1

[21] Jinzhou Lin, Han Gao, Xuxiang Feng, Rongtao Xu, Changwei Wang, Man Zhang, Li Guo, and Shibiao Xu. Advances in embodied navigation using large language models: A survey, 2024. 2

[22] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023. 2

[23] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019. 3

[24] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *ieee transactions on robotics*, 32(1):1–19, 2015. 7

[25] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 6, 7

[26] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward

design via coding large language models. *arXiv preprint arXiv:2310.12931*, 2023. 2

[27] 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. IEEE, 2024. 2

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[29] 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. 5

[30] Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. March in chat: Interactive prompting for remote embodied referring expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15758–15767, 2023. 6

[31] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 7

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[33] 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. In *Proceedings of the International Conference on Automated Planning and Scheduling*, pages 464–474, 2024. 5, 6

[34] Mr D Murahari Reddy, Mr Sk Masthan Basha, Mr M Chinnaiahgari Hari, and Mr N Penchalaiah. Dall-e: Creating images from text. *UGC Care Group I Journal*, 8(14):71–75, 2021. 4

[35] 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*, pages 18924–18933, 2024. 6

[36] 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. 6

[37] Yong Shi, Mengyu Shang, and Zhiquan Qi. Intelligent layout generation based on deep generative models: A comprehensive survey. *Information Fusion*, page 101940, 2023. 4

[38] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022. 2

[39] Soudabeh Taghian Dinani, Doina Caragea, and Nikesh Gyawali. Disaster tweet classification using fine-tuned deep learning models versus zero and few-shot large language models. In *International Conference on Data Management Technologies and Applications*, pages 73–94. Springer, 2023. 1

[40] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[41] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Chatgpt empowered long-step robot control in various environments: A case application. *IEEE Access*, 2023. 2

[42] Shaohu Wang, Fangbo Qin, Yuchuang Tong, Xiuqin Shang, and Zhengtao Zhang. Probabilistic boundary-guided point cloud primitive segmentation network. *IEEE Transactions on Instrumentation and Measurement*, 2023. 3

[43] Shaohu Wang, Yuchuang Tong, Xiuqin Shang, and Zhengtao Zhang. Hierarchical viewpoint planning for complex surfaces in industrial product inspection. *IEEE/ASME Transactions on Mechatronics*, 2023. 3

[44] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023. 2

[45] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023. 2

[46] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Qian Cheng, Bin He, and Shuo Jiang. Robot learning in the era of foundation models: A survey. *arXiv preprint arXiv:2311.14379*, 2023. 5

[47] Taozheng Yang, Ya Jing, Hongtao Wu, Jiafeng Xu, Kuankuan Sima, Guangzeng Chen, Qie Sima, and Tao Kong. Moma-force: Visual-force imitation for real-world mobile manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6847–6852. IEEE, 2023. 2

[48] Takuma Yoneda, Jiading Fang, Peng Li, Huanyu Zhang, Tianchong Jiang, Shengjie Lin, Ben Picker, David Yunis, Hongyuan Mei, and Matthew R Walter. Statler: State-maintaining language models for embodied reasoning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15083–15091. IEEE, 2024. 2

[49] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3554–3560. IEEE, 2023. 6, 7

[50] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 3

[51] 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*, pages 7641–7649, 2024. 5, 6

[52] 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. 6, 7