The Development of LLMs for Embodied Navigation

Jinzhou Lin, Han Gao, Xuxiang Feng, Rongtao Xu†, Changwei Wang, Man Zhang, Li Guo, Shibiao Xu †, *Member, IEEE,*

***Abstract*—In recent years, the rapid advancement of Large Language Models (LLMs) such as the Generative Pre-trained Transformer (GPT) has attracted increasing attention due to their potential in a variety of practical applications. The application of LLMs with Embodied Intelligence has emerged as a significant area of focus. Among the myriad applications of LLMs, navigation tasks are particularly noteworthy because they demand a deep understanding of the 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 capabilities. This article offers an exhaustive summary of the symbiosis between LLMs and embodied intelligence with a focus on navigation. It reviews state-of-the-art models, research methodologies, and assesses the advantages and disadvantages of existing embodied navigation models and datasets. Finally, the article elucidates the role of LLMs in embodied intelligence, based on current research, and forecasts future directions in the field. A comprehensive list of studies in this survey is available at https://github.com/Rongtao-Xu/Awesome-LLM-EN.**

***Index Terms*—Large Language Models, Embodied Intelligence, Navigation.**

I. Introduction

THE development of LLMs for embodied intelligence is a rapidly evolving field with significant potential for advancing both natural language processing and machine learning. Notably, LLMs have already achieved remarkable successes in Few-Shot Planning, enabling effective planning and decision-making for new tasks with minimal or no sample data. However, alongside these achievements, numerous technical and theoretical challenges persist. These include the integration of text, images, and other sensor data simultaneously, the reduction of latency for real-time applications, and the enhancement of training efficiency without sacrificing performance.

To address these challenges, researchers employ a diverse array of methods such as machine learning [25], reinforcement learning(RL) [20], and evolutionary algorithms [34]. These methodologies are aimed at developing agents that are capable of learning from experience and continually improving their performance over time. Concurrently, the application of LLM-based agents in dataset investigations has garnered considerable attention in recent years. Such intelligent agents utilize machine learning models to analyze and derive insights from large-scale datasets, thereby making significant contributions to various domains including data mining, natural language processing, and information retrieval.

In this paper, we review existing studies that have employed LLM-based agents for dataset investigations. These studies have demonstrated the efficacy of LLM-based agents in tasks like sentiment analysis, topic detection, and entity recognition. The success of these studies underscores the utility of LLM-based agents as invaluable tools for data analysis and knowledge extraction. Datasets like MP3D [9], TOUCHDOWN [11], R2R [2], CVDN [70], REVERIE [55], RXR [36], SOON [90], ProcTHOR [18], R3ED [86], and X-Embodiment [16] offer unique opportunities for exploring the capabilities of LLMs in real-world settings. These datasets feature complex and diverse linguistic content and provide rich, authentic environmental information.

Furthermore, we examine various techniques and algorithms employed in LLM-based agents, such as Long Short-Term Memory (LSTM) [5], Convolutional Neural Networks (CNN), Contrastive Language-Image Pre-training (CLIP), and attention mechanisms. These methods enhance the agents' capabilities to process complex datasets, resulting in more accurate and meaningful outcomes. For example, ESC [88] uses a pre-trained commonsense reasoning language model for spatial and object reasoning and employs probabilistic soft logic to model ”soft” commonsense constraints, thus aiding traditional exploration methods in zero-shot decision-making. SayNav [60] leverages commonsense knowledge stored in LLMs for versatile navigation solutions. By constructing a 3D scene graph, generating high-level navigation plans, and executing short-distance point-goal navigation tasks, SayNav enables efficient and flexible navigation in unfamiliar or intricate environments. By understanding these benchmarks in-depth, w e offer comprehensive knowledge and resources for researchers and practitioners.

In summary, this paper serves as a valuable resource for both researchers and practitioners in the field of embodied intelligence. It offers a comprehensive review of recent advancements in research. Our aim is to provide an overview

Jinzhou Lin and Han Gao are co-first authors.

Rongtao Xu and Shibiao Xu are the corresponding authors (xurongtao2019@ia.ac.cn; shibiaoxu@bupt.edu.cn).

Jinzhou Lin, Han Gao, Man zhang, Li Guo and Shibiao Xu are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China.

Xuxiang Feng is with the Aerospace Information Research Institute, Chinese Academy of Science.

Rongtao Xu and Changwei Wang are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China.

of the existing literature in this rapidly expanding field, while also pointing out potential avenues for future exploration. It is structured as follows: Section 1 introduces the paper's purpose and discusses key technical and theoretical challenges. Section 2 focuses on the application of LLMs in Few-Shot Planning, outlining recent successes and ongoing challenges. Section 3 analyzes the structure of each dataset and offers comparisons, emphasizing their unique compositions and strengths. The final section provides an overview of the emerging field of embodied AI, highlighting key challenges and opportunities for future research. Our contributions are as follows.

1) We have summarized the evolutionary trajectory of LLMs, as well as their applications in the realm of embodied intelligence.
2) We have presented a selection of currently popular benchmarks and carried out a comparative evaluation among them.
3) A comparative analysis and introduction of commonly used datasets in LLMs for Embodied Intelligence are provided.

II. Background

A. Large Language Models

The development of LLMs represents a significant milestone in the domains of Natural Language Processing (NLP) [15] and Machine Learning. Tracing the evolution of these models requires examining the early stages of NLP and machine learning, where methods such as Bag-of-Words (BoW) [29] were prevalent for text representation but limited in capturing word order and contextual nuances.

Before the mainstream adoption of deep learning, simpler algorithms like N-grams and decision trees held sway in a variety of language-processing tasks. N-grams served as foundational elements for text and language data analysis by breaking down text into sequences of N contiguous words or characters. Although efficient and straightforward, N-grams suffer from limitations such as failing to capture long-distance dependencies and requiring a large parameter space for substantial N values.

Decision trees have been employed for classification and regression tasks. These trees recursively partition the dataset, where each internal node denotes a feature, each branch symbolizes a decision rule, and each leaf node represents the final output. While decision trees offer ease of interpretation and handle irrelevant features effectively, they are susceptible to overfitting and struggle with non-linear, high-dimensional data.

Both models have their merits and drawbacks, but they are now primarily used as baseline models or for feature engineering in the wake of more complex models enabled by deep learning.

The advent of word embedding models around 2013, notably Word2Vec and GloVe, signaled a transformative shift in NLP. These models allowed for words to be mapped into a continuous vector space, thereby capturing intricate word relationships with remarkable effectiveness.

Word2Vec was particularly impactful, offering an efficient method for encapsulating complex semantic and syntactic relationships between words. Due to its efficient training algorithms, Word2Vec could be applied to vast datasets, yielding word vectors that could be leveraged across numerous NLP applications such as text classification, named entity recognition, and machine translation. Nonetheless, Word2Vec is not without limitations; it provides a single vector representation for each word and requires large, labeled datasets for effective training.

GloVe, proposed by Pennington et al. in 2014, sought to combine global statistical information with local semantic details to produce high-quality word vectors. Unlike Word2Vec, GloVe optimizes a linear model directly, making it highly scalable and applicable to large datasets. However, it shares some limitations with Word2Vec, including the generation of a single vector per word and the requirement for substantial storage and computational resources due to its large co-occurrence matrix.

With the advent of word embedding techniques such as Word2Vec and GloVe, the focus shifted toward incorporating these pre-trained word vectors into sequence models like Recurrent Neural Networks (RNNs) and LSTMs. RNNs are neural architectures specifically designed to handle sequence data by utilizing the temporal context from prior elements in a sequence for processing subsequent inputs. 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.

To mitigate these drawbacks, specialized RNN variants such as LSTMs and Gated Recurrent Units (GRUs) [14] have gained prominence in sequence data analysis, particularly within NLP. LSTMs, equipped with gating mechanisms, effectively circumvent the vanishing gradient issue but at the cost of added computational complexity and the potential for exploding gradients. The increased number of parameters also increases the risk of overfitting and extends training time.

Building on word embeddings and sequence models, a new paradigm emerged with the development of end-to-end learning architectures, most notably the Transformer model. Introduced by Google in 2017 [71], the Transformer architecture eschews recurrent and convolutional layers, opting instead for a Self-Attention mechanism to capture sequence dependencies. This design enables parallel data processing, resulting in a significant speed-up over RNNs and LSTMs. The Transformer's scalability also benefits from its layered architecture, facilitating the addition of layers or parameters to improve performance.

Following the Transformer, Google's Vision Transformer (ViT) in 2020 [22] and OpenAI's CLIP [57] in 2021 marked the model's extension from NLP to visual processing tasks. Both architectures employ Transformer-based text encoders and demonstrate the model's versatility across domains.

Among the various Transformer-based architectures, BERT [19] (Bidirectional Encoder Representations from Transformers), introduced by Google in 2018, garnered attention for its pre-training approach across diverse NLP tasks. Employing bidirectional encoders, BERT captures both preceding and succeeding contextual information for each word. While effective, BERT's complexity results in high computational requirements.

OpenAI's introduction of the GPT series [58] [59] [7] has been another milestone, particularly for text generation tasks. These models also exhibit ”zero-shot learning” capabilities, enabling them to perform certain tasks without task-specific training. However, for tasks requiring higher accuracy, fine-tuning is generally necessary.

GPT models have been deployed in a wide array of applications in NLP and artificial intelligence, such as chatbots, automated writing, and question-answering systems, and have opened new avenues in multi-modal learning and code generation.

In summary, large language models are increasingly shaping various domains, and their ongoing development continues to reveal new potential. The widespread deployment and diverse applications of these models not only underscore their immediate utility but also open exciting possibilities for future research

B. *Embodied Intelligence*

Embodied Intelligence is an emerging field focused on understanding and developing intelligent agents that interact closely with their environment. The field posits that intelligence is fundamentally an emergent property of an agent's interaction with its surroundings, rather than a characteristic inherent in isolated, abstract computations. Drawing from diverse disciplines such as neuroscience, psychology, robotics, and artificial intelligence, Embodied Intelligence aims to create novel models and algorithms to simulate intelligent behavior.

Recent progress in embodied artificial intelligence has leveraged advancements in natural language processing to convert human instructions into formats interpretable by physically embodied agents. Additionally, sophisticated techniques in object recognition and scene understanding have been employed to enhance agents' situational awareness. Research in this area also utilizes state-of-the-art algorithms in planning and decision-making to allow agents to navigate complex environments effectively.

Growing interest exists in merging LLMs like GPT-3 with Embodied Intelligence. Although powerful in natural language processing, LLMs typically lack direct engagement with the physical world, as their training data consist primarily of text. Integrating LLMs with embodied agents aims to create language models with enhanced context-awareness and adaptability, potentially transforming the landscape of natural language processing and intelligent agent behavior.

In the realm of Embodied Intelligence, the concept of an intelligent agent stands as a pivotal element. The control over these agents bifurcates into High-level and Low-level facets. High-level controls encompass task scheduling and strategy development, incorporating the likes of reinforcement learning, deep learning, and the emergent methodologies based on expansive language models. In contrast, Low-level control pertains to the direct command over the agent's operational functions, such as control over position, speed, and force. An agent's capabilities hinge upon its design and its Low-level control parameters [4], while the execution and approach to the completion of tasks are orchestrated by High-level controls [46].

These methodologies have unfolded a spectrum of distinct applications, encompassing terrain recognition [72], prediction of machinery lifespan [76] and emulation of gaze mechanisms [44]. These strides not only reflect the dynamic essence of the field but also signify the tangible impact that these controlled embodied agents procure across a multitude of domains.

The advancement of intelligent agents is contingent upon the harmonious integration of high-level and low-level controls. Utilizing superior algorithms and control techniques is essential for the development of new agents characterized by greater robustness and broader generalization capabilities.

For example, the LM-Nav [66] model proposed by Shah and Dhruv combines a self-supervised robotic control model, a vision-language model, and a large language model. Each component brings specific strengths: visual perception and physical interactivity from the robotic model, grounding of text to images from the vision-language model, and text parsing and translation from the large language model. LM-Nav thus enables long-horizon planning based on raw sensory inputs and free-form textual instructions, facilitating complex tasks in real-world settings.

In addition to research on single-agent systems, there is also research on multi-agent systems [65] [28] [62].Each of them independently focus on the multi-agent cooperation issue. Their respective research efforts are instrumental in broadening the spectrum of tasks achievable by agents, boosting work efficiency, and enhancing the real-world applicability and generalization of Embodied Agents.

One of the significant challenges in Embodied Intelligence lies in designing agents capable of real-time learning and adaptation to their environment. This calls for a deep understanding of sensory-motor coordination and morphological computation, which are foundational elements of embodied cognition. Researchers employ various techniques such as machine learning, reinforcement learning, and evolutionary algorithms to create agents that can learn from their experiences and improve performance over time.

