Large Language Models for Robotics: Opportunities, Challenges, and Perspectives

Jiaqi Wang*, Zihao Wu*, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Huaqin Zhao, Zhengliang Liu, Haixing Dai, Lin Zhao, Bao Ge, Xiang Li, Tianming Liu†, and Shu Zhang†

***Abstract*—Large language models (LLMs) have undergone significant expansion and have been increasingly integrated across various domains. Notably, in the realm of robot task planning, LLMs harness their advanced reasoning and language comprehension capabilities to formulate precise and efficient action plans based on natural language instructions. However, for embodied tasks, where robots interact with complex environments, text-only LLMs often face challenges due to a lack of compatibility with robotic visual perception. This study provides a comprehensive overview of the emerging integration of LLMs and multimodal LLMs into various robotic tasks. Additionally, we propose a framework that utilizes multimodal GPT-4V to enhance embodied task planning through the combination of natural language instructions and robot visual perceptions. Our results, based on diverse datasets, indicate that GPT-4V effectively enhances robot performance in embodied tasks. This extensive survey and evaluation of LLMs and multimodal LLMs across a variety of robotic tasks enriches the understanding of LLM-centric embodied intelligence and provides forward-looking insights toward bridging the gap in Human-Robot-Environment interaction.**

***Index Terms*—Large language model, robotic, GPT-4V, artifical general intelligence.**

I. Introduction

As pre-trained models have expanded in both model size and data volume, some large pre-trained models have demonstrated remarkable capabilities across a spectrum of complex tasks [1], [2]. Large language models (LLMs) have garnered widespread attention in various domains due to their exceptional contextual emergence abilities [2]–[10]. This emergent capability empowers artificial intelligence algorithms in unprecedented ways, reshaping the manner in which people utilize artificial intelligence algorithms and prompting a reevaluation of the possibilities of Artificial General Intelligence (AGI).

With the rapid development of LLMs, the utilization of instruction tuning and alignment tuning has become the primary approach to adapt them for specific objectives. In the field of natural language processing (NLP), LLMs can, to some extent, function as a versatile solution for language-related tasks [3], [5], [11]–[13].These transformer-based large models have demonstrated extraordinary achievements [14]–[17] in multiple domains, profoundly transforming the state of the art of artificial intelligence [3], [12], [18]–[26]. Research paradigms have also shifted towards employing to address subdomain-specific issues. In the realm of computer vision (CV), researchers are also working on developing large models, akin to GPT-4 and Gemini [27], [28], that incorporate both visual and language information, thus supporting multimodal inputs [29]. This strategy of enhancing LLMs not only boosts their performance in downstream tasks but also holds significant guidance for the development of robotics by ensuring alignment with human values and preferences. This method has been extensively adopted in numerous sectors [7], [29]–[32], even in areas where convolutional neural networks (CNNs) have been the primary technology [33]–[40].

The ability of LLMs to process and internalize vast amounts of textual data offers unprecedented potential for enhancing a machine's understanding and natural language analysis capabilities [41], [42]. This extends to comprehending documents like manuals and technical guides and applying this knowledge to engage in coherent, accurate, and human-aligned dialogues [43]–[45]. Through conversation, natural language instructions are translated from text prompts into machine-understandable code that triggers corresponding actions, thereby rendering robots more adaptive and flexible in accommodating a wide array of user commands [46]–[48]. Integrating real-world sensor modalities into language models facilitates establishing connections between words and perceptions, enabling their application across various specific tasks. Nevertheless, text only LLMs lack experiential exposure to the physical world and the empirical outcomes of observation, making it challenging to employ them in decision-making within specific environments. Therefore, incorporating multimodality into LLMs is crucial for the effective execution of robotic tasks. Additionally, the field of robotics presents subtler variations in tasks. Unlike NLP and CV, which can

*Co-first authors.

†Co-corresponding authors: Tianming Liu, Shu Zhang

Jiaqi Wang, Enze Shi, Huawen Hu, Xuhui Wang, Yincheng Yao, and Shu Zhang are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China. Chong Ma is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China. (e-mail: {jiaqi.wang, ezshi, huawenhu, xuhuiwang, yaoyincheng }@mail.nwpu.edu.cn; shu.zhang@nwpu.edu.cn; mc-npu@mail.nwpu.edu.cn).

Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Huaqin Zhao, Zhengliang Liu, Haixing Dai, Lin Zhao, and Tianming Liu are with the School of Computing, The University of Georgia, Athens 30602, USA. Hanqi Jiang is with the College of Engineering, University of Georgia, Athens 30602, USA.(e-mail: {zihao.wu1, yl80817, hj67104, peng.shu, hz33227, zl18864, haixing.dai, lin.zhao, tliu}@uga.edu.)

Yiheng Liu and Bao Ge are with the School of Physics and Information Technology, Shaanxi Normal University, Xi'an 710119 China. Xuan Liu is with the School of Computer Science, Shaanxi Normal University, Xi'an 710119, China. (e-mail: {liuyiheng, bob_ge, xuanliu}@snnu.edu.cn)

Xiang Li is with the Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston 02115, USA. (e-mail: XLI60@mgh.harvard.edu).

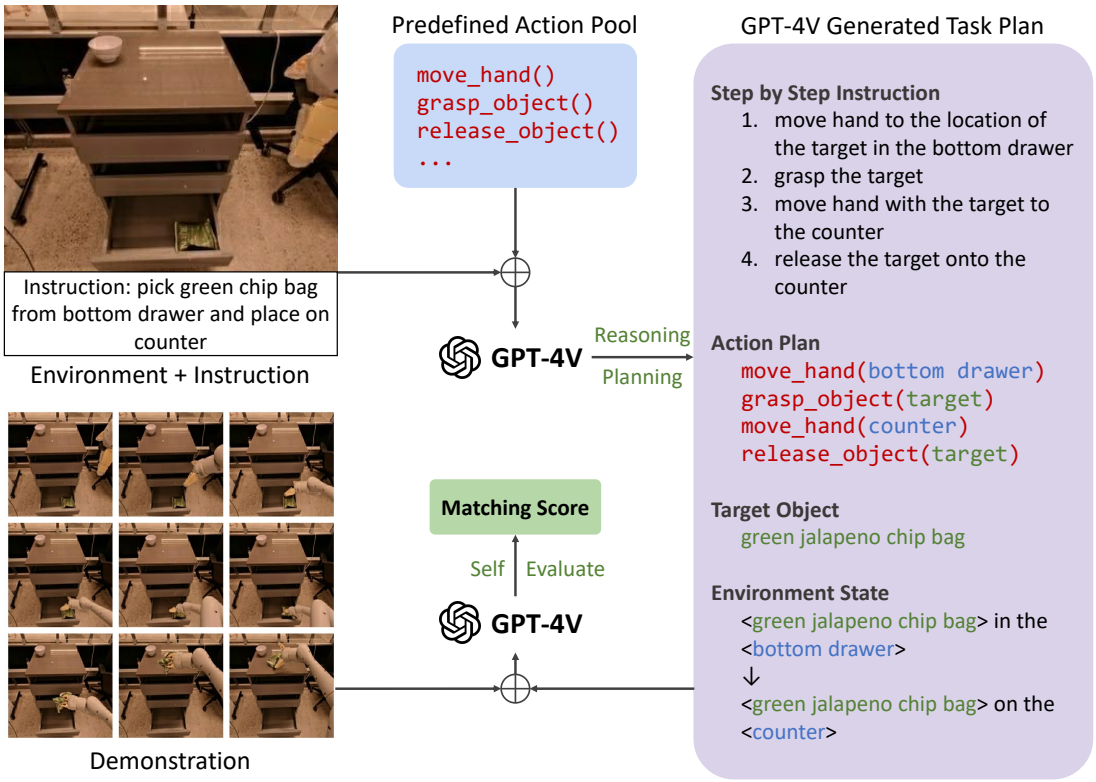


Fig. 1. Framework of the proposed GPT-4V empowered embodied task planning. We use the initial frame of video data along with their corresponding textual instructions as input. Our framework, leveraging GPT-4V, breaks down the instructions into a sequence of task plans and selects corresponding representations from a predefined action pool. Simultaneously, we can analyze the target object related to the instruction and the environmental changes before and after the instruction in the images. Finally, we employ GPT-4V to compare and score the task plan we generated against the ground truth plan.

leverage extensive datasets from the internet, acquiring large and diverse datasets for robot interactions is challenging [49]. These datasets often either focus on a single environment and object or emphasize specific task domains, resulting in substantial differences between them. [50] This intricacy presents more significant challenges when integrating LLMs with robotics.

How to overcome the challenges posed by robotic technology and harness the accomplishments of LLMs in other domains for the benefit of the robotics field is the central inquiry addressed in this review. In this article, the work's contributions can be summarized in four main points.

- We meticulously survey and synthesize existing LLM for robotic literature, exploring the latest advancements in three distinct task categories: planning, manipulation, reasoning.
- We summarize the primary technical approaches that LLMs offer to the realm of robotics, examine the potential for training generalized robot strategies, and provide a foundational survey for researchers in this domain.
- We assess the effectiveness of multimodal GPT-4V in robot task planning across various environments and scenarios.
- We summarize the key findings of our investigation, deliberate upon the outstanding challenges to be tackled in future endeavors, and present a forward-looking perspective.

II. Related Work

A. LLM for Robotics

The field of robotics research based on LLMs has made significant strides. These models demonstrate exceptional natural language understanding and commonsense reasoning capabilities, significantly enhancing a robot's ability to comprehend contexts and execute commands. Current research focuses on leveraging LLMs to parse complex contexts and instructions, including addressing ambiguity, resolving ambiguities, and understanding implicit information. A key advancement in this domain includes the development of vision-language models, [51]–[53] which have markedly improved the performance of tasks like visual question answering [54]–[56] and image captioning. [57], [58] These advancements have greatly boosted a robot's ability to reason in the physical world, particularly in areas such as complex command navigation. [59], [60] Through visual language processing systems, robots are capable of understanding image content and integrating it with relevant linguistic information, such as image descriptions and command execution. This multimodal information processing is similarly applied in audio-visual integration. Another major progress with LLMs is in human-robot interaction, facilitated by interactive learning processes that better align with human needs and preferences. For example, by integrating reinforcement learning with human feedback, robots can continuously improve their task execution, addressing semantic ambiguities encountered in large model applications, by combining human

guidance with large language models, robots can refine instructions more precisely, thereby better achieving autonomous learning and environmental adaptation for more accurate and targeted control. Robots can also learn and adapt to user behavior, preferences, and needs through interaction, providing a more personalized and customized interaction experience. These advancements not only enhance the practicality of robotic technology but also open up new possibilities for future human-machine interactions.

B. Multimodal Task Planning with LLMs

Multimodal Tasks Planning within the domain of LLMs constitutes a sophisticated intersection of artificial intelligence disciplines, engaging an amalgamation of disparate data modalities — such as textual, visual, and auditory inputs — to foster a more holistic and nuanced AI-driven analysis [61]–[65].