III. LLMs in Embodied Agents

A. *LLMs for Grounded Language Understanding*

”Grounded Language Understanding” aims to reconcile the abstract symbols processed by language models with concrete entities, actions, or states in the physical or simulated world. This is pivotal for applications requiring real-world interaction, such as robotic control, natural language interfaces, or advanced research in Embodied Intelligence.

Within this context, LLMs like GPT-3, GPT-4, and BERT can integrate with sensors, databases, or simulated environments to generate and interpret language applicable to real-world scenarios. In a robotic setting, for instance, an LLM could interpret sensor data, process natural language directives, and issue control signals to the robot, thereby bridging high-level language and low-level actions.

LLMs' application in grounded language understanding is an evolving research area, intersecting with disciplines such as robotics, computer vision, and human-computer interaction. Techniques like Few-Shot Learning and Transfer Learning are commonly used to adapt these pre-trained models to specialized grounding tasks with minimal additional training.

Although LLMs have advanced significantly in NLP, they primarily excel in text generation or classification and seldom engage with tangible entities or real-world situations. Their limitations include a lack of foundational knowledge, impairing their ability to process interrelated concepts and function in interactive settings (Mahowald et al. [48]). To mitigate these issues, researchers have employed various strategies, such as 1) fine-tuning pre-trained models, 2) implementing reinforcement learning algorithms for decision-making in complex environments, and 3) devising specialized architectures for multimodal learning.

For example, Yang et al. [81] introduced LLM-Grounder, which integrates GPT-4 with CLIP for applications in 3D visual grounding. Carta et al. [8]utilized online reinforcement learning to boost LLM performance in reinforcement learning tasks, improving both sample efficiency and generalization.

Despite their immense potential and practical applications, LLMs face numerous technical and theoretical challenges in grounded language understanding. These include the integration of text, images, and sensor data, latency reduction for real-time applications, and maintaining training efficiency without sacrificing performance. With ongoing research and emerging technologies, significant advancements in this domain are anticipated.

B. LLMs for Few-Shot Planning

LLMs have achieved notable success in NLP and are increasingly being applied to Few-Shot Planning, a machine learning technique designed for effective planning and decision-making in new tasks with minimal sample data. LLMs excel in zero-shot and few-shot learning, allowing for efficient generalization across tasks without the need for task-specific training or limited examples. Employing LLMs for Few-Shot Planning often involves processing natural language queries, enabling the models to generate actionable plans or steps, thus proving invaluable in zero-shot navigation tasks.

As discussed in research by Zhou et al. in ESC [88] and Song et al. in LLM-Planner [69], the effectiveness of navigation tasks is closely related to semantic scene understanding, which relies heavily on Few-Shot Planning. Planners incorporating LLMs are increasingly vital for enabling embodied agents to execute complex tasks in visually rich environments based on natural language directives. The performance of such planners has a direct bearing on the task's overall success. The role of LLMs as planners in embodied tasks is expanding, thanks in part to the emergence of methodologies like LLM-Planner [69], CoT (Chain-of-Thought) [73], and LLM-DP [17]. These methods complement the already respectable performance of traditional symbolic planners like the Fast-Forward planner [31] and BFS(f) planner [41] in current research.

C. LLMs for Zero-Shot Navigation

In Zero-Shot Navigation, Large Language Models (LLMs) serve several crucial functions:

1) **Natural Language Understanding**: LLMs excel at interpreting intricate natural language instructions or queries, a skill vital for navigation in unfamiliar environments. For instance, given a directive like ”Find the nearest exit,” an LLM can parse this instruction and formulate an appropriate navigation strategy. Moreover, LLMs often contain a broad spectrum of general knowledge, useful for elementary reasoning and decision-making in zero-shot scenarios.
2) **Dynamic Planning**: Although primarily engineered for natural language processing, LLMs can also formulate or recommend a sequence of actions or movements, a capability invaluable for dynamic planning and route optimization.
3) **Multimodal Input**: Some applications enable LLMs to integrate with visual or other sensory data, providing a holistic navigation solution. For example, by processing natural language directives and evaluating visual information, an LLM can enhance accuracy in unfamiliar locales.
4) **Real-time Interaction**: LLMs can swiftly process natural language queries or directives in real-time, an attribute essential for dynamic planning operations in navigation.
5) **Task Generalization**: Due to their proficiency in zero-shot and few-shot learning, LLMs can quickly adapt to and generalize across new, unexplored navigation tasks and settings.

Consequently, in zero-shot navigation endeavors, LLMs function as multifaceted, highly adaptive units, taking on roles ranging from task interpretation and plan creation to real-time interaction.

1) Benchmarks: In recent years, a rapidly expanding array of benchmarks has emerged in the field of Zero-Shot Navigation, each characterized by its unique emphasis and methodologies for performance assessment. Generally, these benchmarks employ approaches that fall into two categories, as illustrated in Fig.2 and Fig.3. The first category uses LLMs as planners, directly generating actions and leveraging exploration policies to guide agents. The second category 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. An architectural diagram delineating these approaches is presented below. This section endeavors to examine the nuanced features and distinctions among these benchmarks, furnishing a comprehensive temporal map of the benchmarks under discussion in this chapter, as depicted in Fig.1. In doing so, we aim to offer an exhaustive understanding that will prove invaluable for both scholars and professionals in this field.

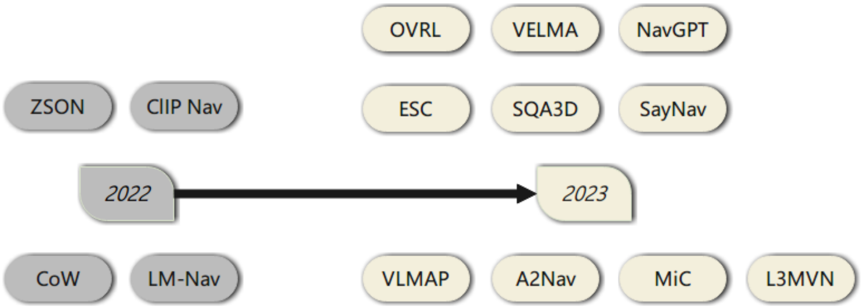

Fig. 1: This presentation will exhibit a temporal map depicting the benchmarks of Embodied Agents from 2022 to 2023. The map illustrates the evolution of major benchmarks, offering valuable insights into the advancement of Embodied Agents. These benchmarks serve as standardized evaluations for assessing the performance of Embodied Agents in tasks such as navigation, perception, and manipulation.

1)**CoW** [27] (CLIP on Wheels): Gadre et al. effectively transition the successes of zero-shot visual models, such as CLIP, to prevalent embodied AI tasks like object navigation. In their experimental setup, an agent must identify an arbitrary target object in unfamiliar environments sourced from diverse datasets, specified via text. A principal insight they present is the modular decomposition of the task into zero-shot object localization and exploration. However, applying the CLIP model directly to the task presents multiple challenges:

1) CLIP fails to offer precise spatial localization when an object is within the field of view, a capability crucial for steering the agent's low-level actions.
2) CLIP solely processes static images and lacks memory or alternate mechanisms to guide the agent's environmental exploration.
3) Conventional fine-tuning of the CLIP model could substantially undermine its robustness and generalizability, traits explicitly sought for leveraging from CLIP.

To surmount these limitations, the CoW framework segments the zero-shot object navigation task into zero-shot object localization and exploration. To maintain CLIP's generalizability, the CoW framework introduces three designs that incorporate the CLIP model directly as an object localizer, eschewing any fine-tuning. Further, Gadre et al. investigate two exploration paradigms: learnable and traditional methods. Their algorithm for the integration of object localization and exploration mirrors human-like strategies: CoW commences exploration when the object remains undetected and advances towards the target upon its identification.

For the localization module, CoW utilizes three techniques: Gradient-based, k-Patch-based, and k-Language-based. The Gradient-based method employs CLIP gradients to generate an image saliency map, while Grad-CAM accentuates essential regions for aligning image features with textual labels. In the absence of the object, Grad-CAM yields low-saliency predictions, mitigating the risk of false positives. The k-Patch-based approach necessitates only a forward pass through the CLIP model, obviating the need for backward gradient data. Its central concept involves discretizing the image into k subpatches and conducting inference on each. The k-Language-based technique employs CLIP to scan the entire image once, matching it with k distinct captions containing pertinent location information.

The exploration module leverages depth maps and employs two modes: Learning-based and Frontier-based. The Frontier-based approach is a top-down map expansion method, initially proposed by Yamauchi et al. [80]. The Learning-based approach incorporates a GRU, linear actor, and critic heads.

Overall, the CoW framework effectively repurposes zero-shot image classification models for popular embodied AI tasks. By modularizing the task into zero-shot object localization and exploration, and judiciously selecting methods for each component, it attains commendable success rates. While exploring deployments with user-specified targets is an intriguing avenue, the ultimate criterion remains real-world performance.

2)**ZSON** [49] introduces an innovative methodology for instructing virtual robots to navigate unfamiliar terrains and identify objects without pre-existing rewards or demonstrations. Distinct from conventional ObjectNav techniques, ZSON exploits image-goal navigation (ImageNav) to transcode goal images into a multimodal, semantic embedding space. This allows for the scalable training of semantic-goal navigation (SemanticNav) agents in unannotated 3D settings. The underpinning theory of ZSON hinges on the principle of semantic similarity. By encoding goal images as semantic embeddings, the approach enables agents to navigate toward objects predicated on their semantic likeness to the goal image. This strategy assumes that objects bearing semantic resemblance to the goal image are likely located in similar spatial contexts.

In the implementation phase, the first step entails using CLIP for pre-training to produce semantic embeddings of image targets. These embeddings encapsulate intricate semantic details about the objectives. Subsequently, a SemanticNav agent undergoes training through reinforcement learning. The agent ingests egocentric RGB observations and semantic target embeddings as input variables and utilizes a ResNet-50 encoder along with a policy network to forecast actions.

In performance evaluation, the team executed extensive experiments on three ObjectNav datasets: Gibson, MP3D, and HM3D. Their zero-shot agent recorded a 31.3% success rate in Gibson settings, reflecting a considerable 20.0% absolute uptick over earlier zero-shot benchmarks. In the MP3D dataset, the agent secured a 15.3% success rate, marking a 4.2% absolute enhancement vis-à-vis existing zero-shot paradigms. On the HM3D dataset, the agent's zero-shot Success weighted by Path Length (SPL) paralleled that of a state-of-the-art ObjectNav technique trained with direct guidance from 40,000 human demonstrations.

To conclude, the approach put forth by the researchers yielded commendable success metrics across the three ObjectNav datasets, constituting a significant divergence from traditional ObjectNav strategies. The model is scalable, zero-shot, and well-suited for application in unannotated 3D worlds, establishing it as a viable avenue for open-world ObjectNav implementations.

3)**LM-Nav** [66] is a navigation architecture designed for robots that leverages pre-existing models specialized in language, vision, and action to enable sophisticated interactions with robots via natural language commands. Remarkably, the system eliminates the need for costly supervision and fine-tuning, relying solely on pre-trained models for navigation, image-language correlation, and language modeling. The architecture of LM-Nav consists of three integrated, pre-trained models to ensure precise instruction execution in complex, real-world scenarios. Specifically, the system employs GPT-3 as the LLM, tasked with decoding verbal instructions into a series of textual landmarks. Concurrently, CLIP serves as the VLM, anchoring these textual landmarks to a topological map. Lastly, the Vision-Action Model (VAM) is a self-supervised robotic control model, responsible for utilizing visual data and executing physical actions based on plans synthesized by the LLM and VLM.

Implemented on a real-world mobile robot, LM-Nav has been shown to accomplish long-horizon navigation in intricate, outdoor settings solely based on natural language directives. A salient feature is the lack of model fine-tuning; all three component models are trained on expansive datasets with self-supervised learning objectives and are deployed as-is. The system has demonstrated its ability to interpret and execute natural language instructions across significant distances in complex, suburban terrain while also offering disambiguation in path selection through detailed commands. Such performance metrics underscore LM-Nav's robust generalization capabilities and its proficiency in navigating complicated outdoor environments.

4)**CLIP-NAV** [21] introduces an innovative ”zero-shot” navigation framework aimed at solving coarse-grained instruction-following tasks. The architecture is structured to dissect the guidance language into critical keyphrases, visually anchor them, and leverage the resulting grounding scores to direct the CLIP-Nav sequence-to-sequence model in predicting the agent's subsequent actions. Initially, a keyphrase extraction module isolates salient terms from the given instructions. Following this, a visual grounding module anchors these keyphrases within the environment, thereby generating a set of grounding scores. These scores serve as the basis for the sequence-to-sequence model in CLIP-Nav, which computes the next set of actions based on the agent's current state and grounding scores.

To augment the efficacy of CLIP-Nav, the model incorporates a backtracking mechanism. This allows the agent to retrace its steps, facilitating revisions to prior decisions. Such a feature is particularly beneficial in the context of coarse-grained instruction-following tasks, wherein corrective backtracking may be essential.

Evaluation metrics confirm the robustness of CLIP-Nav, establishing a zero-shot baseline on the REVERIE task. Remarkably, CLIP-Nav exceeds the performance of the unseen supervised baseline, even without dataset-specific fine-tuning, in both success rate and success weighted by path length. Additionally, the paper introduces a new performance metric, termed Relative Change in Success (RCS), to quantitatively assess generalizability in vision and language navigation tasks. The RCS metric substantiates the superior performance of CLIP-based methodologies over conventional supervised approaches.

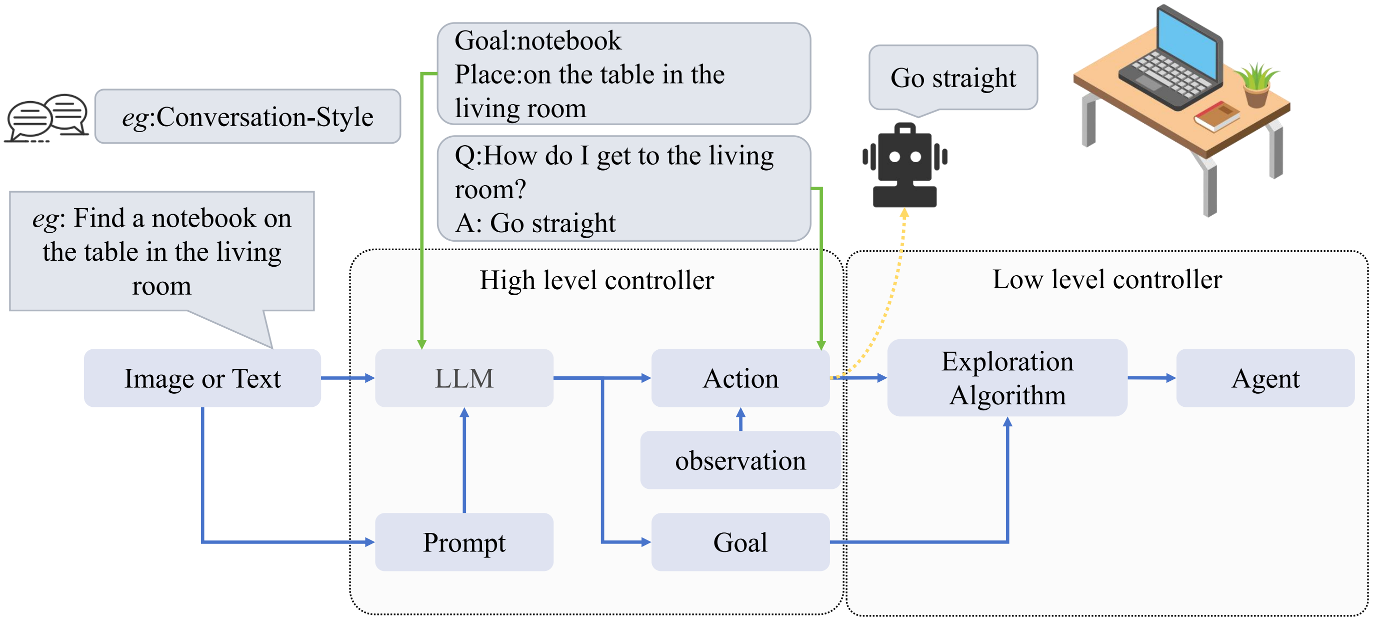

Fig. 2: The first type employs LLMs as planners that directly generate actions, thereby leveraging exploration policies to control agents.

5)**SQA3D** [47] (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.

SQA3D is underpinned by an extensive dataset derived from 650 scenes in ScanNet, comprising 6.8k distinct scenarios, 20.4k textual descriptions, and 33.4k multifaceted reasoning questions. The scope and diversity of the dataset enable the exploration of complex questions that demand a nuanced understanding and facilitate the aggregation of a substantial data corpus.

To address SQA3D's inherent challenges, researchers employed transformer-based vision-language models such as ScanQA, which generates answers by associating 3D scans with questions. In addition, auxiliary tasks have been considered during the training phase to amplify model performance. Specialized models attuned to various 3D scene contexts, including egocentric video clips and bird-eye view images, have also been deployed. Notably, researchers have explored the potential of recent language and vision models, including GPT-3 and Unified QA, to tackle the SQA3D task in a zero-shot setting.

Experimental analyses involving state-of-the-art multimodal reasoning models, such as ScanQA, were conducted to evaluate their efficacy on the SQA3D task against established benchmarks. The outcomes clearly establish SQA3D as a demanding task that necessitates complex reasoning about the agent's environment for accurate question answering.

6)**L3MVN** [82] introduces a pioneering module framework that capitalizes on large language models to enhance visual target navigation. The framework aims to solve the challenge of autonomous exploration in unfamiliar settings, where a robot, provided with an object's name, must proficiently navigate a 3D space to locate that object.

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.

L3MVN's cardinal innovation resides in its utilization of large language models, specifically GPT-3, to imbue the system with common-sense reasoning for object location tasks. GPT-3, a leading-edge language model, undergoes fine-tuning on a limited set of task-specific data to produce a semantic map replete with object-searching-relevant information.

The theoretical underpinning of L3MVN is anchored in the notion that human adeptness in intricate and dynamic environments stems partially from inherent comprehension of the physical world. This common-sense knowledge can be transferred to robots via large language models, enhancing their operational efficiency in unknown terrains.

Efficacy evaluations of L3MVN were conducted on two comprehensive photorealistic 3D scene datasets, namely Gibson and HM3D. Experimental outcomes indicate substantial enhancements in both the success rate and efficiency of visual target navigation, while obviating the need for expansive

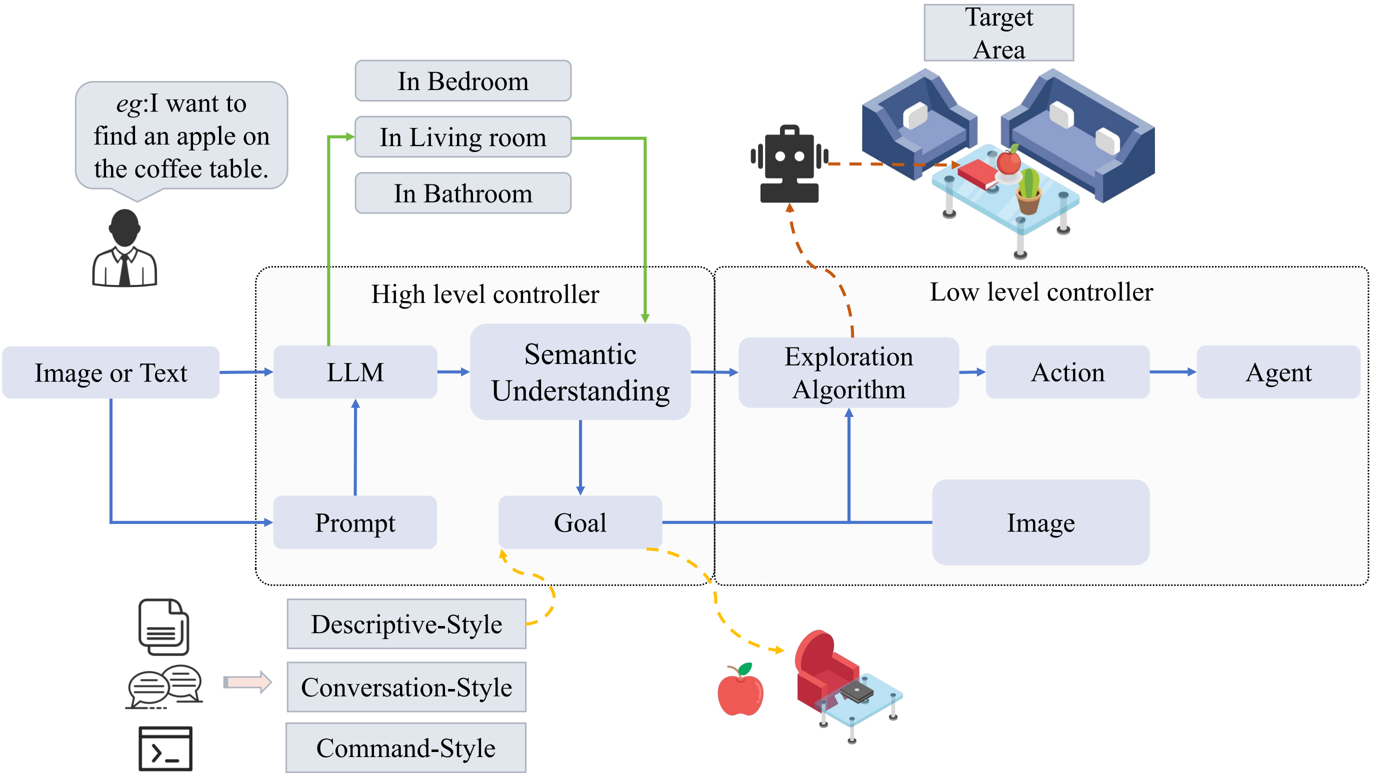

Fig. 3: The second type utilizes LLMs to analyze incoming visual or textual data to extract goal-relevant information, upon which exploration policies subsequently generate appropriate actions to guide agent movement.

learning processes. Additional ablation studies corroborate the efficacy of our language models and cost-utility exploration in facilitating more efficient navigation.

7)**VLMAP** [33] presents an innovative spatial map representation that integrates pretrained visual-language features with a 3D environmental reconstruction, facilitating natural language map indexing without the need for extra labeled data. The central innovation is its capacity to intrinsically combine visual-language attributes with the 3D environmental blueprint, empowering robots to execute spatially precise navigation via natural language directives.

The methodology for VLMAP consists of creating a spatial visual-language map capable of direct landmark or spatial reference localization via natural language. Employing readily available visual-language models and conventional 3D reconstruction libraries, VLMAP is assembled by merging visual-language characteristics with a 3D environmental model, thereby allowing natural language map indexing without supplemental labeled data.

VLMAP advances the fields of computer vision, natural language processing, and robotics by unifying visual-language features with 3D reconstructions for precise, natural language-guided robotic navigation. The incorporation of visual-language features leverages recent progress in deep learning, while the 3D reconstruction component builds on established principles of 3D point cloud processing. The natural language capabilities are supported by an expanding corpus of research in natural language processing and human-robot interaction.

Experimental evaluations of VLMAP substantiate its efficacy in enabling robots to navigate spatially precise paths through natural language commands. Tests in a simulated environment affirm the system's capability to convert natural language directives into a sequence of open-vocabulary navigation objectives that can be directly pinpointed on the map. Further, the adaptability of VLMaps for use among robots with disparate configurations to generate real-time obstacle maps is also demonstrated.

8)**OVRL** [79] introduces a neural network architecture specifically aimed at visual navigation challenges. Notably, the architecture is composed of task-agnostic elements such as ViTs, convolutions, and LSTMs, obviating the need for task-specific modules. The standout feature of OVRL-V2 is a novel compression layer designed to preserve spatial details in visual navigation assignments. This layer compresses high-dimensional image attributes from the ViT model into low-dimensional features, maintaining the essential spatial context.

The incorporation of the compression layer is grounded in the theoretical premise that retaining spatial information is vital for navigation endeavors, and this layer effectively conserves this attribute. The model further employs self-supervised learning for pretraining, which enhances its generalization performance across new environments.

Architecturally, the paper outlines a universal agent structure consisting of a visual encoder, a goal encoder, and a recurrent policy network. The visual encoder utilizes a ViT-based module for RGB data processing. This output is then amalgamated with a goal representation and channeled through a recurrent

LSTM policy network to predict actions. Training methods vary between tasks: IMAGENAV agents are trained via reinforcement learning and DD-PPO [74], while OBJECTNAV agents employ human demonstrations and behavior cloning for training. The visual encoder can either be trained de novo in an end-to-end fashion or pretrained via the MAE algorithm.

Experimental validations affirm OVRL-V2's state-of-the-art performance in both IMAGENAV and OBJECTNAV tasks, registering success rates of 82.0% and 64.0%, respectively. These results signify marked advancements over preceding methodologies, corroborating the efficacy of both the proposed architecture and its associated training techniques.