This interdisciplinary approach transcends the traditional boundaries of LLMs, which predominantly focused on textual comprehension and generation, ushering in an era where these models are adept at interpreting, correlating, and interacting with multiple data streams in unison. In this context, the LLM's role evolves from mere language processing to a more integrative function, synthesizing and responding to complex data interplays. In the realm of Multimodal Tasks Planning with LLMs, recent advancements exemplified by projects like Inner Monologue and SayCan demonstrate the burgeoning complexity and sophistication in this field. Inner Monologue's [65] methodology represents a significant leap in this domain, as it integrates multi-modal feedback sources from the environment. This integration enables the generation of more reliable and contextually aware task planning, harmonizing different sensory inputs to create a more cohesive understanding of the AI's surroundings. Similarly, the SayCan's [61] framework introduces a novel dimension to LLM applications. This system employs LLMs as a proxy for the model's "hands and eyes," generating the optimal long-horizon instructions and effectively scoring the affordance probability of the instruction on the current scene. This methodology not only enhances the AI's ability to understand and interact with its immediate environment but also leverages the nuanced understanding of LLMs to plan and execute complex sequences of actions over extended periods.

The integration of these advanced techniques in Inner Monologue and SayCan within multimodal task planning with LLMs represents a significant stride towards creating AI systems that are not only more cognizant of multiple data streams but are also capable of synthesizing these streams into actionable intelligence. This progression points towards a future where AI can navigate and interact with the real world in a manner that is far more dynamic, context-aware, and autonomous [61], [65]–[67], pushing the boundaries of what is achievable in AI-driven innovation and interdisciplinary synthesis.

III. Scope of Robotic Tasks

A. Planning

1) Natural Language Understanding : In robot planning, Large Language Models excel due to their advanced natural language comprehension. They translate natural language instructions into executable action sequences for robots, a crucial aspect of robot planning [61], [68]. This study reveals that LLMs can generate accurate action sequences based on linguistic instructions alone, even without visual input [69]. However, their performance is enhanced significantly with a modest amount of visual information, enabling them to create precise visual-semantic plans. These plans transform high-level natural language instructions into actionable guidance for virtual agents to undertake complex tasks. This ability underscores the potential of LLMs to integrate multimodal information, thereby improving their comprehension. It also demonstrates their capacity to interpret and incorporate information from various modalities, leading to a more comprehensive task understanding [70]. Moreover, research in generating action sequences from a large language model for natural language understanding further confirms the efficacy of LLMs in robot planning. LLMs also show great promise in interpreting natural language commands in sync with the physical environment. Employing the Grounded Decoding approach, they can produce behavior sequences that align with the probabilities of the physical model, showcasing this method's effectiveness in robot planning tasks [71].

The research in complex sequential task planning has highlighted significant advancements in the capabilities of LLMs. Text2Motion's studies demonstrate that LLMs are adept not only at processing linguistic information but also at addressing dependencies in skill sequences [72]. This is achieved through geometrically feasible planning, marking a crucial advancement in the interpretation of abstract instructions and the comprehension of intricate task structures. In addition, LLM-Planner research enhances the natural language understanding abilities of LLMs in robotic planning by integrating them with conventional planners [73]. This synergy illustrates how the NLP proficiencies of LLMs can be harnessed to boost the efficiency and precision of planning tasks. Moreover, LLM+P harnesses the capabilities of classical planners, employing the Planning Domain Definition Language (PDDL) and problem cues to create task-specific problem files for LLMs [44]. This integration significantly amplifies LLMs' efficacy in addressing long-term planning tasks. Also, SayPlan addresses the issue of planning horizon by integrating a classical path planner. By doing so, SayPlan is capable of grounding large-scale, long-horizon task plans derived from abstract and natural language instructions, enabling a mobile manipulator robot to execute them successfully [74]. Furthermore, LLMs have shown promise in acting as heuristic strategies within search algorithms, while also serving as reservoirs of common-sense knowledge. This dual role of LLMs not only enhances the reasoning capabilities within these algorithms but also aids in forecasting potential outcomes. Such an approach harnesses the full potential of LLMs, leveraging their advanced reasoning capabilities for the effective planning of complex tasks [66].

This dual application underscores the extensive and versatile potential of large language models in task planning and problem-solving.

The research conducted on LLMs has showcased their remarkable ability to parse and comprehend natural language understanding. This capability extends beyond mere text matching to a profound semantic understanding, encompassing the tasks' purpose and context. A critical aspect of LLMs is translating the instructions they comprehend into executable action sequences for robots, an essential feature in robot task planning. LLMs significantly enhance the quality and adaptability of instruction generation, enabling the creation of complex action sequences that are both context-aware and environment-specific. These models demonstrate versatility in managing various task-planning complexities and types, from straightforward physical interactions to intricate, long-term sequence planning. The studies highlight LLMs' potential as both independent decision-makers and collaborators with other modalities and planning algorithms. This collaboration is pivotal in interpreting natural language and advancing robotic planning. As research progresses, LLMs are expected to play an increasingly vital role in the fields of robotics and automated systems.

2) Complex Task Reasoning and Decision-making: In the realm of complex task reasoning and decision-making, robots empowered by LLMs have shown remarkable proficiency. These LLM-based robotic planning tasks have significantly transcended the realms of mere text generation and language comprehension. Recent research highlights the immense capabilities of Language Models in managing intricate tasks, engaging in logical reasoning, making informed decisions, and partaking in interactive learning. [3], [75] These breakthroughs have not only expanded our comprehension of LLM-based robotic planning's potential but also opened the door to innovative practical applications.

In exploring the application of pre-trained language models (PLMs) in interactive decision-making, research has demonstrated how targets and observations are transformed into embedding sequences, initializing the network with PLMs. This strategy's generalization ability is particularly effective in multivariate environments and supervised modalities [76]. A notable advancement in the multimodal domain is the development of the LM-Nav system [59]. This system, grounded in PLMs, integrates language, vision, and action models to guide robotic navigation via high-level natural language commands. Significantly, it reduces dependency on costly trajectory annotation supervision by merging pre-trained visual navigation, image-verbal correlation, and language understanding models. Focusing on LLMs in specific environments, researchers [65] have examined their capacity for reasoning with natural language feedback and complex task planning. This capability is crucial for following high-level task instructions and enhancing the model's applicability in real-world scenarios. Addressing the issue of consistency fault-tolerance in natural language understanding and decision-making, the innovative ReAct model [77] overcomes prior limitations of linguistic reasoning in interactive settings. It tackles challenges like hallucination generation and misinformation propagation. By leveraging LLMs' potential to maintain working memory and abstractly conceptualize high-level goals, the ReAct model achieves significant performance improvements across various tasks. In parallel, to address confidently hallucinated predictions in large language models (LLMs) applied to robotics, KnowNo [78] provides statistical guarantees for task completion while minimizing the need for human assistance in complex multi-step planning scenarios. Notably, KnowNo seamlessly integrates with LLMs without requiring model-finetuning, offering a lightweight and promising method to model uncertainty. This approach aligns with the constantly evolving capabilities of foundation models, providing a scalable solution. Further, a strategy involving preconditioned error cues has been proposed, enabling LLMs to extract executable plans. This approach offers a fresh perspective on the independence and adaptability of agents in task execution. In terms of multi-agent collaboration, the integration of language models with action agents is increasingly being explored. By pairing LLMs with agents executing tasks in specific environments, a system comprising planners, executors, and reporters is established. This arrangement markedly enhances the efficiency of reasoning and execution in complex tasks.

The burgeoning field of large pre-trained LMs is witnessing a notable trend: these models are increasingly adept at understanding and performing complex tasks, closely aligning with real-world scenarios. This advancement not only underscores the adaptability and versatility of pre-trained models but also heralds the advent of next-generation AI. As these technologies evolve, we anticipate a surge in innovative applications, poised to revolutionize various industries. A key aspect of these tasks is the utilization of LLMs' robust language comprehension and generation capabilities for intricate reasoning and decision-making processes. Each study in this domain explores the potential of LLMs in complex cognitive functions. Many models employ self-supervised learning, with some incorporating fine-tuning to better align with specific tasks. This approach enables LLMs to excel in downstream task-assisted reasoning, leading to more precise and tailored decisions. Despite the widespread use of LLMs in complex reasoning and decision-making, the specific techniques and approaches vary, particularly in terms of task handling, learning strategies, and feedback mechanisms. These models find applications in diverse real-world contexts, including home automation, robot navigation, and task planning, demonstrating their broad and evolving utility.

3) Human-robot interaction: In the realm of human-robot interaction, the advanced reasoning capabilities of AGI language models empower robots with a significant degree of generalization ability. [79] This enables them to adapt to new task planning in previously unseen environments and tasks. Furthermore, the natural language understanding interface of LLMs facilitates communication with humans, opening new possibilities for human-robot interactions. [80] Extensive research has underscored the progress made by LLMs in aiding intelligent task planning, which in turn enhances multi-intelligence collaborative communication. Studies have found that using natural language to boost the efficiency of multi-intelligence cooperation is an effective method to

enhance communication efficiency. A notable example of this is OpenAI's ChatGPT, whose capabilities in robotics applications were evaluated through rigorous experiments. The findings revealed that ChatGPT excels in complex tasks such as logical, geometric, and mathematical reasoning, along with airborne navigation, manipulation, and controlling embodied agents [48]. It achieves this through techniques like free-form dialogue, parsing XML tags, and synthesizing code. Furthermore, ChatGPT allows user interaction via natural language commands, providing vital guidance and insights for the development of innovative robotic systems that interact with humans in a natural and intuitive way. In a similar vein, there is a proposed framework that leverages large-scale language models for collaborative embodied intelligence [81]. This framework enables the use of language models for efficient planning and communication, facilitating collaboration between various intelligences and humans to tackle complex tasks. Experimental results demonstrate that this approach significantly outperforms traditional methods in the field.

B. Manipulation

1) Natural Language Understanding: In the field of robot control, the natural language understanding capabilities of LLM can help robots make common-sense analyses. For example, LLM-GROP demonstrates how semantic information can be extracted from LLM and used as a way to make common-sense, semantically valid decisions about object placement as part of a task and motion planner that performs multistep tasks in complex environments in response to natural language commands [82]. The research proposes a framework for placing language at the core of an intelligent body [83]. By utilizing the prior knowledge contained in these models, better robotic agents can be designed that are able to solve challenging tasks directly in the real world. Through a series of experiments, it is demonstrated how the framework can be used to solve a variety of problems with greater efficiency and versatility by utilizing the knowledge and functionality of the underlying models. At the same time, the study introduces Linguistically Conditional Collision Function (LACO), a novel method to learn collision functions using only single-view image, language prompt, and robot configuration. LACO predicts collisions between robots and the environment, enabling flexible conditional path planning [84].

Outside of natural language understanding capabilities, the powerful reasoning capabilities of LLM also have a prominent role. For example, in the VIMA work [85], a novel multimodal cueing formulation is introduced to convert different robot manipulation tasks into a unified sequence modeling problem and instantiated in a diverse benchmark with multimodal tasks and system generalization evaluation protocols. Experiments show that VIMA is capable of solving tasks such as visual goal realization, one-off video imitation, and novel conceptual foundations using a single model, with robust model scalability and zero-sample generalization. Similarly, TIP proposes Text-Image Cueing [86], a bimodal cueing framework that connects LLMs to multimodal generative models for rational multimodal program plan generation.