9)**ESC** [88] (Exploration with Soft Commonsense Constraints for Zero-shot Object Navigation) presents a groundbreaking approach to zero-shot object navigation by leveraging commonsense knowledge from pre-trained models for open-world navigation. Zhou et al. assert that effective object navigation hinges on two essential faculties: (1) Semantic scene understanding, crucial for recognizing objects and rooms, and (2) Commonsense reasoning, needed for logical inferences about probable locations of target objects based on general knowledge.

Previous zero-shot object navigation methods often fall short in commonsense reasoning. They either necessitate training on distinct goal-oriented navigation tasks and scenarios or depend on basic exploration heuristics. These approaches underperform when encountering novel objects or settings. Unlike these methods, ESC employs pre-trained visual and language models for open-world, hint-based grounding via the GLIP model [39], enabling scene comprehension and object grounding. Large-scale image-text pre-training further facilitates generalization to new objects.

Moreover, ESC deploys a pre-trained commonsense reasoning language model to deduce interrelationships between rooms and objects, applying this contextual data for spatial and object reasoning. However, translating this inferred commonsense knowledge into practical actions remains a challenge. Given that relationships between entities tend to be probabilistic rather than absolute, ESC utilizes Probabilistic Soft Logic (PSL [3]) to formulate ”soft” commonsense constraints. These are amalgamated with conventional exploration techniques like Frontier-Based Exploration (FBE) for informed zero-shot decisions on subsequent exploration frontiers.

By setting a new technical benchmark and pioneering the use of pre-trained commonsense knowledge in LLMs for object navigation, ESC paves the way for future research. This could include extracting intricate commonsense relationships, such as spatial connections between rooms, from LLMs. Additionally, future work might explore relaxing the zero-shot stipulations to permit limited fine-tuning in the learning of frontier selection strategies.

10)**NavGPT** [87] is an instruction-following visual navigation agent founded on Language Learning Models (LLMs). It aims to explore the reasoning capabilities of GPT models in intricate, embodied contexts via zero-shot sequential action prediction. Zhou et al. conducted extensive experiments to show that NavGPT excels in advanced navigation planning, including the dissection of instructions into sub-goals, the incorporation of common-sense knowledge pertinent to navigation tasks, landmark identification in observed scenes, navigation progress tracking, and adjustments to plans based on unexpected developments. The study also revealed LLMs' aptitude for generating precise navigation instructions from observations and actions, as well as mapping accurate top-down metric trajectories from navigation history.

Despite its robust capabilities, NavGPT's performance in zero-shot Room-to-Room (R2R) tasks does not yet rival that of specialized models. To address this, Zhou et al. advocate the use of LLMs with multi-modal inputs as visual navigation agents and suggest applying LLMs' explicit reasoning to learning-based models to boost performance.

Previous research efforts have aimed to harness GPT models for enhancing navigation tasks. LLMs have served as parsers for managing various language inputs, extracting landmarks from instructions for visual matching and planning, and capitalizing on their common-sense reasoning skills to enrich the agent's perception and decision-making capabilities. Nevertheless, the full potential of LLMs in navigation, especially their comprehension of interactive worlds, actions, and outcomes in text formats, and their utilization for resolving navigation challenges, remains unexplored.

Zhou et al. introduced NavGPT as a fully automated, LLM-based system designed for language-guided visual navigation. The system accommodates multi-modal inputs, unrestricted language guidance, interactions with open-world settings, and maintains a navigation history. Utilizing GPT-4 for intricate planning, NavGPT demonstrates proficiency in sequential action prediction. It excels in subdividing instructions into sub-goals, integrating navigation-relevant common-sense knowledge, recognizing landmarks in observed settings, tracking ongoing navigation progress, and making plan adjustments based on unexpected events. Furthermore, NavGPT demonstrates spatial and historical awareness by generating trajectory instructions and plotting navigated viewpoints in an overhead view.

The authors' exhaustive experimentation revealed LLMs' remarkable abilities for intricate navigation planning, including subdividing instructions into various sub-goals, incorporating relevant common-sense knowledge, recognizing landmarks in observed settings, continually monitoring navigation progress, and adjusting plans based on unforeseen occurrences. The study also confirmed that LLMs can construct navigation trajectories on metric maps and regenerate navigation instructions, thus demonstrating historical and spatial awareness in navigation assignments.

However, the zero-shot performance of currently available open-source LLMs lags behind specialized models. A significant bottleneck in NavGPT is the information loss incurred when translating visual signals into natural language and condensing observations into historical records. Consequently, Zhou et al. propose future research directions that include employing LLMs with multi-modal inputs and navigation systems that leverage the advanced planning, historical, and spatial awareness capabilities of LLMs.

11)**VELMA** [64] is an LLM-based embodied intelligence agent designed for urban Visual Language Navigation (VLN)

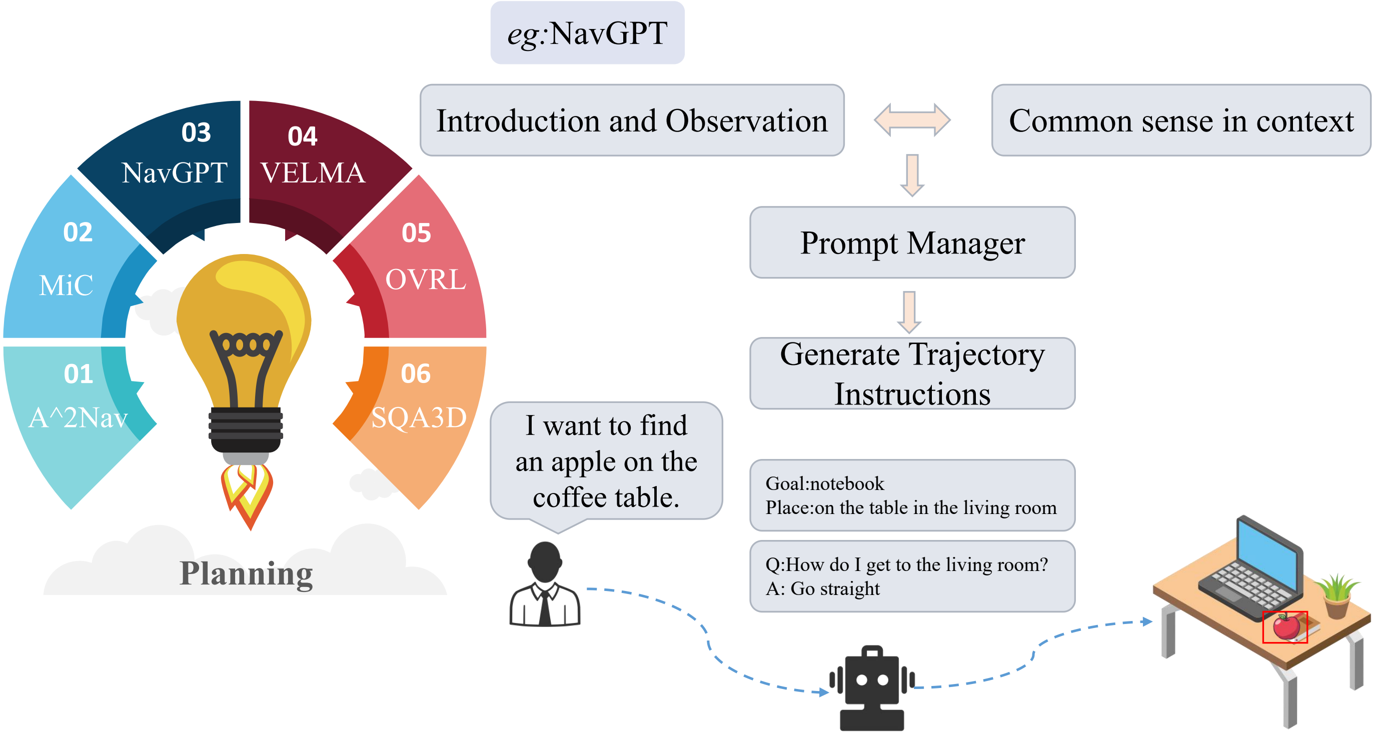
eg:NavGPT
Introduction and Observation
Common sense in context
Prompt Manager
Generate Trajectory Instructions
01 A^2Nav
02 MiC
03 NavGPT
04 VELMA
05 OVRL
06 SQA3D
Planning
I want to find an apple on the coffee table.
Goal:notebook
Place:on the table in the living room
Q:How do I get to the living room?
A: Go straight

Fig. 4: This figure is an example diagram for Planning.

in Street View settings. The agent navigates based on human-generated instructions, which include landmark references and directional cues. Schumann et al. developed VELMA to use verbal cues, contextualized by visual observations and trajectories, to determine subsequent actions. 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.

LLMs have been increasingly utilized as reasoning engines for embodied agents in diverse applications, such as domestic robotics, video games, and indoor navigation. However, the challenge of interfacing optimally with interactive visual environments persists. This study addresses the issue through verbalization embodiment. Here, the agent's task, trajectory, and visual observations are verbalized in natural language, allowing LLMs to comprehend navigation instructions and execute corresponding actions.

The paper introduces a landmark scorer to assess landmark visibility within panoramic images. This scorer calculates 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.

The landmark classification is unsupervised, given the absence of ground-truth labels for visibility. The scorer's threshold is its only tunable parameter. The agent also assesses views to its current left and right orientations. Each directional view is linked to specific textual information. Visible landmarks, along with their associated directional texts, are fed into a language expresser, which then generates verbal descriptions of the environment.

These verbal observations are of two kinds: street intersections and landmarks visible within the current view. These strings encode information about the number of outgoing edges at the current node, the names of visible landmarks, and their associated directional texts.

The study evaluated the performance of the proposed LLM agent in both fine-tuning and contextual learning scenarios. A GPT-3-based landmark extractor was used for a single run and its extracted landmarks were employed in all subsequent experiments. For the landmark scorer, the CLIP-ViT-bigG-14-laion2B-39B-b160k model was chosen as the CLIP model.

12) A^2**Nav** [12] is an action-aware vision-and-language navigation method that consists of an instruction parser and an action-aware navigation strategy. The instruction parser, which employs large language models such as GPT-3, decomposes complex navigation instructions into a sequence of object-specific navigation sub-tasks that relate to particular actions. These sub-tasks necessitate that the agent identify objects and navigate to target locations based on associated action requirements. Despite advancements in VLN, current approaches have failed to sufficiently account for diverse action requirements within instructions, such as ”continue beyond” and ”exit,” thus ignoring the agent-object relationship in the scene.

To tackle this problem, the agent must accurately identify and execute the action requirements associated with each milestone. However, the complexity of grammatical structures and the diversity of action expressions in instructions make understanding these requirements challenging. Additionally,

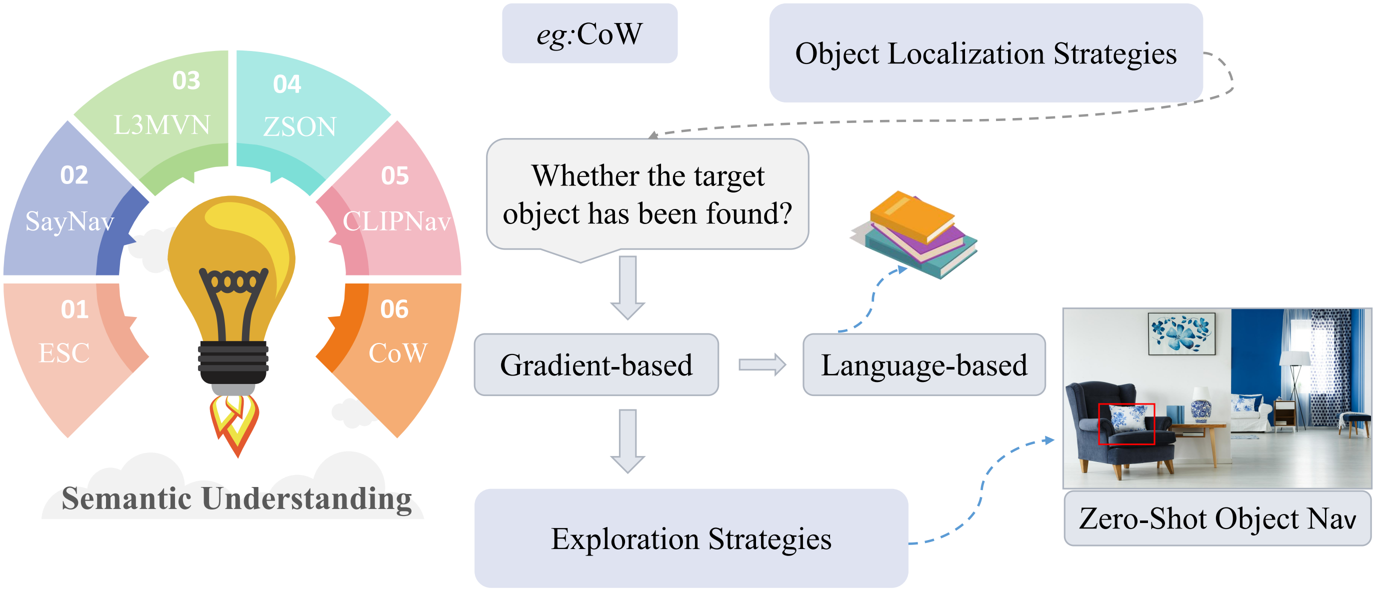

Fig. 5: This figure is an example diagram for semantic understanding.

the development of navigation strategies capable of meeting these action requirements without annotated path instructions remains an open question.