In addition to prompt methods, fine-tuning downstream tasks based on pre-trained LMs is also a common approach in the field of robot control. For example, the work demonstrated that pre-trained visual language representations can effectively improve the sample efficiency of existing exploratory methods [87]. R3M investigates how pre-trained visual representations on different human video data can enable data-efficient learning of downstream robot manipulation tasks [88]. The LIV is trained on a large generalized human video dataset, and fine-tuned on a small robot data set, is fine-tuned to outperform state-of-the-art methods in three different evaluation settings, and successfully performs real-world robotics tasks [89].

This collection of studies collectively illustrates the significant role of LLMs and Natural Language Understanding techniques in advancing robotic intelligence, particularly in comprehending and executing complex, language-based tasks. A key emphasis of these studies is on the importance of model generalization and the ability to apply these models across various domains. Each study, while sharing this common theme, diverges in its specific focus and application methodology. For instance, LLM-GROP is dedicated to the extraction and application of semantic information. In contrast, VIMA and TIP concentrate on multimodal processing and learning without prior examples. Furthermore, methodologies that fine-tune pre-trained LMs are directed toward enhancing application efficiency and task-specific optimization. Collectively, these studies demonstrate that integrating sophisticated NLP techniques with machine learning strategies can substantially enhance the efficiency of robotic systems, particularly in their ability to understand and perform intricate tasks. This advancement is a crucial stride towards achieving greater intelligence and autonomy in robotic manipulation.

2) Interactive Strategies: In the realm of interactive strategies, the TEXT2REWARD framework introduces an innovative approach for generating interactive reward codes using LLMs [83]. This method automatically produces dense reward codes, enhancing reinforcement learning. Also, By utilizing Large Language Models to define reward parameters that can be optimized to accomplish a variety of robotic tasks, the gap between high-level language instructions or corrections and low-level robot actions can be effectively bridged. The rewards generated by the language models serve as an intermediate interface, enabling seamless communication and coordination between high-level instructions and low-level actions of the robot [90]. Furthermore, VoxPoser presents a versatile framework for robot manipulation [64], distinct in its ability to extract manipulability and constraints directly from LLMs. This approach significantly enhances the adaptability of robots to open-set instructions and diverse objects. By integrating LLMs with vision-language models and leveraging online interactions, VoxPoser efficiently learns to interact with complex task dynamics models. The application of LLMs extends to human-robot interaction as well. The LILAC system exemplifies this through a scalable [63], language-driven interaction mechanism between humans and robots. It translates natural language discourse into actionable commands within a low-dimensional control space, enabling precise and user-friendly guidance of robots. Importantly, each user correction

refines this control space, allowing for increasingly targeted and accurate commands. InstructRL offers another innovative framework designed to enhance human-AI collaboration [91]. It focuses on training reinforcement learning agents to interpret and act on natural language instructions provided by humans. This system employs LLMs to formulate initial policies based on these instructions, guiding reinforcement learning agents toward achieving an optimal balance in coordination. Lastly, for language-based human-machine interfaces, a novel, flexible interface LILAC has been developed. It permits users to alter robot trajectories using textual input and scene imagery [92]. This system synergizes pre-trained language and image models like BERT and CLIP, employing transformer encoders and decoders to manipulate robot trajectories in both 3D and velocity spaces. Proving effective in simulated environments, this approach has also demonstrated its practicality through real-world applications.

All of these techniques and approaches depend, to varying degrees, on advanced language modeling to enhance human-robot interaction and robot control. They collectively underscore the crucial role of LLMs in interpreting and executing human intentions. Each method aims to boost the adaptability and flexibility of robots, enabling them to handle diverse tasks and environments more effectively. Specifically, TEXT2REWARD centers on generating and optimizing reward codes. This enhances the efficacy of reinforcement learning strategies. Conversely, VoxPoser focuses on extracting operants and constraints from LLMs. Meanwhile, LILAC and InstructRL adopt distinct approaches to interpreting and executing natural language commands. LILAC prioritizes mapping discourse to a control space, whereas StructRL dedicates itself to training reinforcement learning agents to comprehend and follow natural language instructions. Additionally, the last discussed language-based Human-Machine Interaction research investigates how to directly extract user intentions from text and images, applying them across various robot platforms. This aspect sets it apart from other approaches that might not incorporate this feature. Collectively, these studies mark substantial advancements in integrating LLMs techniques into robotics. While their application areas and methodologies have distinct focal points, they collectively demonstrate the potential for innovation in artificial intelligence. Furthermore, they pave the way for future explorations in human-robot interaction.

3) Modular Approaches: Recent advancements in robot control emphasize modular approaches, allowing the creation of more complex and feature-rich robotic systems. Key aspects of this trend have been highlighted in recent research. PROGRAMPORT proposes a program-based modular framework focused on robot manipulation [93]. It interprets and executes linguistic concepts by translating natural language's semantic structure into programming elements. The framework comprises neural modules that excel in learning both general visual concepts and task-specific operational strategies. This structured approach distinctly enhances learning of visual foundations and operational strategies, improving generalization to unseen samples and synthetic environments.

Next, researchers have explored the use of LLMs to expedite strategy adaptation in robotic systems [94], particularly when encountering new tools. By generating geometrical shapes and descriptive tool models, and then converting these into vector representations, LLMs facilitate rapid adaptation. This integration of linguistic information and meta-learning has shown significant performance improvements in adapting to unfamiliar tools.

In addition, Combining NLMap [95], a visual language model based on ViLD and CLIP, with the SayCan framework, has led to a more flexible scene representation. This combination is particularly effective for long-term planning, especially when processing natural language commands in open-world scenarios. NLMap enhances the capability of LLM-based planners to understand their environments.

The "Scaling Up and Distilling Down" framework combines the advantages of LLMs [96], sampling-based planners, and policy learning. It automates the generation, labeling, and extraction of rich robot exploration experiences into a versatile visual-linguistic motion strategy. This multi-task strategy not only inherits long-term behavior and robust manipulation skills but also shows improved performance in scenarios outside the training distribution.

MetaMorph introduces a Transformer-based method for learning a generalized controller applicable to a vast modular robotic design space [97]. This approach enables the use of robot morphology as a Transformer model output. By pre-training on a diverse range of morphologies, strategies generated through this approach demonstrate broad generalizability to new morphologies and tasks. This showcases the potential for extensive pre-training and fine-tuning in robotics, akin to developments in vision and language fields.

In each of these studies, a modular approach has been adopted, enhancing the system's flexibility and adaptability to new tasks and environments. These works extensively utilize deep learning techniques, notably in synergy with LLMs, to augment the robotic system's understanding and decision-making abilities. Moreover, a significant focus of these studies is the application of NLP. This is evident either through the direct interpretation of linguistic commands or via linguistically enriched learning and adaptation processes. The primary objective is to enhance the robot's capability for quick generalization and adaptation in novel environments and tasks. While all the studies employ deep learning and LLMs, their specific implementations and applications are diverse. Some are centered on linguistic description and comprehension, while others explore the fusion of vision and language. The research goals are varied, addressing challenges from adapting to new tools, to long-term strategic planning, to polymorphic robot control. Despite differences in technical approaches, application areas, and targeted tasks, each study significantly contributes to advancing the intelligence and adaptive capabilities of robotic systems.

C. Reasoning

1) Natural Language Understanding: In the realm of robotic reasoning tasks, LLMs based on natural language understanding serve as an essential knowledge base, providing common sense insights crucial for various tasks. Extensive

research has shown that LLMs effectively simulate human-like states and behaviors, especially relevant in the study of robots performing household cleaning functions. This approach deviates from traditional methods, which typically require costly data gathering and model training. Instead, LLMs leverage off-the-shelf methods for generalization in robotics, benefiting from their robust summarization abilities honed from extensive textual data analysis. Moreover, the common sense reasoning and code comprehension capabilities of LLMs foster connections between robots and the physical world. For instance, Progprompt introducing programming language features in LLMs has been shown to enhance task performance. This approach is not only intuitive but also sufficiently flexible to adapt to new scenarios, agents, and tasks, including actual robot deployments [98]. Concurrently, GIRAF harnesses the power of large language models to more flexibly interpret gestures and language commands, enabling accurate inference of human intentions and contextualization of gesture meanings for more effective human-machine collaboration [99].

One innovative development in this field is Cap (Code as Policies) [47], which advocates for robot-centric language model generation programs. These programs can be adapted to specific layers of the robot's operational stack: interpreting natural language commands, processing perceptual data, and parameterizing low-dimensional inputs for the original language control. The underlying principle of this approach is that layered code generation facilitates the creation of more intricate code, thereby advancing the state-of-the-art in this area.

Both the home cleaning application and the robot-centric language model generation programs in Cap highlight the strengths of LLMs in providing common sense knowledge and interpreting natural language instructions. Traditional robotics often necessitates extensive data collection and specialized model training. In contrast, LLMs mitigate this need by utilizing their extensive training on textual data. The code comprehension and generation abilities of LLMs are particularly crucial, enabling robots to interact more effectively with the physical world and execute complex tasks. However, there is a distinction in application focus: the home cleaning function tends to emphasize everyday tasks and environmental adaptability, whereas Cap centers on programming and controlling the robot's more technical behaviors through Language Model Generation Programs (LMPs).

In summary, the integration of LLMs into robotic reasoning tasks underscores their remarkable capabilities in natural language understanding, common sense knowledge provision, and code comprehension and generation. These features not only alleviate the data collection and model training burdens typically associated with traditional robotics but also enhance robots' generalization and flexibility. With adequate training and adjustment, LLMs can be applied across various scenarios and tasks, demonstrating their vast potential and wide-ranging applicability in the future of robotics and artificial intelligence.

2) Complex Task Reasoning and Decision-making: In the realm of complex task reasoning and decision-making, various studies have leveraged the reasoning abilities of LLMs to augment the refinement of specific downstream tasks. For instance, SayCan utilizes the extensive knowledge embedded in LLMs for concretization tasks alongside reinforcement learning [61]. This method involves using reinforcement learning to uncover insights about an individual's skill value function. It then employs textual labels of these skills as potential responses, while the LLM provides overarching semantic guidance for task completion.

Another notable development is the Instruct2Act framework [100]. It offers a user-friendly, general-purpose robotics system that employs LLMs to translate multimodal commands into a sequence of actions in the robotics field. This system uses policy code generated by LLMs, which make API calls to various visual base models, thus attaining a visual comprehension of the task set.

The usage of LLMs for self-planning and in PDDL (Planning Domain Definition Language) planning has also been explored [101]. It has been shown that LLM outputs can guide heuristic search planners effectively.

In the domain of failure explanation and correction tasks, the REFLECT framework leverages a hierarchical summary of the robot's past experiences generated from multisensory observations to query an LLM for failure reasoning [102]. The failure explanation obtained can then guide a language-based planner to correct the failure and successfully accomplish the task.

Furthermore, the adaptation of pre-trained multimodal models is a common strategy. By integrating the pre-training of vision-language models with robot data to train Visual-Linguistic-Action (VLA) models [62], researchers have found that models trained on internet data with up to 55 billion parameters can generate efficient robot strategies. These models exhibit enhanced generalization performance and benefit from the extensive visual-linguistic pre-training capabilities available on the web.

Socratic Models represent another approach [67], where structured dialogues between multiple large pre-trained models facilitate joint predictions for new multimodal tasks. This method has achieved zero-shot performance across multiple tasks.

In these studies, the primary focus has been on harnessing LLMs for automating reasoning and decision-making processes. This is achieved by leveraging LLMs' capacity to provide or utilize high-level semantic knowledge, thereby enhancing task execution. Some approaches integrate LLMs with other modalities, like vision and action, to deepen task understanding and execution. Others demonstrate effective performance on previously unseen tasks, showcasing zero-shot or few-shot learning capabilities.

Each study adopts a unique approach to integrate LLMs. For example, SayCan incorporates reinforcement learning, whereas Instruct2Act is centered on the direct mapping of multimodal instructions. The techniques employed—ranging from reinforcement learning and heuristic search to multimodal pretraining—vary significantly across different application domains like robot manipulation, planning, and automated decision-making. These studies collectively illustrate LLMs' vast potential in managing complex task reasoning and decision-making. By amalgamating LLMs with other

techniques, such as reinforcement learning and multimodal data processing, a deeper semantic understanding and more effective decision support can be achieved. This is particularly evident in robotics and automation, where such integrated approaches are paving the way for novel applications. However, the efficacy of these methods is highly contingent on the specific nature of the task, the data utilized, and the model training approach. Hence, the selection and application of each method must be meticulously tailored to the specific context.

3) Interactive Strategies: The recent advancements in LLMs have significantly contributed to the development of interactive strategies, showcasing impressive capabilities in language generation and human-like reasoning. Matcha [103], utilizing LLMs, enhances interactive multimodal perception, illustrating the potential of LLMs in understanding various types of input data, such as visual and auditory. This approach proposes an augmented LLM multimodal interactive agent. This agent not only leverages commonsense knowledge inherent in LLMs for more plausible interactive multimodal perception but also demonstrates the practical application of LLMs in conducting such perception and interpreting behavior.

Generative agents, as introduced are interactive computational agents designed to simulate human behavior [104]. The architecture of these agents is engineered to store, synthesize, and apply relevant memories, thereby generating plausible behaviors using large language models. The integration of LLMs with these computational agents facilitates the creation of advanced architectures and interaction patterns. This combination enables more realistic simulations of human behavior, extending the potential applications of LLMs.

The emphasis in LLM-based interactive strategies is on the fusion of LLMs with other perceptual systems, such as image recognition and speech processing. This amalgamation aims to mimic or augment human abilities, enhancing cognitive and processing capabilities. Such advancements have profound implications in the realms of intelligent assistants, robotics, and augmented reality systems.

In the discussed work, a notable emphasis is placed on multimodal perception, focusing on improving the system's ability to understand and interact with its environment. Additionally, the simulation of human behavior seeks to replicate human thought and action processes in AI. The convergence of these two directions holds the promise of creating more powerful and versatile intelligent systems. These systems are envisioned to interact with humans at a more complex and humanized level, presenting significant technical challenges as well as raising crucial ethical and social adaptation questions.

IV. GPT-4V Empowered Embodied Task Planning

Based on the aforementioned investigation into embodied tasks and LLMs, we developed an embodied task planning framework based on GPT-4V in this study and conducted evaluation experiments, as shown in Fig. 1. The following section provides detailed information on the datasets, prompt design, and experimental results.

[1]See https://docs.google.com/spreadsheets/d/1rPBD77tk60AEIGZrGSODwyyzs5FgCU9Uz3h-3_t2A9g/edit?gid=0 for more information.

A. Datasets

To comprehensively evaluate the multimodal embodied task planning capabilities of GPT-4V, over 40 cases from 9 datasets are selected, focusing on manipulation and grasping. These actions are fundamental in instruction-following robotics, involving a variety of human instructions across diverse scenarios, such as kitchen pickups and tabletop rearrangements. The selected datasets are accessed through the Google Open X-Embodiment Dataset [49]. In each case, video demonstrations and natural language instructions serve as inputs to assess GPT-4V as a robotic brain. This setup enables robust planning based on natural language instructions for generating robot actions.

B. Prompt Design

The design of prompts plays a crucial role in querying LLMs. A meticulously crafted prompt, rich in information and structured clearly, yields more precise and consistent outputs aligned with the given instructions. Here we update the text prompt from [114] by incorporating images, creating a multimodal prompt that guides GPT-4V to produce robot task plans. The multimodal prompt consists of five parts:

- **System Role Explanation**: Specifies the task and the persona GPT-4V adopts in its responses.
- **Predefined Action Pool**: A set of predefined robot actions from which GPT-4V can select and sequence to complete tasks step-by-step. To address vocabulary limitations, GPT-4V is prompted to create new actions if necessary.
- **Example Output**: An example in JSON format to illustrate the expected output and ensure consistency.
- **Case-by-Case Environment Image and Natural Language Instruction**: Includes the first frame extracted from the video demonstration as the environment image.
- **Evaluation**: GPT-4V is tasked to assess the generated task plan against the ground truth video demonstration, scoring the plan based on its alignment with the video and providing an explanation.

The first three components are input as system messages for each query, while the last two as user messages vary according to the test data. The complete prompt is depicted in Fig. 4 of Appendix.

V. Experimental Results

In our experimental framework, the Large Language Models (LLMs) first generate step-by-step instructions tailored to the objectives of each robotic task. Subsequently, guided by these generated instructions, the model selects the most appropriate action from a predefined action pool and action objects to form the action plan for each step. After obtaining the instructions generated by the LLMs, we quantitatively evaluated the generated results by comparing them with the Ground-Truth instructions from the respective video dataset. Rigorous testing was conducted on 9 publicly available robot datasets, resulting in profound and insightful findings.

TABLE I

DESCRIPTION [1] OF THE DATASET AND THE AVERAGE MATCHING SCORES SELF-EVALUATED BY GPT-4V, COMPARING THE TASK PLANS IT GENERATED WITH THE GROUND TRUTH DEMONSTRATIONS ACROSS NINE TESTED DATASETS.

Dataset	Description	Matching Score
RT-1 Robot Action [105]	Robot picks, places and moves 17 objects from the google micro kitchens.	9/10
QT-Opt [106]	Kuka robot picking objects in a bin.	8/10
Berkeley Bridge [107]	The robot interacts with household environments including kitchens, sinks, and tabletops.	8.7/10
TOTO Benchmark [108]	The TOTO Benchmark Dataset contains trajectories of two tasks: scooping and pouring.	8.5/10
BC-Z [109]	The robot attempts picking, wiping, and placing tasks on a diverse set of objects on a tabletop,	7.7/10
Berkeley Autolab UR5 [110]	The data consists of 4 robot manipulation tasks.	8.3/10
NYU VINN [111]	The robot arm performs diverse manipulation tasks on a tabletop.	9/10
Freiburg Franka Play [112]	The robot interacts with toy blocks, it pick and places them, stacks them,	10/10
USC Jaco Play [113]	The robot performs pick-place tasks in a tabletop toy kitchen environment.	9/10
All	-	8.7/10

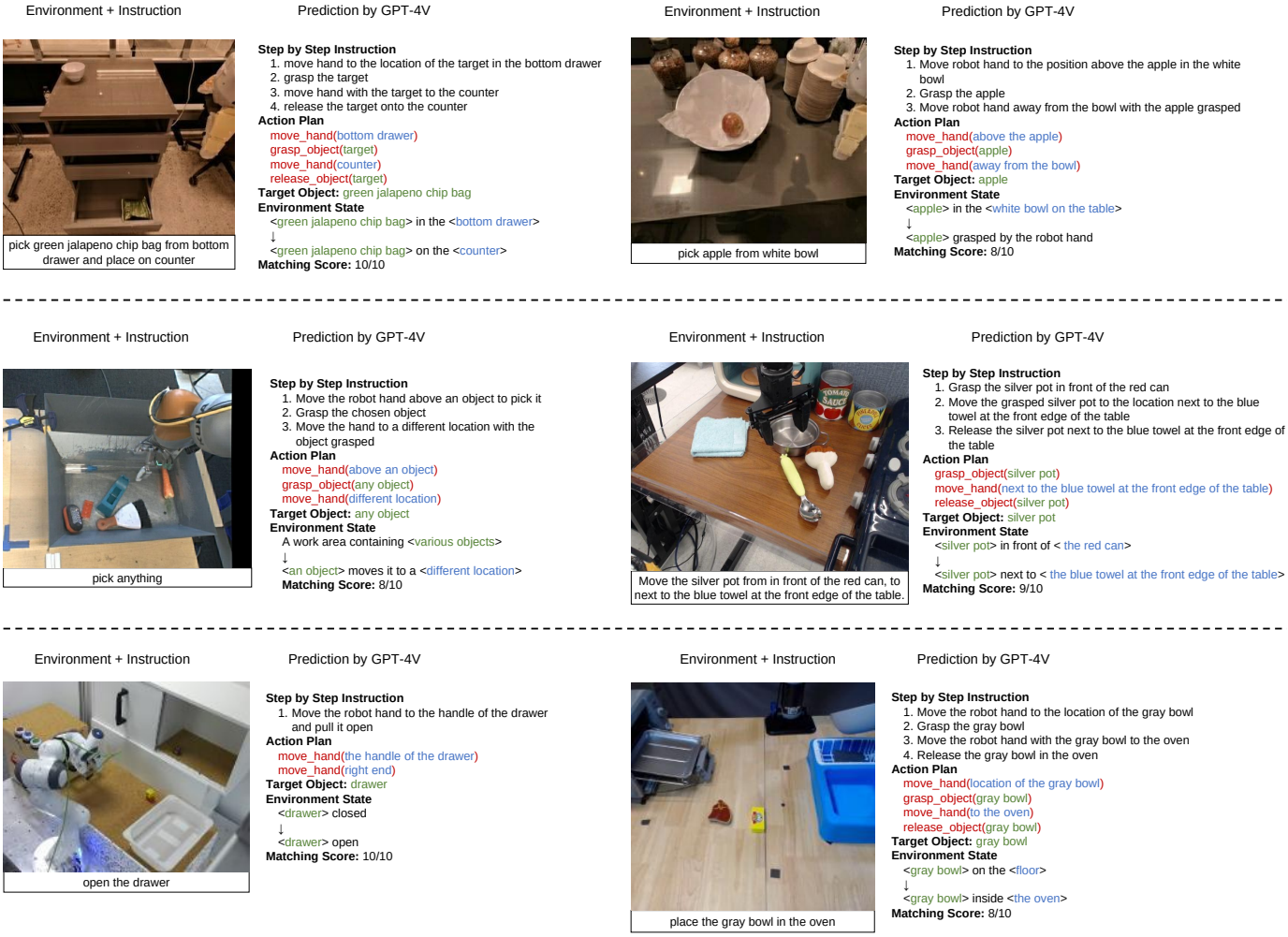


Fig. 2. Generated task plans for different datasets: RT-1 Robot Action (Top Panel), QT-Opt (Middle Left), Berkeley Bridge (Middle Right), Freiburg Franka Play (Bottom Left), and USC Jaco Play (Bottom Right).