A^2Nav employs the reasoning capabilities of large language models to parse instructions into action-specific object navigation sub-tasks. In these sub-tasks, the agent identifies landmarks and navigates based on pertinent action requirements. Addressing the limitations of existing zero-shot navigation agents, Chen et al. identified five fundamental action requirements and developed distinct navigators for each. This approach, which obviates the need for manual path instruction annotations, yields more accurate and interpretable navigation outcomes. The system utilizes GPT-3 to parse all landmarks and their associated action requirements, effectively decomposing the instruction into a series of sub-tasks specific to actions.

To execute these sub-tasks sequentially, A^2Nav adopts a zero-shot learning approach, comprising five action-specific navigators. While parsing instructions, GPT-3 predicts the ”ACTION,” which typically aligns with one of the predefined ”ACTION” for a sub-task, allowing direct mapping to the corresponding sub-task type. Chen et al. encode all predicted and predefined ”ACTION” using BERT [35] and select the sub-task with the highest cosine similarity score as the predicted sub-task.

For action-aware navigation, Chen et al. fine-tuned ZSON [49] to individually train five distinct action-specific navigators. In each specific action dataset, the positions of landmarks and paths reveal features pertinent to the respective action requirement. In any given task, the current sub-task type is first identified, and the corresponding action-specific navigator is selected to predict low-level actions. The process iterates until either all sub-tasks are completed or the maximum step limit for the episode is reached.

In summary, A^2Nav effectively addresses action requirements within VLN tasks by breaking them down into action-specific object navigation sub-tasks. It achieves superior zero-shot VLN performance on the RxR and R2R datasets and outperforms state-of-the-art supervised methods on RxR. Additionally, A^2Nav excels in navigating according to instructions with specific action requirements, underscoring its potential in scenarios requiring human-agent interactions.

13)**MiC** [56] (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 modules: Generalized Object-and-Scene-Oriented Dynamic Planning (GOSP), Step-by-step Object-and-Scene-Oriented Dynamic Planning (SODP), and Room-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 perception, 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.

Concerning ROASP, it not only classifies the room type but also identifies the visible objects in the agent's immediate environment. Rather than employing separate classifiers and detectors for predicting room types and visible objects, ROASP leverages the CLIP model. With CLIP's robust zero-shot image classification capabilities, ROASP efficiently performs both tasks. Specifically, it extracts room-type labels from MatterPort3D and object-type labels from REVERIE, then utilizes CLIP for feature extraction.

In operational terms, during a task, GOSP identifies the target object by querying the LLM and utilizes the LLM's extensive world knowledge to deduce the object's probable location. Subsequently, ROASP interprets the current scene and prompts the LLM to generate detailed step-by-step plans for the next navigation action. If ROASP discerns a room change, SODP re-queries the LLM to generate a fresh set of fine-grained instructions, aligning them with prior LLM responses. The agent then continues, following these dynamically generated instructions, executing GOSP once and iterating through ROASP and SODP until task completion.

In summary, MiC provides concise high-level instructions to VLN agents, enabling REVERIE agents to dynamically interact with the LLM to formulate plans for upcoming steps. This dynamic planning methodology posits a promising avenue for future research and development.

14)**SayNav** [60] is a groundbreaking framework designed to harness the common-sense knowledge encapsulated in Large Language Models (LLMs) for generalizing intricate navigation tasks in expansive and unfamiliar environments. SayNav employs a unique grounding mechanism that incrementally constructs a 3D scene graph of the explored territory, subsequently feeding this information to LLMs. This results in the generation of high-level navigation plans that are context-aware and practically implementable, executed thereafter by a pretrained low-level planner as a series of short-distance point-goal navigation sub-tasks.

Incorporating a two-tiered planning architecture and relying on the ProcTHOR [18] dataset, SayNav's high-level planner extracts subgraphs from the comprehensive 3D scene graph, concentrating on spatial relations in the immediate vicinity of the agent. These subgraphs are converted into textual prompts and presented to the LLM, which in turn produces short-term, stepwise instructions in pseudo-code format. The plans thus generated recommend efficient search strategies, ranking various locations in the room based on their likelihood of containing target objects.

The low-level planner is tasked with generating brief motion control commands. It operates on RGBD images (320x240 resolution) and the agent's pose data, outputting basic navigational commands in alignment with standard PointNav configurations. To unify the operations of the high-level and low-level planners, SayNav treats each pseudo-code instruction from the LLM as a short-distance point-goal navigation subtask. The 3D coordinates of the objects specified in each planned step determine the target points for these sub-tasks.

SayNav distinguishes itself by its ability to dynamically generate navigation instructions and iteratively refine its future steps based on newly collected data. Performance metrics on multi-object navigation tasks demonstrate its substantial advantage over Oracle-based PointNav methods, confirming its proficiency in dynamically planning and successfully executing object-finding tasks in large, unfamiliar environments. Moreover, SayNav has proven its efficacy in generalizing from simulated to real-world conditions. In conclusion, SayNav emphasizes the utilization of human-like reasoning and advanced semantic understanding for achieving efficient and adaptive navigation, especially in complex or uncharted settings.

2) Comparison: We have compared the benchmarks mentioned above, evaluating aspects such as the large models employed, application domains, and additional features. These comparisons are summarized in Table III.At the same time, we have divided the aforementioned benchmarks into two categories: Planning, as illustrated in fig.4, and Semantic Understanding, as shown in fig.5. Corresponding examples are provided for each category.Furthermore, we have analyzed the performance of these benchmarks across various datasets, the details of which are provided in Tables I and II. It is important to clarify that comprehensive evaluation was unfeasible for some benchmarks due to the unavailability of public code. Consequently, the results are based solely on performance metrics reported in the corresponding publications. For an in-depth understanding of each benchmark, readers are directed to the cited literature.

TABLE I: Performance Comparison (Part 1)

Benchmark	Habitat		RoboTHOR		Gibson	
	SPL	SR	SPL	SR	SPL	SR
CoW	6.3%	-	10%	16.3%	-	-
ZSON	-	-	-	-	28%	36.9%
L3MVN	-	-	-	-	37.7%	76.1%
VLMAP	6%	2.5%	-	-	-	-
ESC	-	-	22.2%	38.1%	-	-

TABLE II: Performance Comparison (Part 2)

Benchmark	HM3D		MP3D		REVERIE		R2R	
	SPL	SR	SPL	SR	SPL	SR	SPL	SR
ZSON	12.6%	25.5%	4.8%	15.3%	-	-	-	
CLIP-Nav	-	-	-	-	4.56%	15.2%		
L3MVN	-	-	23.1%	50.4%	-	-	-	
OVRL	29%	64%	-	-	-	-	-	
ESC	22%	38.5%	13.7%	28.6%	-			
NavGPT	-	-	-	-	-	-	34%	29%
A^2Nav	-	-	-	-	-	-	22.6%	11%
MIC	-	-	-	-	55.74%	41.97%		
SayNav	60.32%	34%	-	-	-	-	-	

D. Other Applications of LLMs in Embodied Intelligence

1)**VoxPoser** [34] focuses not on embodied navigation benchmarks but on embodied AI. Proposed by Huang et al., VoxPoser exploits the actionable knowledge embedded in LLMs for robot manipulation tasks, particularly in reasoning and planning. While progress exists in this domain, most methods continue to rely on predefined motion primitives for interacting with the environment— a limitation VoxPoser seeks to address. The primary aim is to synthesize robotic trajectories composed of dense sequences of 6-DoF end-effector path points for a wide range of operational tasks, given an open set of instructions and object domains.

To accomplish this, VoxPoser utilizes LLMs to extract hints and constraints from free-form language instructions. Interaction with the LLM is facilitated through code composition, which culminates in the formation of 3D value maps. These maps are subsequently used in a model-based planning framework to zero-shot synthesize closed-loop robotic trajectories robust to dynamic perturbations.

Regarding system implementation challenges, Huang et al. point out the impracticality of LLMs directly outputting control actions in text form due to the high-frequency control

TABLE III: **Comparison of Benchmarks** This table provides a comparison of different benchmarks, the models used, the datasets, and their respective applications. It’s worth noting that each benchmark uses different evaluation criteria, which are not listed here. OVRL is not a zero-shot model but rather a fully supervised learning model, included here for reference. For specific details on each benchmark, please refer to the citations at the end of the article.

Benchmark	Multimodal	Design Structure	Dataset	Application
CoW	✓	CLIP	RoboTHOR	indoor scenes
ZSON	✓	CLIP	MP3D	open-world object-goal navigation
LM-Nav	✓	ViNG,CLIP,GPT-3	-	outdoor environments
CLIP-Nav	✓	CLIP	REVERIE	Household environments
L3MVN	✓	RoBERTa	Gibson,HM3D	indoor scenes
VLMAP	✓	CLIP	Matterport 3D,Habitat	navigation,generate obstacles map
OVRL	✓	ViT,LSTM	HM3D,Gibson	IMAGENAV,OBJECTNAV
ESC	✓	PSL,GLIP	HM3D,RoboTHOR	indoor scenes
NavGPT	✓	GPT-4	R2R, R2R-VLN	indoor scenes
VELMA	✓	CLIP	Touchdown,Map2seq	urban VLN
A^2Nav	✓	GPT-3	R2R-Habitat, RxR-Habitat	Zero-Shot Robot Navigation
MIC	✓	CLIP	REVERIE	remote object localization
SayNav	✓	GPT	Habitat,Gibson-4+	multi-object navigation
SQA3D	✓	CLIP,BERT,GPT-3	ScanNet	Situated Question Answering

signals required in high-dimensional spaces. However, LLMs are adept at deducing hints and constraints based on language conditions. By leveraging their code-generation capabilities, dense 3D voxel maps can be assembled by coordinating perceptual calls, like those enabled by CLIP or open-vocabulary detectors, and positioning them within the visual space.

For example, an instruction such as ”Open the topmost drawer, but be careful of the vase,” allows the LLM to deduce that: 1) the handle of the topmost drawer should be grasped, 2) the handle needs to be pulled outward, and 3) caution must be exercised to avoid the vase. Once transcribed into text, the LLM can generate Python code to invoke perception APIs for determining spatial geometry of pertinent objects or components (e.g., ”handle”). Subsequently, 3D voxels are manipulated to allocate rewards or costs to relevant positions within the observed space. The resulting value maps then serve as the objective function for motion planners, facilitating the direct synthesis of robotic trajectories that fulfill given instructions, all without the need for additional training data.

VoxPoser uniquely enables the formation of voxel value maps in a 3D observational space, guiding robotic interaction with the environment. It employs LLMs to tackle critical aspects of robotic trajectory generation, circumventing the need to train policies on typically sparse or inconsistent robotic data. This yields effective zero-shot generalization to open-set instructions. The online learning methodology contributes significantly to enhancing both the robustness and generalizability of the agent.

2)**ALFRED** [68] (Action Learning From Realistic Environments and Directives) serves as both a benchmark and dataset for the training and evaluation of models across a spectrum of embodied intelligence tasks. The central focus of this benchmark is to examine the efficacy of AI agents in carrying out domestic tasks, based on detailed natural language directives. The dataset comprises 25,743 English instructions that correspond to 8,055 expert demonstrations. Each demonstration encompasses approximately 50 steps, culminating in a total of 428,322 paired image-action sequences.

ALFRED tasks are highly complex, often necessitating that agents execute a chain of sub-tasks involving irreversible state alterations, such as breaking an object or transferring food between containers. Included in the dataset are high-level objectives like ”rinse a mug and place it in the coffee maker,” as well as more granular instructions such as ”walk to the coffee maker on the right.”

Robots interacting with human habitats must learn to associate natural language with their surroundings. ALFRED outperforms existing vision and language task datasets by offering a wealth of expert demonstrations in simulation environments that are more intricate in terms of sequence length, action space, and language complexity.