For instance, in the RT-1 Robot Action [105] dataset, as depicted in Fig. 2 top panel, the multi-modality LLMs accurately identified the target object and proficiently decomposed and executed the task. As shown in Fig. 2 top left, based on the given environment and instruction, the instructions generated by LLMs were as follows: 1) Move the hand to the target's location in the bottom drawer; 2) Grasp the target; 3) Move the hand with the target to the counter; 4) Release the target onto the counter. After providing detailed step-by-step textual instructions, the LLMs select from the action pool and lists a set of instructions and objects that comply with the current strategy. For example, "$move\_hand(bottom\ drawer)$" is the functional expression of the first textual instruction, facilitating subsequent direct usage of this action plan with the interface code controlling the robotic arm. Additionally, through the "Environment State" generated by the LLMs, it is evident that

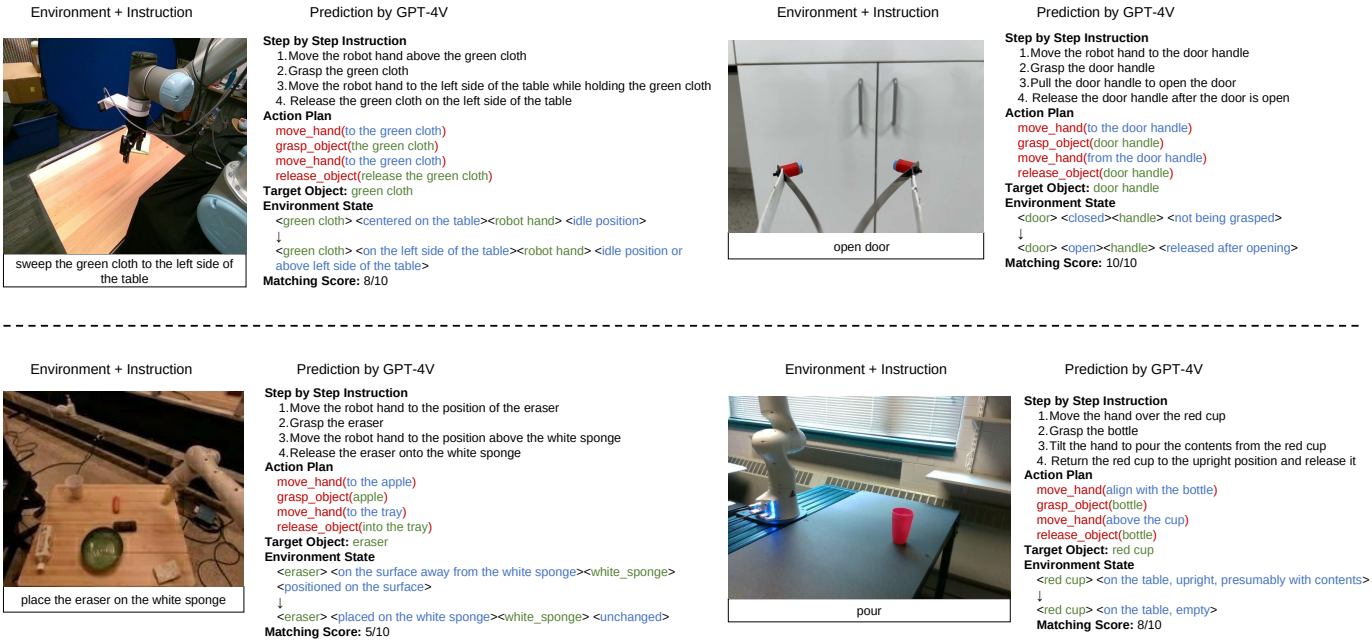


Fig. 3. Generated task plans for different datasets: Berkeley Autolab UR5 (Top Left), NYU VINN (Top Right), BC-Z (Bottom Left), and TOTO Benchmark (Bottom Right).

the models can effectively comprehend the changing spatial relationships of key objects in the environment after a series of operations. The "Matching Score" in Fig. 2 also demonstrates the model's precision.

In the aforementioned testing cases, the scenarios involved fewer objects and relatively concise and clear task instructions. Therefore, we further conducted tests involving semantically vague task descriptions and complex scenes. Fig. 2 middle left represents a test case from the QT-Opt dataset [106], where the instruction is simply "pick anything" without specifying any entities in the scene. From the results generated by LLMs, it produced a series of generalized instructions suitable for grasping any object, maintaining a high level of consistency with the ground truth. For complex scenes, as illustrated in Fig. 2 middle right, we tested an exemplary case from the Berkeley Bridge dataset [107]. The input instruction "Move the silver pot from in front of the red can, to next to the blue towel at the front edge of the table" involves multiple objects and their spatial relationships within the scene. Here, the LLMs not only grasped the task's purpose but also executed the task details adeptly, exemplifying their advanced image comprehension and logical reasoning abilities.

Further evidence of the LLMs' effectiveness in diverse and complex scenarios (including datasets [108]–[113]) is presented in Fig. 2 and Fig. 3. Across these experiments, the LLMs demonstrated remarkable performance, even in tasks with intricate settings or specific requirements. Table I presents the average matching score, self-evaluated by GPT-4V across nine diverse datasets, indicating a consistently high level of agreement between the generated task plans and the ground truth demonstrations. This consolidates the validity of our approach and underscores the potent image understanding and logical reasoning capacities of multimodal LLMs in robotic task execution. Additional test results can be found in the Appendix.

VI. Limitation, Discussion and Future Work

We present an overview of integrating Large Language Models (LLMs) into robotic systems for various tasks and environments and evaluate GPT-4V in multimodal task planning. Although GPT-4V exhibits impressive multimodal reasoning and understanding capabilities as a robot brain for task planning, it faces several limitations: 1) The generated plans are homogenous, lacking in detailed embodiment and specific, robust designs to manage complex environments and tasks. 2) Current multimodal LLMs, such as GPT-4V and Google Gemini [28], necessitate carefully crafted, lengthy prompts to produce reliable outputs, which require domain expertise and extensive tricks. 3) The robot is constrained by predefined actions, limiting its executional freedom and robustness. 4) The closed-source nature of the GPT-4V API and associated time delays may impede embedded system development and real-time commercial applications. Future research should aim to address these challenges to develop more robust AGI robotic systems.

On the other hand, the advanced reasoning and vision-language understanding abilities exhibited by multimodal GPT-4V in robotics highlight the potential of LLM-centric AGI robotic systems. Moving forward, multimodal-LLM-centric AGI robots hold potential for application across various domains. In the realm of precision agriculture, these robots could supplant human labor in various labor-intensive tasks, especially in harvesting. This encompasses tasks like fruit picking and crop phenotyping [115], [116], which require advanced reasoning and precise action in the intricate environment of farms [117]. In the healthcare domain, the critical

need for safety and precision imposes greater demands on the perceptual and reasoning abilities of multimodal LLMs. This aspect is especially vital in robot-assisted screening and surgeries, where custom tasks tailored to individual needs are paramount [118]. Furthermore, leveraging contrastive learning models like CLIP [119] to align brain signals with natural language suggests a pathway for developing Brain-Computer Interfaces (BCIs) in LLM-centric AGI robotic systems [120]. These systems could be capable of reading and interpreting human brain signals, such as EEG and fMRI, for self-planning and control in complex task completion [80], [121]. This advancement could significantly bridge the gap in human-environment interaction and alleviate physical and cognitive labor.

VII. Conclusion

In this paper, we have provided an overview of the integration of Large Language Models (LLMs) into various robotic systems and tasks. Our analysis reveals that LLMs demonstrate impressive reasoning, language understanding, and multimodal processing abilities that can significantly enhance robots' comprehension of instructions, environments, and required actions.

We evaluated the recently released GPT-4V model on over 30 cases across 9 datasets for embodied task planning. The results indicate that GPT-4V can effectively leverage natural language instructions and visual perceptions to generate detailed action plans to accomplish manipulation tasks. This suggests the viability of using multimodal LLMs as robotic brains for embodied intelligence.

However, some challenges remain to be addressed regarding model transparency, robustness, safety, and real-world applicability as we progress towards more practical and capable LLM-based AI systems. Specifically, the black-box nature of large neural models makes it difficult to fully understand their internal reasoning processes and failure modes. Additionally, bridging the gap between simulation and the real-world poses persisting difficulties in transferring policies without performance degradation. Extensive research is still needed to address these issues through techniques like standardized testing, adversarial training, policy adaptation methods, and safer model architectures. Accountability and oversight protocols for autonomous intelligent systems relying on LLMs also warrant thoughtful consideration. Overcoming these multifaceted challenges in a careful, ethical and socially responsible manner remains imperative as we advance progress in this domain.

As language models continue to accumulate extensive grounded knowledge from multimodal data, we anticipate rapid innovations in integrating them with robotics and simulation-based learning. This could enable intuitive development and validation of intelligent robots entirely in simulation using sim-to-real techniques before deployment. Such developments could profoundly enhance and transform how we build, test and deploy intelligent robotic systems.

Overall, the synergistic integration of natural language processing and robotics is a promising frontier filled with opportunities and challenges that warrant extensive future interdisciplinary research.

References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[2] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu *et al.*, "Prompt engineering for healthcare: Methodologies and applications," *arXiv preprint arXiv:2304.14670*, 2023.

[3] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu *et al.*, "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-Radiology*, p. 100017, 2023.

[4] Y. Liu, H. He, T. Han, X. Zhang, M. Liu, J. Tian, Y. Zhang, J. Wang, X. Gao, T. Zhong, Y. Pan, S. Xu, Z. Wu, Z. Liu, X. Zhang, S. Zhang, X. Hu, T. Zhang, N. Qiang, T. Liu, and B. Ge, "Understanding llms: A comprehensive overview from training to inference," *arXiv preprint arXiv:2401.02038*, 2024.

[5] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," 2023.

[6] L. Zhao, L. Zhang, Z. Wu, Y. Chen, H. Dai, X. Yu, Z. Liu, T. Zhang, X. Hu, X. Jiang *et al.*, "When brain-inspired ai meets agi," *Meta-Radiology*, p. 100005, 2023.

[7] Z. Liu, M. He, Z. Jiang, Z. Wu, H. Dai, L. Zhang, S. Luo, T. Han, X. Li, X. Jiang *et al.*, "Survey on natural language processing in medical image analysis." *Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical Sciences*, vol. 47, no. 8, pp. 981–993, 2022.

[8] D. Rothman and A. Gulli, *Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3*. Packt Publishing Ltd, 2022.

[9] M. S. Rahaman, M. T. Ahsan, N. Anjum, H. J. R. Terano, and M. M. Rahman, "From chatgpt-3 to gpt-4: a significant advancement in ai-driven nlp tools," *Journal of Engineering and Emerging Technologies*, vol. 2, no. 1, pp. 1–11, 2023.

[10] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," *arXiv preprint arXiv:2307.10169*, 2023.

[11] Z. Liu, Y. Li, P. Shu, A. Zhong, L. Yang, C. Ju, Z. Wu, C. Ma, J. Luo, C. Chen *et al.*, "Radiology-llama2: Best-in-class large language model for radiology," *arXiv preprint arXiv:2309.06419*, 2023.

[12] C. Ma, Z. Wu, J. Wang, S. Xu, Y. Wei, Z. Liu, L. Guo, X. Cai, S. Zhang, T. Zhang *et al.*, "Impressiongpt: an iterative optimizing framework for radiology report summarization with chatgpt," *arXiv preprint arXiv:2304.08448*, 2023.

[13] Z. Liu, T. Zhong, Y. Li, Y. Zhang, Y. Pan, Z. Zhao, P. Dong, C. Cao, Y. Liu, P. Shu *et al.*, "Evaluating large language models for radiology natural language processing," *arXiv preprint arXiv:2307.13693*, 2023.

[14] S. Rezayi, H. Dai, Z. Liu, Z. Wu, A. Hebbar, A. H. Burns, L. Zhao, D. Zhu, Q. Li, W. Liu, S. Li, T. Liu, and X. Li, "Clinicalradiobert: Knowledge-infused few shot learning for clinical notes named entity recognition," in *Machine Learning in Medical Imaging*, C. Lian, X. Cao, I. Rekik, X. Xu, and Z. Cui, Eds. Cham: Springer Nature Switzerland, 2022, pp. 269–278.

[15] Z. Liu, M. He, Z. Jiang, Z. Wu, H. Dai, L. Zhang, S. Luo, T. Han, X. Li, X. Jiang, D. Zhu, X. Cai, B. Ge, W. Liu, J. Liu, D. Shen, and T. Liu, "Survey on natural language processing in medical image analysis," *Zhong nan da xue xue bao. Yi xue ban = Journal of Central South University. Medical sciences*, vol. 47, no. 8, p. 981—993, August 2022. [Online]. Available: https://doi.org/10.11817/j.issn.1672-7347.2022.220376

[16] W. Liao, Z. Liu, H. Dai, Z. Wu, Y. Zhang, X. Huang, Y. Chen, X. Jiang, W. Liu, D. Zhu, T. Liu, S. Li, X. Li, and H. Cai, "Mask-guided bert for few shot text classification," 2023.

[17] S. Rezayi, Z. Liu, Z. Wu, C. Dhakal, B. Ge, H. Dai, G. Mai, N. Liu, C. Zhen, T. Liu *et al.*, "Exploring new frontiers in agricultural nlp: Investigating the potential of large language models for food applications," *arXiv preprint arXiv:2306.11892*, 2023.

[18] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu *et al.*, "Auggpt: Leveraging chatgpt for text data augmentation," *arXiv preprint arXiv:2302.13007*, 2023.

[19] Z. Liu, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, W. Liu, D. Shen, Q. Li, T. Liu, D. Zhu, and X. Li, "Deid-gpt: Zero-shot medical text de-identification by gpt-4," 2023.

[20] W. Liao, Z. Liu, H. Dai, S. Xu, Z. Wu, Y. Zhang, X. Huang, D. Zhu, H. Cai, T. Liu *et al.*, "Differentiate chatgpt-generated and human-written medical texts," *arXiv preprint arXiv:2304.11567*, 2023.

[21] H. Dai, Y. Li, Z. Liu, L. Zhao, Z. Wu, S. Song, Y. Shen, D. Zhu, X. Li, S. Li, X. Yao, L. Shi, Q. Li, Z. Chen, D. Zhang, G. Mai, and T. Liu, "Ad-autogpt: An autonomous gpt for alzheimer's disease infodemiology," 2023.

[22] Z. Guan, Z. Wu, Z. Liu, D. Wu, H. Ren, Q. Li, X. Li, and N. Liu, "Cohortgpt: An enhanced gpt for participant recruitment in clinical study," *arXiv preprint arXiv:2307.11346*, 2023.

[23] H. Cai, W. Liao, Z. Liu, Y. Zhang, X. Huang, S. Ding, H. Ren, Z. Wu, H. Dai, S. Li *et al.*, "Coarse-to-fine knowledge graph domain adaptation based on distantly-supervised iterative training," *arXiv preprint arXiv:2211.02849*, 2022.

[24] Z. Liu, Z. Wu, M. Hu, B. Zhao, L. Zhao, T. Zhang, H. Dai, X. Chen, Y. Shen, S. Li *et al.*, "Pharmacygpt: The ai pharmacist," *arXiv preprint arXiv:2307.10432*, 2023.

[25] Y. Shi, S. Xu, Z. Liu, T. Liu, X. Li, and N. Liu, "Mededit: Model editing for medical question answering with external knowledge bases," *arXiv preprint arXiv:2309.16035*, 2023.

[26] X. Gong, J. Holmes, Y. Li, Z. Liu, Q. Gan, Z. Wu, J. Zhang, Y. Zou, Y. Teng, T. Jiang *et al.*, "Evaluating the potential of leading large language models in reasoning biology questions," *arXiv preprint arXiv:2311.07582*, 2023.

[27] OpenAI, "Introducing ChatGPT — openai.com," https://openai.com/blog/chatgpt, [Accessed 28-08-2023].

[28] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[29] J. Wang, Z. Liu, L. Zhao, Z. Wu, C. Ma, S. Yu, H. Dai, Q. Yang, Y. Liu, S. Zhang *et al.*, "Review of large vision models and visual prompt engineering," *arXiv preprint arXiv:2307.00855*, 2023.

[30] H. Dai, C. Ma, Z. Liu, Y. Li, P. Shu, X. Wei, L. Zhao, Z. Wu, D. Zhu, W. Liu *et al.*, "Samaug: Point prompt augmentation for segment anything model," *arXiv preprint arXiv:2307.01187*, 2023.

[31] L. Zhang, Z. Liu, L. Zhang, Z. Wu, X. Yu, J. Holmes, H. Feng, H. Dai, X. Li, Q. Li *et al.*, "Segment anything model (sam) for radiation oncology," *arXiv preprint arXiv:2306.11730*, 2023.

[32] Z. Xiao, Y. Chen, L. Zhang, J. Yao, Z. Wu, X. Yu, Y. Pan, L. Zhao, C. Ma, X. Liu, W. Liu, X. Li, Y. Yuan, D. Shen, D. Zhu, T. Liu, and X. Jiang, "Instruction-vit: Multi-modal prompts for instruction learning in vit," 2023.

[33] L. Zhao, Z. Wu, H. Dai, Z. Liu, T. Zhang, D. Zhu, and T. Liu, "Embedding human brain function via transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer Nature Switzerland Cham, 2022, pp. 366–375.

[34] H. Dai, Q. Li, L. Zhao, L. Pan, C. Shi, Z. Liu, Z. Wu, L. Zhang, S. Zhao, X. Wu *et al.*, "Graph representation neural architecture search for optimal spatial/temporal functional brain network decomposition," in *International Workshop on Machine Learning in Medical Imaging*. Springer Nature Switzerland Cham, 2022, pp. 279–287.

[35] Y. Liu, E. Ge, M. He, Z. Liu, S. Zhao, X. Hu, D. Zhu, T. Liu, and B. Ge, "Discovering dynamic functional brain networks via spatial and channel-wise attention," *arXiv preprint arXiv:2205.09576*, 2022.

[36] L. Zhang, J. M. Holmes, Z. Liu, S. A. Vora, T. T. Sio, C. E. Vargas, N. Y. Yu, S. R. Keole, S. E. Schild, M. Bues *et al.*, "Beam mask and sliding window-facilitated deep learning-based accurate and efficient dose prediction for pencil beam scanning proton therapy," *arXiv preprint arXiv:2305.18572*, 2023.

[37] Z. Liu, R. J. Crouser, and A. Ottley, "Survey on individual differences in visualization," in *Computer Graphics Forum, 39: 693-712. doi:10.1111/cgf.14033*, 2020.

[38] X.-A. Bi, K. Chen, S. Jiang, S. Luo, W. Zhou, Z. Xing, L. Xu, Z. Liu, and T. Liu, "Community graph convolution neural network for alzheimer's disease classification and pathogenetic factors identification," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[39] Y. Ding, H. Feng, Y. Yang, J. Holmes, Z. Liu, D. Liu, W. W. Wong, N. Y. Yu, T. T. Sio, S. E. Schild *et al.*, "Deep-learning based fast and accurate 3d ct deformable image registration in lung cancer," *Medical Physics*, 2023.

[40] Y. Ding, Z. Liu, H. Feng, J. Holmes, Y. Yang, N. Yu, T. Sio, S. Schild, B. Li, and W. Liu, "Accurate and efficient deep neural network based deformable image registration method in lung cancer," in *MEDICAL PHYSICS*, vol. 49, no. 6. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2022, pp. E148–E148.

[41] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023.

[42] H. Li, Y. Jiao, K. Davey, and S.-Z. Qiao, "Data-driven machine learning for understanding surface structures of heterogeneous catalysts," *Angewandte Chemie*, vol. 135, no. 9, p. e202216383, 2023.

[43] H. Huang, Y. Feng, C. Shi, L. Xu, J. Yu, and S. Yang, "Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator," *arXiv preprint arXiv:2309.14494*, 2023.

[44] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone, "Llm+p: Empowering large language models with optimal planning proficiency," 2023.

[45] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu, "Aligning large language models with human: A survey," *arXiv preprint arXiv:2307.12966*, 2023.

[46] T. Yoneda, J. Fang, P. Li, H. Zhang, T. Jiang, S. Lin, B. Picker, D. Yunis, H. Mei, and M. R. Walter, "Statler: State-maintaining language models for embodied reasoning," *arXiv preprint arXiv:2306.17840*, 2023.

[47] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

[48] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *Microsoft Auton. Syst. Robot. Res*, vol. 2, p. 20, 2023.

[49] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[50] J. Mai, J. Chen, B. Li, G. Qian, M. Elhoseiny, and B. Ghanem, "Llm as a robotic brain: Unifying egocentric memory and control," *arXiv preprint arXiv:2304.09349*, 2023.

[51] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5579–5588.

[52] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[53] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[54] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, "Medical visual question answering: A survey," *Artificial Intelligence in Medicine*, p. 102611, 2023.

[55] S. Ravi, A. Chinchure, L. Sigal, R. Liao, and V. Shwartz, "Vlc-bert: visual question answering with contextualized commonsense knowledge," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1155–1165.

[56] Z. Shao, Z. Yu, M. Wang, and J. Yu, "Prompting large language models with answer heuristics for knowledge-based visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 974–14 983.

[57] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[58] V. Gorokhovatskyi, I. Tvoroshenko, O. Kobylin, and N. Vlasenko, "Search for visual objects by request in the form of a cluster representation for the structural image description," *Advances in Electrical and Electronic Engineering*, vol. 21, no. 1, pp. 19–27, 2023.

[59] D. Shah, B. Osiński, S. Levine *et al.*, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Conference on Robot Learning*. PMLR, 2023, pp. 492–504.

[60] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 10 608–10 615.

[61] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*. PMLR, 2023, pp. 287–318.

[62] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[63] Y. Cui, S. Karamcheti, R. Palleti, N. Shivakumar, P. Liang, and D. Sadigh, "No, to the right: Online language corrections for robotic manipulation via shared autonomy," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 93–101.