Contrary to the advancements made possible by datasets like MP3D, these resources often fail to address the complexities arising from task-oriented object interactions. ALFRED mandates that agents extract essential information from natural language directives, encompassing both the goals and potential execution steps for tasks. Agents are also required to deconstruct high-level tasks into executable actions or sub-tasks and plan an optimal path for task completion. Additionally, they must identify various objects and environmental features from a first-person perspective and engage in complex interactions with environmental objects, such as picking up items, manipulating switches, or navigating to specific locations.

By offering a visually and physically realistic simulated environment that enables translation from language to action sequences and interactions, ALFRED tackles numerous challenges evident in real-world scenarios. This includes translating human language into robotic actions for household tasks. Models capable of overcoming these challenges will begin to narrow the divide between real-world, language-driven robots and their simulated analogs.

3)**PaLM-E** [23] (Pre-training across Language and Modality Embodied) serves as a multimodal language model proficient in integrating visual, linguistic, and sensor modalities to facilitate more grounded reasoning for complex real-world challenges. Specializing in robotic planning, visual question-answering, image description, and linguistic tasks, PaLM-E exhibits robust transfer learning capabilities. It trains as a generalized, efficient embodied reasoner by merging embeddings of real-world entities in its dataset, making it skillful in multi-agent decision-making and transfer learning. Experiments indicate that concurrent training on diverse robotic tasks and datasets bolsters its performance and generalization capacity. Moreover, PaLM-E excels in language-conditioned policies and self-supervised learning, enabling both zero-shot and one-shot generalizations.

While LLMs display substantial reasoning capabilities across an array of domains, they are constrained by a grounding issue. Training LLMs on expansive textual data may yield representations pertaining to the physical world, yet integrating these with real-world visual and sensor modalities is essential for addressing broader, grounded challenges in computer vision and robotics. Previous research, such as work by Ahn et al. [1], combined the outputs of LLMs with learned robotic policies and affordance functions. However, these efforts were limited as they supplied only textual inputs to the LLM, insufficient for tasks demanding geometric scene understanding.

PaLM-E augments this by directly assimilating continuous sensor modalities within the agent, thereby enriching the language model's capacity for informed, sequential decision-making in real-world contexts. Inputs like images and state estimates are encoded in the same latent space as linguistic tokens, and are processed through the self-attention layers of a Transformer-based LLM analogously to text. Driess et al. initiated this process with a pre-trained LLM, incorporating continuous inputs via encoders. These encoders undergo end-to-end training to produce sequential decisions in natural language, interpretable by agents through fine-tuned, low-level policies or by providing specific query responses.

Utilizing the Transformer architecture and drawing from a diverse set of training data that encompasses text, images, and audio, PaLM-E can comprehend and manage multiple information types. It boasts zero-shot learning capabilities, enabling it to tackle previously unencountered tasks via inference and generation. The model accommodates a broad array of inputs like text, images, and audio, rendering it versatile for complex tasks. Its generative prowess extends to the production of natural language text, image descriptions, and beyond, thereby making it applicable across various natural language generation domains.

4)**RT-2** [6] (Robotics Transformer 2) serves as an avant-garde model crafted to adapt web-sourced visual and linguistic insights into robotic control functions. Composed of dual core elements—a pre-trained vision-language model and a reinforcement learning-based robotic control policy—RT-2 leverages expansive web data to cultivate comprehensive visual and linguistic awareness. Exhibiting remarkable proficiency in both simulated and authentic environments, the model excels in executing intricate operations such as object manipulation and door-opening tasks.

Contemporary researchers like Brohan et al. contend that LLMs function primarily at a macroscopic level, essentially acting as state machines that dictate orders to independent, microscopic controllers. These low-level controllers often lack access to the rich semantic knowledge harbored within web-scale models during their developmental stages. They probe the feasibility of incorporating pre-trained, large-scale Visual Language Models directly into these microscopic controllers to amplify their generalization and emergent semantic reasoning faculties. In response, they develop visual-language models tailored for open-vocabulary visual question-answering and dialogue, programmed to generate low-level robotic actions, while concurrently addressing broader web-scale visual-language challenges.

Through the refinement of existing models such as PaLM-E [23] and PaLI-X [13] into RT-2-PaLM-E and RT-2-PaLI-X, Brohan et al. engineer high-efficacy robotic schemes that inherit the advanced generalization and emergency response features facilitated by web-scale visual-language pre-training. This unambiguous yet multifaceted strategy elucidates the immense potential for robotic technologies to gain directly from the strides made in visual-language models, thereby aligning the progression of robotics with concurrent advancements in complementary fields.

Nonetheless, deploying large-scale Visual-Language-Action (VLA) models for real-time operations necessitates considerable computational resources, thereby raising issues for scenarios that demand high-frequency control. A promising avenue for ensuing research encompasses the exploration of quantization and distillation techniques, which could facilitate the functioning of these models at elevated speeds or on more economically viable hardware platforms.

IV. State-of-the-Art Datasets

Recently, with the accelerated advancement of AI technologies, the focus of interest has transitioned from digital interfaces to tangible environments, catalyzing the emergence of a novel research frontier—Embodied AI. As implied by its nomenclature, Embodied AI primarily investigates the dynamic interplay between an autonomous agent and its real-world environment. To execute tasks both successfully and robustly, agents require advanced proficiencies in environmental perception, encompassing, but not limited to, faculties in vision, language, reasoning, and planning. Consequently, the selection of an appropriate dataset for pre-training these agents becomes a critical determinant of their performance.

Current explorations within the domain of Embodied AI encompass a diverse array of applications including navigation, object manipulation, and multimodal learning. Correspondingly, the datasets requisite for these distinct research areas exhibit considerable variation. This section will present an overview of multiple datasets pertinent to embodied intelligence: MP3D [9], TOUCHDOWN [11], R2R [2], CVDN [70], REVERIE [55], RXR [36], SOON [90], ProcTHOR [18], R3ED [86], and X-Embodiment [16]. Each dataset's structure will be analyzed, providing readers with an opportunity to delve deeper into the characteristics and utility of these resources by consulting the references furnished at the article's conclusion. The chronological development of these datasets is depicted in Figure 6.

1)**MatterPort3D** [9] (MP3D) is a comprehensive dataset, predominantly comprising RGB-D scenes. It features 10,800

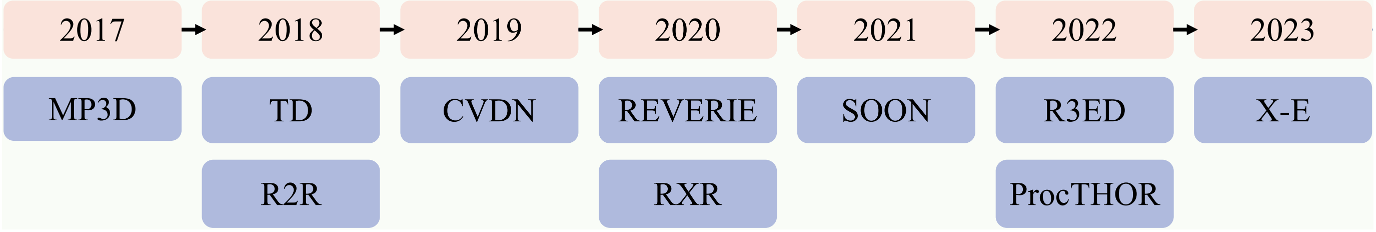

Fig. 6: The timeline below provides an overview of the key, commonly referenced datasets discussed in this paper, covering a period from 2017 to 2023.

panoramic views generated from 194,400 RGB-D images and spans 90 expansive architectural settings, primarily focusing on 3D indoor environments. The dataset encompasses various data modalities such as RGB images, depth images, and semantic annotations. Annotations are supplied for facets like scene surfaces, camera poses, and semantic segmentation. Given its voluminous collection of scenes and precise global perspectives, MP3D serves as a pivotal resource for tasks reliant on self-supervision or alternative methods.

As a cornerstone in the field, MP3D has influenced the development of numerous subsequent datasets, which have either adapted or expanded upon its structure. This renders it an invaluable asset for simulating Embodied AI in computational settings. Nevertheless, the dataset's primary focus on indoor residential environments circumscribes its applicability. Unique to MP3D, each viewpoint incorporates a 360-degree panoramic image, encompassing both depth and color information. The dataset meticulously samples viewpoints at human height throughout these spaces, ensuring camera poses are aligned. Data are amassed across a range of applications, from semantic segmentation to object and regional categorization.

For each object surface, MP3D furnishes multiple views captured from diverse angles and distances, thereby enabling the learning and prediction of surface properties contingent on these perspectives. The dataset places a significant emphasis on visual perception and scene understanding, making it highly versatile for a plethora of computer vision tasks including object detection, scene segmentation, and 3D reconstruction. For Embodied AI research, MP3D stands as an exemplary dataset, offering a data-rich environment that enhances agent-environment interactions.

2)**TOUCHDOWN** [11] is a specialized dataset designed for navigation and spatial reasoning, utilizing Google Street View to construct an expansive outdoor environment. The dataset encompasses 39,641 panoramic images and 61,319 edges, all originating from New York City. Each viewpoint in the dataset features a 360-degree RGB panorama. To compile this dataset, TOUCHDOWN employs a ”treasure hunt” methodology, involving two roles: the writer and the finder. The writer crafts instructions to enable the finder to navigate toward and locate a specific target, without prior knowledge of the correct route or destination. Upon task completion, the finder documents the followed path and elucidates their navigation choices.

TOUCHDOWN consists of two primary tasks: Navigation and Spatial Description Resolution (SDR). The Navigation task emphasizes spatial reasoning, with instructions focusing on the spatial relationships between the agent and its environment. The SDR task, on the other hand, demands linguistic expertise to decipher relationships among objects and their spatial configurations. Each of these tasks presents its own set of unique challenges.

Contrary to MP3D [9], which is confined to indoor scenarios, TOUCHDOWN predominantly features outdoor settings. This makes it an ideal dataset for training and evaluating models in tasks that require outdoor navigation and spatial understanding.

3)**Room-to-Room** [2](R2R) is a specialized dataset designed for Vision-and-Language Navigation (VLN), extending the data from MP3D to include a substantial corpus of natural language data pertinent to navigation. R2R characterizes a series of navigational paths within these 3D indoor settings, each consisting of a sequence of specific locations and orientations, termed ’viewpoints.’ Accompanying each path are one or more natural language instructions that guide navigation from the origin to the destination.

Unlike MP3D, which is more general-purpose, R2R employs these indoor settings for highly specialized navigation tasks, necessitating a nuanced understanding of natural language and complex decision-making. Navigating intricate indoor terrains often demands capabilities in long-term memory and reasoning. A model utilizing R2R not only needs to comprehend natural language directives but must also correlate them with the visual landscape. Upon this synthesis, the model is required to strategize an efficient route from the starting location to the specified endpoint. Through these multiple dimensions, R2R offers an intricate and demanding research platform for those specializing in vision-and-language navigation.

4)**CVDN** [70] is an extensive dataset comprising over 2,000 tailored ”human-to-human” dialogues, set in a simulated, photorealistic domestic environment. Developed as an extension of the R2R framework and formulated in English, CVDN aims to cultivate navigation agents capable of functioning in residential and commercial spaces via linguistic directives. Notably, these agents have the ability to pose focused questions during instances of uncertainty regarding subsequent navigational steps.

Furthermore, CVDN can be leveraged to develop agents endowed with specialized environmental knowledge, thereby facilitating automated linguistic guidance for individuals in unfamiliar contexts—such as direction-seeking within an office complex. The high fidelity of the simulated environment ensures that agents trained on CVDN can generalize skills to real-world settings.

Diverging from R2R, dialogues within CVDN typically contain nearly triple the word count and describe paths averaging over thrice the length. Thus, CVDN provides a significant avenue for the exploration of in-situ navigational interactions and holds potential for the training of agents proficient in both navigation and dynamic inquiry within human-centric environments.

5)**REVERIE** [55] is a comprehensive dataset featuring 10,567 panoramic images spread across 90 buildings and encompassing 4,140 target objects. Accompanying these are 21,702 crowd-sourced instructions, averaging 18 words each. While it shares visual data attributes like RGB images with earlier datasets, REVERIE also incorporates enhanced semantic particulars related to the Goal Object, such as color, along with natural language directives.

Distinct from conventional VLN datasets, REVERIE includes a myriad of complex and diverse indoor environments. Moreover, its instruction set diverges from the step-by-step commands typical of R2R datasets to offer semantic-level directives more congruent with real-world tasks and human-agent interactions. For example, instead of sequential directives like 'Start at the entrance and proceed to the red sofa,' REVERIE provides semantically rich tasks such as 'Fold the towel in the bathroom with the fishing theme.'