[64] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv preprint arXiv:2307.05973*, 2023.

[65] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *arXiv preprint arXiv:2207.05608*, 2022.

[66] Z. Zhao, W. S. Lee, and D. Hsu, "Large language models as commonsense knowledge for large-scale task planning," *arXiv preprint arXiv:2305.14078*, 2023.

[67] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[68] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[69] J. A. Abdulsaheb and D. J. Kadhim, "Classical and heuristic approaches for mobile robot path planning: A survey," *Robotics*, vol. 12, no. 4, p. 93, 2023.

[70] P. A. Jansen, "Visually-grounded planning without vision: Language models infer detailed plans from high-level instructions," *arXiv preprint arXiv:2009.14259*, 2020.

[71] W. Huang, F. Xia, D. Shah, D. Driess, A. Zeng, Y. Lu, P. Florence, I. Mordatch, S. Levine, K. Hausman *et al.*, "Grounded decoding: Guiding text generation with grounded models for robot control," *arXiv preprint arXiv:2303.00855*, 2023.

[72] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2motion: From natural language instructions to feasible plans," *arXiv preprint arXiv:2303.12153*, 2023.

[73] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2998–3009.

[74] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," in *Conference on Robot Learning*. PMLR, 2023, pp. 23–72.

[75] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *arXiv preprint arXiv:2307.03109*, 2023.

[76] S. Li, X. Puig, C. Paxton, Y. Du, C. Wang, L. Fan, T. Chen, D.-A. Huang, E. Akyürek, A. Anandkumar *et al.*, "Pre-trained language models for interactive decision-making," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 199–31 212, 2022.

[77] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.

[78] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," *arXiv preprint arXiv:2307.01928*, 2023.

[79] N. Rane, "Transformers in industry 4.0, industry 5.0, and society 5.0: Roles and challenges," 2023.

[80] H. Qiao, Y.-X. Wu, S.-L. Zhong, P.-J. Yin, and J.-H. Chen, "Brain-inspired intelligent robotics: Theoretical analysis and systematic application," *Machine Intelligence Research*, vol. 20, no. 1, pp. 1–18, 2023.

[81] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J. B. Tenenbaum, T. Shu, and C. Gan, "Building cooperative embodied agents modularly with large language models," *arXiv preprint arXiv:2307.02485*, 2023.

[82] Y. Ding, X. Zhang, C. Paxton, and S. Zhang, "Task and motion planning with large language models for object rearrangement," *arXiv preprint arXiv:2303.06247*, 2023.

[83] N. Di Palo, A. Byravan, L. Hasenclever, M. Wulfmeier, N. Heess, and M. Riedmiller, "Towards a unified agent with foundation models," in *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023.

[84] A. Xie, Y. Lee, P. Abbeel, and S. James, "Language-conditioned path planning," in *Conference on Robot Learning*. PMLR, 2023, pp. 3384–3396.

[85] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," *arXiv*, 2022.

[86] Y. Lu, P. Lu, Z. Chen, W. Zhu, X. E. Wang, and W. Y. Wang, "Multimodal procedural planning via dual text-image prompting," *arXiv preprint arXiv:2305.01795*, 2023.

[87] A. Tam, N. Rabinowitz, A. Lampinen, N. A. Roy, S. Chan, D. Strouse, J. Wang, A. Banino, and F. Hill, "Semantic exploration from language abstractions and pretrained representations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 377–25 389, 2022.

[88] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[89] Y. J. Ma, W. Liang, V. Som, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman, "Liv: Language-image representations and rewards for robotic control," *arXiv preprint arXiv:2306.00958*, 2023.

[90] W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever, J. Humplik *et al.*, "Language to rewards for robotic skill synthesis," *arXiv preprint arXiv:2306.08647*, 2023.

[91] H. Hu and D. Sadigh, "Language instructed reinforcement learning for human-ai coordination," *arXiv preprint arXiv:2304.07297*, 2023.

[92] A. Bucker, L. Figueredo, S. Haddadin, A. Kapoor, S. Ma, S. Vemprala, and R. Bonatti, "Latte: Language trajectory transformer," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7287–7294.

[93] R. Wang, J. Mao, J. Hsu, H. Zhao, J. Wu, and Y. Gao, "Programmatically grounded, compositionally generalizable robotic manipulation," *arXiv preprint arXiv:2304.13826*, 2023.

[94] A. Z. Ren, B. Govil, T.-Y. Yang, K. R. Narasimhan, and A. Majumdar, "Leveraging language for accelerated learning of tool manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1531–1541.

[95] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, "Open-vocabulary queryable scene representations for real world planning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 509–11 522.

[96] H. Ha, P. Florence, and S. Song, "Scaling up and distilling down: Language-guided robot skill acquisition," *arXiv preprint arXiv:2307.14535*, 2023.

[97] A. Gupta, L. Fan, S. Ganguli, and L. Fei-Fei, "Metamorph: Learning universal controllers with transformers," *arXiv preprint arXiv:2203.11931*, 2022.

[98] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 523–11 530.

[99] L.-H. Lin, Y. Cui, Y. Hao, F. Xia, and D. Sadigh, "Gesture-informed robot assistance via foundation models," in *Conference on Robot Learning*. PMLR, 2023, pp. 3061–3082.

[100] S. Huang, Z. Jiang, H. Dong, Y. Qiao, P. Gao, and H. Li, "Instruct2act: Mapping multi-modality instructions to robotic actions with large language model," *arXiv preprint arXiv:2305.11176*, 2023.

[101] T. Silver, V. Hariprasad, R. S. Shuttleworth, N. Kumar, T. Lozano-Pérez, and L. P. Kaelbling, "Pddl planning with pretrained large language models," in *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[102] Z. Liu, A. Bahety, and S. Song, "Reflect: Summarizing robot experiences for failure explanation and correction," *arXiv preprint arXiv:2306.15724*, 2023.

[103] X. Zhao, M. Li, C. Weber, M. B. Hafez, and S. Wermter, "Chat with the environment: Interactive multimodal perception using large language models," *arXiv preprint arXiv:2303.08268*, 2023.

[104] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," *arXiv preprint arXiv:2304.03442*, 2023.

[105] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[106] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[107] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and

S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[108] G. Zhou, V. Dean, M. K. Srirama, A. Rajeswaran, J. Pari, K. Hatch, A. Jain, T. Yu, P. Abbeel, L. Pinto, C. Finn, and A. Gupta, "Train offline, test online: A real robot learning benchmark," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[109] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-z: Zero-shot task generalization with robotic imitation learning," in *5th Annual Conference on Robot Learning*, 2021. [Online]. Available: https://openreview.net/forum?id=8kbp23tSGYv

[110] L. Y. Chen, S. Adebola, and K. Goldberg, "Berkeley UR5 demonstration dataset," https://sites.google.com/view/berkeley-ur5/home.

[111] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," *arXiv preprint arXiv:2112.01511*, 2021.

[112] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard, "Latent plans for task agnostic offline reinforcement learning," in *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[113] S. Dass, J. Yapeter, J. Zhang, J. Zhang, K. Pertsch, S. Nikolaidis, and J. J. Lim, "Clvr jaco play dataset," 2023. [Online]. Available: https://github.com/clvrai/clvr_jaco_play_dataset

[114] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Chatgpt empowered long-step robot control in various environments: A case application," *arXiv preprint arXiv:2304.03893*, 2023.

[115] Y. Liu, J. Hou, C. Li, and X. Wang, "Intelligent soft robotic grippers for agricultural and food product handling: A brief review with a focus on design and control," *Advanced Intelligent Systems*, p. 2300233, 2023.

[116] D. Liu, Z. Li, Z. Wu, and C. Li, "Dt/mars-cyclegan: Improved object detection for mars phenotyping robot," *arXiv preprint arXiv:2310.12787*, 2023.

[117] G. Lu, S. Li, G. Mai, J. Sun, D. Zhu, L. Chai, H. Sun, X. Wang, H. Dai, N. Liu *et al.*, "Agi for agriculture," *arXiv preprint arXiv:2304.06136*, 2023.

[118] C. Batailler, A. Fernandez, J. Swan, E. Servien, F. S. Haddad, F. Catani, and S. Lustig, "Mako ct-based robotic arm-assisted system is a reliable procedure for total knee arthroplasty: a systematic review," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 29, pp. 3585–3598, 2021.

[119] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[120] Y. Duan, J. Zhou, Z. Wang, Y.-K. Wang, and C.-T. Lin, "Dewave: Discrete eeg waves encoding for brain dynamics to text translation," *arXiv preprint arXiv:2309.14030*, 2023.

[121] R. Zhang, S. Lee, M. Hwang, A. Hiranaka, C. Wang, W. Ai, J. J. R. Tan, S. Gupta, Y. Hao, G. Levine *et al.*, "Noir: Neural signal operated intelligent robots for everyday activities," *arXiv preprint arXiv:2311.01454*, 2023.

APPENDIX

In this Appendix, the complete prompt (Fig. 4) in our framework and additional experimental results are presented (Fig. 5∼Fig. 7).

SYSTEM

You are an excellent robot task planner. Given a natural language instruction and information about the working environment, you break it down into a sequence of step-by-step instructions and corresponding robot actions.

Predefined Action Pool:
* `move_hand():` Move the robot hand from one position to another with/without grasping an object.
* `grasp_object():` Grab an object.
* `release_object():` Release an object in the robot hand.

Necessary robot actions are defined as above. Note some actions can be done by `move_hand()` like open a drawer by pull back and close it by push forward. If necessary, add new actions to the pool to complete the task.

You generate the task plan and output it as JSON with three keys:
* `JSON{"task_cohesion"}`: A JSON containing information about the robot's actions that have been split up.
* `JSON{"environment_before"}`: The state of the environment before the actions.
* `JSON{"environment_after"}`: The state of the environment after the actions.

Three keys exist in `JSON{"task_cohesion"}`.
* `JSON{"task_cohesion"}{"task_sequence"}`: Contains a list of robot actions. Only the behaviors defined in the "Predefined Action Pool" will be used.
* `JSON{"task_cohesion"}{"step_instructions"}`: Contains a list of step-by-step instructions corresponding to the `JSON{"task_cohesion"}{"task_sequence"}`.
* `JSON{"task_cohesion"}{"object_name"}`: The name of the target object.

You should only return the JSON, without any explanation or notes.

USER

Instruction: <natural language instruction>.

Working Env: as shown in the given image. <image>

ASSISTANT

Task planning response.

USER

Robot task instruction: <natural language instruction>
Task plan generated by GPT-4: <task plan>
Frames sampled from a reference demonstration for the above robot task: <images>

Generate a description of the video.
Score the generated actions on a scale from 0 to 10, based on how well they match the demo video:
a perfect match scores 10.
You should only return a JSON, which has three keys:

* `JSON{"demo_video_description"}`: The description of the reference demonstration video.
* `JSON{"matching_score"}`: The matching score.
* `JSON{"explanation"}`: The explanation of your score.

ASSISTANT

Evaluation response.

Fig. 4. Full prompts.

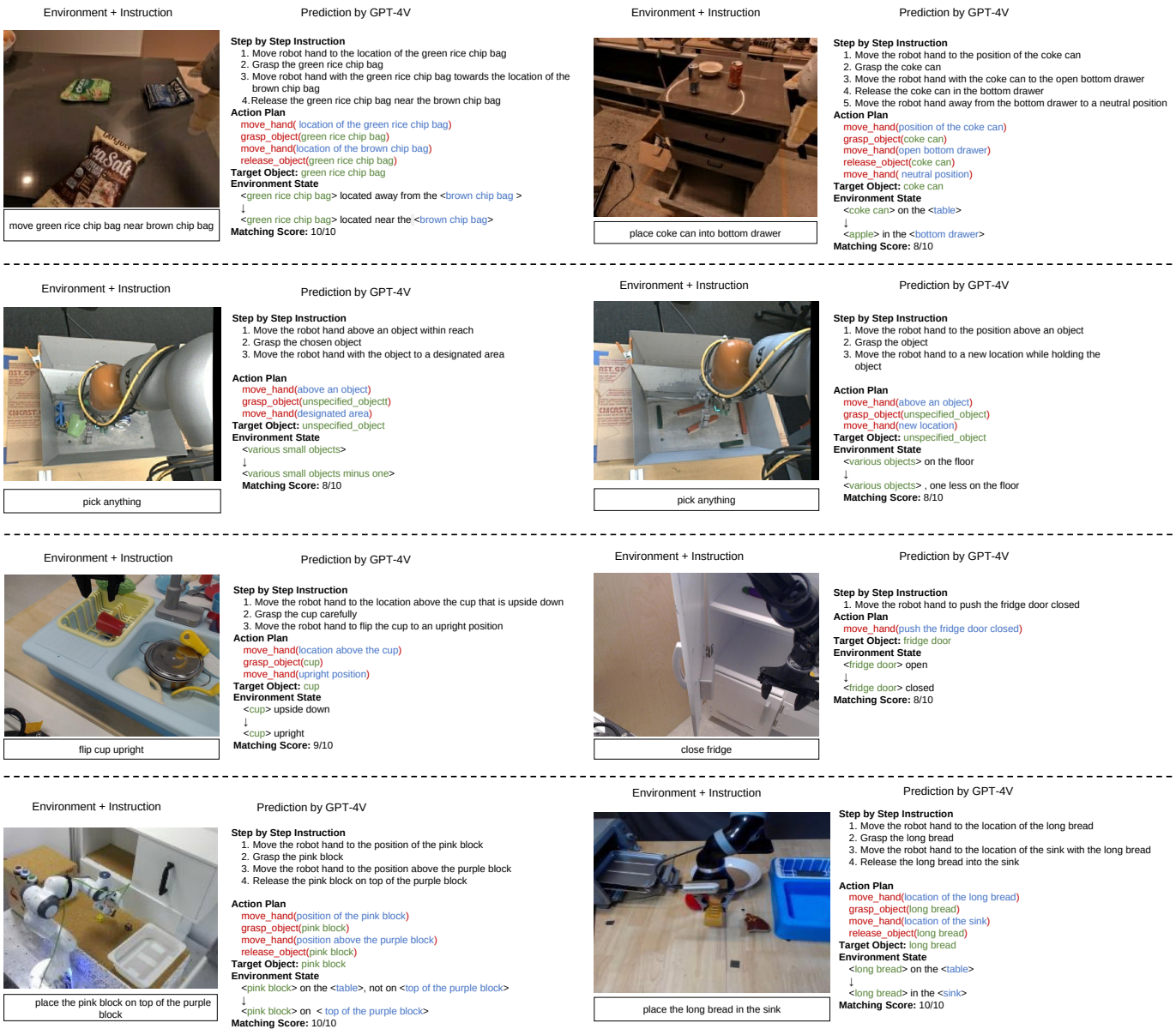
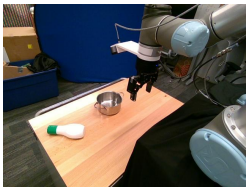


Fig. 5. Generated task plans for different datasets: RT-1 Robot Action (Top Panel), QT-Opt (Second Panel), Berkeley Bridge (Third Panel), Freiburg Franka Play (Bottom Left), and USC Jaco Play (Bottom Right).

Environment + Instruction



put the ranch bottle into the pot

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to the position above the ranch bottle
2. Grasp the ranch bottle
3. Move the robot hand with the grasped ranch bottle to position it above the pot
4. Release the ranch bottle into the pot

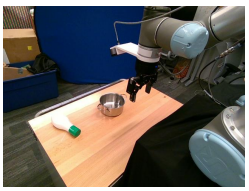
Action Plan
move_hand(move the robot hand to the position above one end of the cable)
grasp_object(grasp the end of the cable)
move_hand(route the cable along the desired path)
release_object(release the cable at the end of the path)

Target Object: ranch bottle

Environment State
<ranch_bottle> <on the table beside the pot><pot> empty<on the table>
↓
<ranch_bottle> <inside the pot><pot> <containing the ranch bottle>

Matching Score: 9/10

Environment + Instruction



put the ranch bottle into the pot

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to the position above the ranch bottle
2. Grasp the ranch bottle
3. Move the robot hand with the grasped ranch bottle to position it above the pot
4. Release the ranch bottle into the pot

Action Plan
move_hand(to the position above the ranch bottle)
grasp_object(ranch bottle)
move_hand(to position above the pot)
release_object(ranch bottle into the pot)

Target Object: ranch bottle

Environment State
<ranch_bottle> <on the table><pot> <on the table>
↓
<ranch_bottle> <in the pot><pot> <on the table with the ranch bottle in it>

Matching Score: 8/10

Environment + Instruction

pour

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to align with the bottle
2. Grasp the bottle
3. Move the robot hand with the bottle above the cup
4. Tilt the bottle to pour the content into the cup and then bring back it to upright position before release

Action Plan
move_hand(align with the bottle)
grasp_object(bottle)
move_hand(above the cup)
release_object(bottle)

Target Object: bottle

Environment State
<bottle> <held by robot arm but not positioned above the cup><cup>
<positioned on the table, empty>
↓
<bottle> <returned to initial position after pouring><cup> <positioned on the table, now contains the poured content>

Matching Score: 9/10

Environment + Instruction

pour

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to align with the bottle
2. Grasp the bottle
3. Move the robot hand with the bottle above the cup
4. Tilt the bottle to pour the content into the cup and then bring back it to upright position before release

Action Plan
move_hand(align with the bottle)
grasp_object(bottle)
move_hand(above the cup)
release_object(bottle)

Target Object: bottle

Environment State
<bottle> <held by robot arm but not positioned above the cup><cup>
<positioned on the table, empty>
↓
<bottle> <returned to initial position after pouring><cup> <positioned on the table, now contains the poured content>

Matching Score: 9/10

Environment + Instruction

drag the ceramic bowl in a circle

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand above the ceramic bowl
2. Grasp the ceramic bowl
3. Move the hand with the bowl in a circular motion to start the circular drag
4. Continue the circular motion to the halfway point
5. Continue the circular motion for three-quarters of the circle
6. Complete the circular dragging motion by finishing the circle
7. Release the ceramic bowl

Action Plan
move_hand(to the bowl)
grasp_object(ceramic bowl)
move_hand(the circular motion)
move_hand(the circular motion)
move_hand(the circular motion)
move_hand(the circular motion)
release_object(ceramic bow)

Target Object: ceramic bowl

Environment State
<ceramic bowl> <stationary on table>
↓
<door> <stationary on table after being moved in a circle>

Matching Score: 10/10

Environment + Instruction

place apple in tray

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to the location of the apple
2. Grasp the apple
3. Move the robot hand with the apple to the tray
4. Release the apple into the tray

Action Plan
move_hand(to the apple)
grasp_object(apple)
move_hand(to the tray)
release_object(into the tray)

Target Object: apple

Environment State
<apple> <on the table, not in the tray><tray> <on the table, empty>
↓
<apple> <in the tray><tray> <on the table, containing the apple>

Matching Score: 8/10

Environment + Instruction

put the ranch bottle into the pot

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand towards the door handle
2. Grasp the door handle
3. Pull the door handle to open the door

Action Plan
move_hand(towards the door handle)
grasp_object(door handle)
move_hand(from the door handle)

Target Object: door handle

Environment State
<door> <closed><robot> <positioned in front of the door>
↓
<door> <open><handle> <still positioned in front of the now open door>

Matching Score: 8/10

Environment + Instruction

pick rxbar chocolate from bottom drawer and place on counter

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to the rxbar chocolate in the bottom drawer
2. Grasp the rxbar chocolate
3. Move the robot hand with the rxbar chocolate to the counter
4. Release the rxbar chocolate onto the counter

Action Plan
move_hand(in the bottom drawer)
grasp_object(rxbar chocolate)
move_hand(the counter)
release_object(rxbar chocolate)

Target Object: rxbar chocolate

Environment State
<bottom_drawer> <closed with rxbar chocolate inside>
<counter> <empty>
↓
<bottom_drawer> <open with no rxbar chocolate>
<counter> <having rxbar chocolate on it>

Matching Score: 9/10

Fig. 6. Generated task plans for different datasets: Berkeley Autolab UR5 (Top Panel), TOTO Benchmar (Second Panel), BC-Z (Third Panel), NYU VINN (Bottom Left). and RT-1 Robot Action (Bottom Right)

Environment + Instruction

close middle drawer

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to the middle drawer and push it closed

Action Plan
move_hand(the middle drawer and push it closed)

Target Object: middle drawer

Environment State
<middle_drawer> <open>
↓
<middle_drawer> <closed>

Matching Score: 9/10

Environment + Instruction

pick banana from white bowl

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to the white bowl on the counter
2. Grasp the banana from the white bowl
3. Move the robot hand away from the bowl with the banana

Action Plan
move_hand(the white bowl on the counter)
grasp_object(banana)
move_hand(away)

Target Object: banana

Environment State
<counter> <white bowl with banana in it>
↓
<counter> <white bowl without the banana>

Matching Score: 9/10

Environment + Instruction

knock pepsi can over

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to the location of the Pepsi can
2. Move the robot hand to slightly push the Pepsi can to knock it over

Action Plan
move_hand(the location of the Pepsi can)
move_hand(lightly push the Pepsi can)

Target Object: Pepsi can

Environment State
<counter> <with a Pepsi can standing upright>
↓
<counter> <with a Pepsi can knocked over>

Matching Score: 10/10

Environment + Instruction

pick brown chip bag from top drawer and place on counter

Prediction by GPT-4V

Step by Step Instruction
1. Move the robot hand to the brown chip bag in the top drawer
2. Grasp the brown chip bag
3. Move the robot hand with the brown chip bag to the counter
4. Release the brown chip bag onto the counter

Action Plan
move_hand(in the top drawer)
grasp_object(brown chip bag)
move_hand(the counter)
release_object(brown chip bag)

Target Object: brown chip bag

Environment State
<top_drawer> <closed with brown chip bag inside>
<counter> <occupied with a bowl of bananas>
↓
<top_drawer> <open with no brown chip bag>
<counter> <occupied with a bowl of bananas and a brown chip bag on it>

Matching Score: 10/10

Fig. 7. Generated task plans for different datasets: RT-1 Robot Action.