This semantic depth, however, introduces a layer of complexity due to potential ambiguities (e.g., 'Fold the towel in the bathroom with the fishing theme' could signify either 'Fold the fishing-themed towel' or 'in the fishing-themed bathroom'). Consequently, agents trained on the REVERIE dataset require enhanced long-term memory, semantic comprehension, and environmental reasoning abilities. Such attributes make REVERIE particularly suited for remote entity referencing and multi-task learning endeavors. In summary, REVERIE offers a multifaceted, challenging platform conducive to advancements in VLN and multimodal learning.

6)**RxR** [36](Room-Across-Room) is an expansive dataset tailored for Multilingual Vision-and-Language Navigation (VLN) tasks. As an extension of the Room-to-Room (R2R) dataset, it inherits substantial visual and navigational data, thereby enriching the research landscape with multilingual capabilities. The dataset features natural language navigation instructions in multiple languages, including but not limited to, English, German, and Spanish. This linguistic variety significantly broadens the scope for multilingual VLN endeavors.

Furthermore, RxR introduces dense spatiotemporal grounding, which entails comprehensive action execution and natural language interpretation directives, calibrated to specific temporal intervals and spatial coordinates within the navigation trajectory. Similar to REVERIE, RxR augments the dataset with additional layers of semantic data pertaining to target objects and multiple languages.

This compounded complexity poses new sets of challenges: the navigation agents must master multilingual understanding, spatiotemporal reasoning, and long-term planning. Owing to its linguistic diversity, RxR serves as a multifaceted, demanding research substrate, suitable for advancing the state-of-the-art in multilingual VLN.

7)**SOON** [90], also known as the FAO dataset, is a language-guided navigation dataset grounded in 3D housing environments featuring authentic image panoramas. The dataset encompasses 90 distinct residential settings, each paired with a natural language instruction that delineates a target location. Unique to FAO, the construction of these instructions is organized around five subtasks: delineation of target attributes, explication of relationships between the target and neighboring objects, exploration of the target's regional setting, scrutiny and description of adjacent areas, and the subsequent condensation of these elements into a three-sentence narrative.

In addition to offering an extensive array of 3,848 instructions characterized by a diverse vocabulary of 1,649 words, the FAO dataset furnishes high-fidelity annotations, which include 3D environmental models, RGB-D imagery, semantic segmentation, and depth data. Remarkably, the average instruction length in FAO stands at 38.6 words, surpassing the 26.3-word and 18.3-word averages observed in REVERIE and R2R, respectively. Most instructions oscillate between 20 and 60 words, affording robust descriptive potential.

In contrast with other datasets, FAO stands out for its elevated linguistic intricacy and ecological diversity, coupled with a rich, realistic representation of environmental contexts. Consequently, the FAO dataset provides an exhaustive and reliable substrate for advancing research in the field of vision and language-guided navigation.

8)**ProcTHOR** [18], also known as PROCTHOR-10K, is a dataset comprising 10,000 algorithmically generated 3D residential scenes specifically tailored for the training and assessment of embodied AI agents. The dataset is a byproduct of the PROCTHOR framework, which is designed to facilitate the large-scale generation of diverse, interactive, and customizable virtual environments. Each scene within the dataset is distinct, furnished with a multiplicity of objects, textural variations, and illumination conditions, thereby constituting a rich and complex milieu for embodied agents to operate, interact, and manipulate.

Intended to support an array of embodied AI tasks—ranging from navigation and object manipulation to comprehensive scene understanding—PROCTHOR-10K has been employed for the training and evaluation of cutting-edge embodied AI models. The dataset has demonstrated robust performance across various industry benchmarks, including but not limited to Habitat 2022, AI2-THOR Rearrangement 2022, and RoboTHOR challenges.

9)**R3ED** [86] is a dataset encompassing over 5,800 point clouds and approximately 22,400 annotated 3D bounding boxes, sourced from seven distinct real-world indoor settings via a dense sampling technique. This dataset aspires to address the limitations inherent in previous embodied AI datasets, which predominantly rely on synthetic data to mimic indoor landscapes and robotic motions. While synthetic data offers convenience, it tends to undermine real-world performance.

Conversely, R3ED offers authentic point cloud data following each agent's move, furnishing researchers with the opportunity to examine active vision through a more realistic and standardized lens. The point cloud data is meticulously annotated with 3D bounding boxes that categorize 12 types of objects, inclusive of prevalent furniture and appliances. Furthermore, the geometrical relationships among these point clouds are explicitly annotated using a specialized storage path format. To augment the dataset's utility, the authors have introduced a 3D divergency policy aimed at directing robotic movements for data collection, thereby enhancing visual performance in unfamiliar environments.

10)**X-Embodiment Dataset** [16] comprises over one million robot trajectories from 22 disparate robot instances, illustrating 527 skills across 160,266 tasks. The dataset amalgamates data from 60 pre-existing robot datasets, sourced from 34 international robotics research labs, and standardizes them into a unified RLDS data format [61]. This format efficiently serializes record files and accommodates a myriad of action spaces and input modalities, from varying numbers of RGB cameras to depth cameras and point clouds, thereby facilitating streamlined data loading across all primary deep learning frameworks. Transformer-based policies trained on this dataset have shown marked positive transfer capabilities across different robotic instances.

Recent developments in machine learning and artificial intelligence indicate that large-scale, diversified datasets are pivotal for the creation of competent AI systems. Such is evident in domains like natural language processing and computer vision, where large, high-capacity models trained on diverse data have effectively addressed a range of downstream applications. This trend has catalyzed the widespread adoption of pretrained models as foundational assets across applications. In the realm of robotics, models have traditionally been specialized for specific applications, robots, and environments. However, there is burgeoning interest in the development of "generalist" robotic policies capable of efficiently adapting to novel robots, tasks, and settings. Large-scale, multipurpose models, often trained on diverse datasets, consistently outperform task-specific counterparts trained on more limited datasets.

Prompted by the efficacy of pretraining expansive vision or language models on varied data, DeepMind posits that the construction of universally applicable robotic policies necessitates X-embodiment training. This approach utilizes data aggregated from multiple robotic platforms. Although individual robotic learning datasets might be too specialized, their combination provides a more comprehensive representation of environmental and robotic variations. The development of methodologies to optimally leverage this X-embodiment data from diverse labs, robots, and settings is indispensable for the creation of universally applicable robotic policies. While current datasets might not yet match the generalization capabilities exhibited by large language models, the collaborative utilization of such resources has the prospective capacity to achieve comparable coverage. Consequently, DeepMind contends that research into X-embodiment robotic learning is of paramount importance at this stage.

This section provides an overview of nine seminal datasets, delineating their structures and merits. Table IV offers a comparative analysis among these datasets, with more exhaustive information available in the References. Broadly, both contemporary and forthcoming datasets are gravitating towards the inclusion of real-world 3D data, multi-modal inputs, higher-order instructions, and provisions for human oversight and direction.

V. Challenges and Limitations

Examining the challenges and limitations of Embodied Intelligence in the realm of navigation provides a comprehensive perspective into the intricacies and the prevailing research gaps within this domain. The following conclusion aims to offer researchers an extensive viewpoint on the hurdles and existing shortcomings in employing EI for navigation purposes.

1) **Ambiguity in Definition:**There's a lingering debate surrounding the definition of Embodied Intelligence, particularly when it comes to whether a physical body is essential for an intelligence to be deemed "embodied" [63]. This ambiguity in definition might sway research directions, experimental design, and the evaluation and comparison of embodied navigation systems.
2) **Consideration of Safety:**The actions of embodied agents could have long-lasting impacts on the environment; thus, they need to learn and interact safely to prevent potential catastrophic events. This necessitates embodied navigation systems to possess adequate self-monitoring and risk assessment capabilities.
3) **Diversity in Visual Navigation:**Visual navigation encompasses a variety of tasks like visual language navigation, embodied question answering, scenario navigation, etc [84]. This diversity demands embodied navigation systems to have sufficient flexibility and adaptability.
4) **Challenges in Practical Applications:**For instance, there are many exciting challenges in practical applications and experiments in visual navigation, vision-and-language, and audio-visual navigation.

The inherent challenges and limitations associated with LLMs in Embodied Intelligence highlight a substantial research trajectory within the field of navigation. The primary challenge stems from the disconnect between LMs and the physical world, as these models predominantly rely on text-based datasets, which constrains their effectiveness in tasks requiring embodied intelligence. Another significant hurdle is the extensive volume of training data required; acquiring and annotating large datasets can be both time-intensive and expensive, limiting the scalability and practical application of LMs. To accelerate progress in this domain, researchers must explore the aforementioned issues in depth and pursue innovative solutions and methodologies. Simultaneously, undertaking more experiments and promoting practical applications are vital for advancing research in this field.

VI. Future Directions

Concerning the future of LLMs for embodied intelligence, multiple promising avenues of research are emerging. One pivotal focus is the development of more robust and adaptable language models proficient in understanding and generating

TABLE IV: **Comparison of Datasets**. This table is a comparison among nine datasets, including seven evaluation methods.

Dataset	Size			Content		Real(Physical)
	Panoramic	Instruction	Point Cloud	Human	Inclusion	
MP3D	10,800	-	-	✓	View,Annotation	✓
TOUCHDOWN	39,641	-	-	✓	Images,Edges	✓
R2R	-	21,567	-	✓	Nav-Instruction	✓
REVERIE	10,567	21,702	-	✓	View,Instruction	✓
RxR	-	126,000	-	✓	Nav-Instruction	✓
FAO	-	3,848	-	✓	Instruction,Location	✓
ProcTHOR	10,000	-	-	✗	View	✗
R3ED	-	-	5800	✗	PointCloud,Bbox	✓
CVDN	-	-	-	✓	Dialog	Real

natural language across various contexts. This entails integrating LLMs with computer vision and robotics to construct agents capable of effective environmental perception and interaction. Another avenue involves the investigation of novel training techniques and architectures for LLMs, with a focus on enhancing the efficiency and scalability of training processes through the use of unsupervised and transfer learning methods. Moreover, an increasing interest in developing LLMs capable of learning from multimodal data, such as text, images, and audio, suggests potential for more holistic language and environmental representations.

Beyond the use of LLMs, Reinforcement Learning (RL) offers significant advantages for Embodied Navigation [32] [45]. RL allows agents to learn decision-making by interacting with complex, dynamic environments to achieve goals, demonstrating exceptional adaptability [26], [52], [77]. For example, Liu et al. [42] constructed a robotic state successor representation model using RL, combined with a goal-based reward function, which allows for the re-planning of the shortest path to adapt to environmental changes.The online learning capabilities of RL enable real-time adjustment of navigation strategies based on environmental feedback. Furthermore, its goal-oriented nature makes RL particularly suitable for tasks requiring target achievement or specific objectives.

However, RL also has notable drawbacks for Embodied Navigation. Its sample inefficiency requires extensive data to learn effective policies, posing substantial challenges in real-world settings. Training RL agents is time-consuming due to the trial-and-error learning process. Stability and generalization of policies remain critical issues, with RL models often struggling to adapt to new, unseen environments.Researchers are actively addressing these issues [85] [43]. For instance, Zhu et al. [91] have combined RL with rule-based methods to form a Rule-based RL (RuRL) algorithm, reducing sample complexity and time costs.

In Embodied Navigation applications, an ideal approach may involve leveraging the strengths of both methods, such as using LLMs for processing complex instructions and RL for interactive and adaptive environmental learning. This multimodal and multi-technique fusion could enhance overall system performance, enabling machines to understand and execute navigation tasks more effectively.

Additionally, the need for standardized benchmarks and evaluative metrics in embodied intelligence tasks is apparent. Such benchmarks will enable comparative assessments of various models and algorithms, thereby expediting the evolution of more effective and efficient agents. Lastly, the ethical and societal implications of deploying LLMs for embodied intelligence are receiving growing attention. Interdisciplinary partnerships among researchers in AI, robotics, and social sciences are essential to guarantee the responsible and ethical evolution and implementation of LLMs.

With respect to the advancement of general-purpose embodied intelligence, several challenges loom large. Chief among these is the diversity and complexity of real-world settings. Embodied agents need the capacity for perceiving and interacting with a wide array of objects and scenarios, as well as for real-time adaptation to dynamic conditions. This mandates the incorporation of multiple sensory modalities, including vision, auditory perception, and tactile sensing, to generate a comprehensive environmental representation. The pursuit of versatile physical intelligence remains a challenging yet exhilarating research domain. By addressing existing limitations of physical agents and exploring the adaptability of task-specific agents, we can foster the development of more competent and intelligent agents. This advancement not only deepens our comprehension of the world but also elevates our quality of life.

In summary, the prospects for LLMs in the realm of embodied intelligence are promising, providing myriad opportunities for innovation and progress. By confronting existing shortcomings and venturing into new areas of research, we can develop more intelligent and capable agents, thereby enriching our understanding of both language and the world.

VII. Conclusion

This paper delves into the rapidly evolving field of Embodied Intelligence, focusing on the premise that intelligence emanates from the interaction between an agent and its environment rather than being a purely internal, abstract construct. The paper scrutinizes 14 benchmarks in embodied artificial intelligence, delineating their variances and commonalities. Furthermore, it offers an in-depth analysis of how these benchmarks catalyze contemporary advancements in research tasks in embodied AI, shedding light on their underlying design principles and methodologies. Experimental outcomes indicate significant success rate improvements across various benchmark evaluations. For instance, L3MVN recorded a success rate of 56.5% on the R2R dataset, outperforming existing state-of-the-art methods. Similarly, OVRL-V2 posted success rates of 82.0% and 64.0% in the IMAGENAV and OBJECTNAV tasks respectively, signifying marked performance enhancements over preceding methods. Additional methods like ESC, NavGPT, and VELMA also exhibited superior navigation planning and scene comprehension abilities. These experimental findings underscore the pivotal role of LLMs in zero-shot navigation tasks.

The paper also outlines challenges and limitations inherent to LLMs in the realm of embodied intelligence. These include the absence of a direct linkage between LLMs and the physical world, the necessity for extensive training data, and the complexity of understanding and generating natural language in a diverse array of contexts. Notwithstanding these impediments, several promising research trajectories exist that could propel advancements in embodied intelligence. These encompass the formulation of more robust and adaptable language models, the investigation of innovative training methodologies and architectures, the establishment of standardized benchmarks and evaluation metrics, and the critical need for cross-disciplinary collaboration among researchers in AI, robotics, and social sciences.

REFERENCES

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022.

[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.

[3] Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18:1–67, 2017.

[4] Richard Beranek, Masoud Karimi, and Mojtaba Ahmadi. A behavior-based reinforcement learning approach to control walking bipedal robots under unknown disturbances. *IEEE/ASME Transactions on Mechatronics*, 27(5):2710–2720, 2022.

[5] Ravali Boorugu and G. Ramesh. A survey on nlp based text summarization for summarizing product reviews. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 352–356, 2020.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

[7] 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.

[8] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. *arXiv preprint arXiv:2302.02662*, 2023.

[9] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017.

[10] Haoyao Chen, Dong Sun, Jie Yang, and Jian Chen. Localization for multirobot formations in indoor environment. *IEEE/ASME Transactions on Mechatronics*, 15(4):561–574, 2010.

[11] 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.

[12] Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. a^2nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models. *arXiv preprint arXiv:2308.07997*, 2023.

[13] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023.

[14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

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

[16] Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. Open X-Embodiment: Robotic learning datasets and RT-X models. https://robotics-transformer-x.github.io, 2023.

[17] Gautier Dagan, Frank Keller, and Alex Lascarides. Dynamic planning with a llm, 2023.

[18] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedu-

ral generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

[19] 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.

[20] Alessandro Devo, Giacomo Mezzetti, Gabriele Costante, Mario L Fravolini, and Paolo Valigi. Towards generalization in target-driven visual navigation by using deep reinforcement learning. *IEEE Transactions on Robotics*, 36(5):1546–1561, 2020.

[21] 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.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

[24] Vikranth Dwaracherla, Shantanu Thakar, Leena Vachhani, Abhishek Gupta, Aayush Yadav, and Sahil Modi. Motion planning for point-to-point navigation of spherical robot using position feedback. *IEEE/ASME Transactions on Mechatronics*, 24(5):2416–2426, 2019.

[25] Cem Eteke, Doğancan Kebüde, and Barış Akgün. Reward learning from very few demonstrations. *IEEE Transactions on Robotics*, 37(3):893–904, 2020.

[26] Anthony Francis, Aleksandra Faust, Hao-Tien Lewis Chiang, Jasmine Hsu, J Chase Kew, Marek Fiser, and Tsang-Wei Edward Lee. Long-range indoor navigation with prm-rl. *IEEE Transactions on Robotics*, 36(4):1115–1134, 2020.

[27] 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.

[28] Yuan Gao, Junfeng Chen, Xi Chen, Chongyang Wang, Junjie Hu, Fuqin Deng, and Tin Lun Lam. Asymmetric self-play-enabled intelligent heterogeneous multirobot catching system using deep multiagent reinforcement learning. *IEEE Transactions on Robotics*, 2023.

[29] Yaakov HaCohen-Kerner, Daniel Miller, and Yair Yigal. The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, 15(5):e0232525, 2020.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Jörg Hoffmann. Ff: The fast-forward planning system. *AI magazine*, 22(3):57–57, 2001.

[32] Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. n^2m^2: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments. *IEEE Transactions on Robotics*, 2023.

[33] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023.

[34] 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. In *7th Annual Conference on Robot Learning*, 2023.

[35] 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*, pages 4171–4186, 2019.

[36] 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.

[37] Guoqiang Li, Jun Wu, Chao Deng, Xuebing Xu, and Xinyu Shao. Deep reinforcement learning-based online domain adaptation method for fault diagnosis of rotating machinery. *IEEE/ASME Transactions on Mechatronics*, 27(5):2796–2805, 2022.

[38] Jiangang Li, Changgui Qi, Yanan Li, and Zenghao Wu. Prediction and compensation of contour error of cnc systems based on lstm neural-network. *IEEE/ASME Transactions on Mechatronics*, 27(1):572–581, 2022.

[39] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[40] Xinwu Liang, Hesheng Wang, Yun-Hui Liu, Zhe Liu, and Weidong Chen. Leader-following formation control of nonholonomic mobile robots with velocity observers. *IEEE/ASME Transactions on Mechatronics*, 25(4):1747–1755, 2020.

[41] Nir Lipovetzky, Miquel Ramirez, Christian Muise, and Hector Geffner. Width and inference based planners: Siw, bfs (f), and probe. *Proceedings of the 8th International Planning Competition (IPC-2014)*, page 43, 2014.

[42] Dong Liu, Zhi Lyu, Qiang Zou, Xue Bian, Ming Cong, and Yu Du. Robotic navigation based on experiences and predictive map inspired by spatial cognition. *IEEE/ASME Transactions on Mechatronics*, 27(6):4316–4326, 2022.

[43] Weiwei Liu, Shanqi Liu, Junjie Cao, Qi Wang, Xiaolei Lang, and Yong Liu. Learning communication for cooperation in dynamic agent-number environment. *IEEE/ASME Transactions on Mechatronics*, 26(4):1846–1857, 2021.

[44] Xiaorui Liu, Wanyue Jiang, Hang Su, Wen Qi, and Shuzhi Sam Ge. A control strategy of robot eye-head coordinated gaze behavior achieved for minimized neural transmission noise. *IEEE/ASME Transactions on Mechatronics*, 28(2):956–966, 2023.

[45] Maria Lombardi, Davide Liuzza, and Mario di Bernardo. Using learning to control artificial avatars in human motor coordination tasks. *IEEE Transactions on Robotics*, 37(6):2067–2082, 2021.

[46] Ruichen Ma, Yu Wang, Chong Tang, Shuo Wang, and Rui Wang. Position and attitude tracking control of a biomimetic underwater vehicle via deep reinforcement learning. *IEEE/ASME Transactions on Mechatronics*, 28(5):2810–2819, 2023.

[47] 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.

[48] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*, 2023.

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

[50] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[51] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[52] Rhys Newbury, Morris Gu, Lachlan Chumbley, Arsalan Mousavian, Clemens Eppner, Jürgen Leitner, Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, et al. Deep learning approaches to grasp synthesis: A review. *IEEE Transactions on Robotics*, 2023.

[53] 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.

[54] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[55] 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.

[56] Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. March in chat: Interactive prompting for remote embodied referring expression. *arXiv preprint arXiv:2308.10141*, 2023.

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[58] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[59] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[60] 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. *arXiv preprint arXiv:2309.04077*, 2023.

[61] Sabela Ramos, Sertan Girgin, Léonard Hussenot, Damien Vincent, Hanna Yakubovich, Daniel Toyama, Anita Gergely, Piotr Stanczyk, Raphael Marinier, Jeremiah Harmsen, et al. Rlds: an ecosystem to generate, share and use datasets in reinforcement learning. *arXiv preprint arXiv:2111.02767*, 2021.

[62] Zhongqiang Ren, Sivakumar Rathinam, and Howie Choset. Cbss: A new approach for multiagent combinatorial path finding. *IEEE Transactions on Robotics*, 2023.

[63] Nicholas Roy, Ingmar Posner, Tim Barfoot, Philippe Beaudoin, Yoshua Bengio, Jeannette Bohg, Oliver Brock, Isabelle Depatie, Dieter Fox, Dan Koditschek, Tomas Lozano-Perez, Vikash Mansinghka, Christopher Pal, Blake Richards, Dorsa Sadigh, Stefan Schaal, Gaurav Sukhatme, Denis Therien, Marc Toussaint, and Michiel Van de Panne. From machine learning to robotics: Challenges and opportunities for embodied intelligence, 2021.

[64] 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. *arXiv preprint arXiv:2307.06082*, 2023.

[65] Esmaeil Seraj, Letian Chen, and Matthew C Gombolay. A hierarchical coordination framework for joint perception-action tasks in composite robot teams. *IEEE Transactions on Robotics*, 38(1):139–158, 2021.

[66] Dhruv Shah. Robotic navigation with large pre-trained models of language, vision, and action. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[67] Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Ving: Learning open-world navigation with visual goals. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13215–13222. IEEE, 2021.

[68] 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.

[69] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models, 2023.

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

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[72] Murilo Mendonça Venâncio, Rogério Sales Gonçalves, and Reinaldo Augusto da Costa Bianchi. Terrain identification for humanoid robots applying convolutional neural networks. *IEEE/ASME Transactions on Mechatronics*, 26(3):1433–1444, 2021.

[73] 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.

[74] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations*, 2019.

[75] Jin Wu, Miaomiao Wang, Yi Jiang, Bowen Yi, Rui Fan, and Ming Liu. Simultaneous hand–eye/robot–world/camera–imu calibration. *IEEE/ASME Transactions on Mechatronics*, 27(4):2278–2289, 2022.

[76] Sheng Xiang, Jianghong Zhou, Jun Luo, Fuqiang Liu, and Yi Qin. Cocktail lstm and its application into machine remaining useful life prediction. *IEEE/ASME Transactions on Mechatronics*, 28(5):2425–2436, 2023.

[77] Zhanteng Xie and Philip Dames. Drl-vo: Learning to navigate through crowded dynamic scenes using velocity obstacles. *IEEE Transactions on Robotics*, 2023.

[78] Xuanhui Xu, Mingyu You, Hongjun Zhou, Zhifeng Qian, and Bin He. Robot imitation learning from image-only observation without real-world interaction. *IEEE/ASME Transactions on Mechatronics*, 2022.

[79] Karmesh Yadav, Arjun Majumdar, Ram Ramrakhya, Naoki Yokoyama, Alexei Baevski, Zsolt Kira, Oleksandr Maksymets, and Dhruv Batra. Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav. *arXiv preprint arXiv:2303.07798*, 2023.

[80] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pages 146–151. IEEE, 1997.

[81] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. *arXiv preprint arXiv:2309.12311*, 2023.

[82] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. *arXiv preprint arXiv:2304.05501*, 2023.

[83] Sheng Yu, Di-Hua Zhai, and Yuanqing Xia. Robotic grasp detection based on category-level object pose estimation with self-supervised learning. *IEEE/ASME Transactions on Mechatronics*, 2023.

[84] 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.

[85] Wei Zhang, Yunfeng Zhang, Ning Liu, Kai Ren, and Pengfei Wang. Ipaprec: A promising tool for learning high-performance mapless navigation skills with deep reinforcement learning. *IEEE/ASME Transactions on Mechatronics*, 27(6):5451–5461, 2022.

[86] Qianfan Zhao, Lu Zhang, Lingxi Wu, Hong Qiao, and Zhiyong Liu. A real 3d embodied dataset for robotic active visual learning. *IEEE Robotics and Automation Letters*, 7(3):6646–6652, 2022.

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

[88] 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. *arXiv preprint arXiv:2301.13166*, 2023.

[89] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[90] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12684–12694, 2021.

[91] Yuanyang Zhu, Zhi Wang, Chunlin Chen, and Daoyi Dong. Rule-based reinforcement learning for efficient robot navigation with space reduction. *IEEE/ASME Transactions on Mechatronics*, 27(2):846–857, 2022.