Look Before You Leap: Unveiling the Power of GPT-4V in Robotic Vision-Language Planning

Yingdong Hu[1,2,3*], Fanqi Lin[1,2,3*], Tong Zhang[1,2,3], Li Yi[1,2,3], Yang Gao[1,2,3†]
[1]Tsinghua University [2]Shanghai Artificial Intelligence Laboratory [3]Shanghai Qi Zhi Institute
{huyd21, lfq20, zhangton20}@mails.tsinghua.edu.cn, {ericyi, gaoyangiiis}@mail.tsinghua.edu.cn
robot-vila.github.io

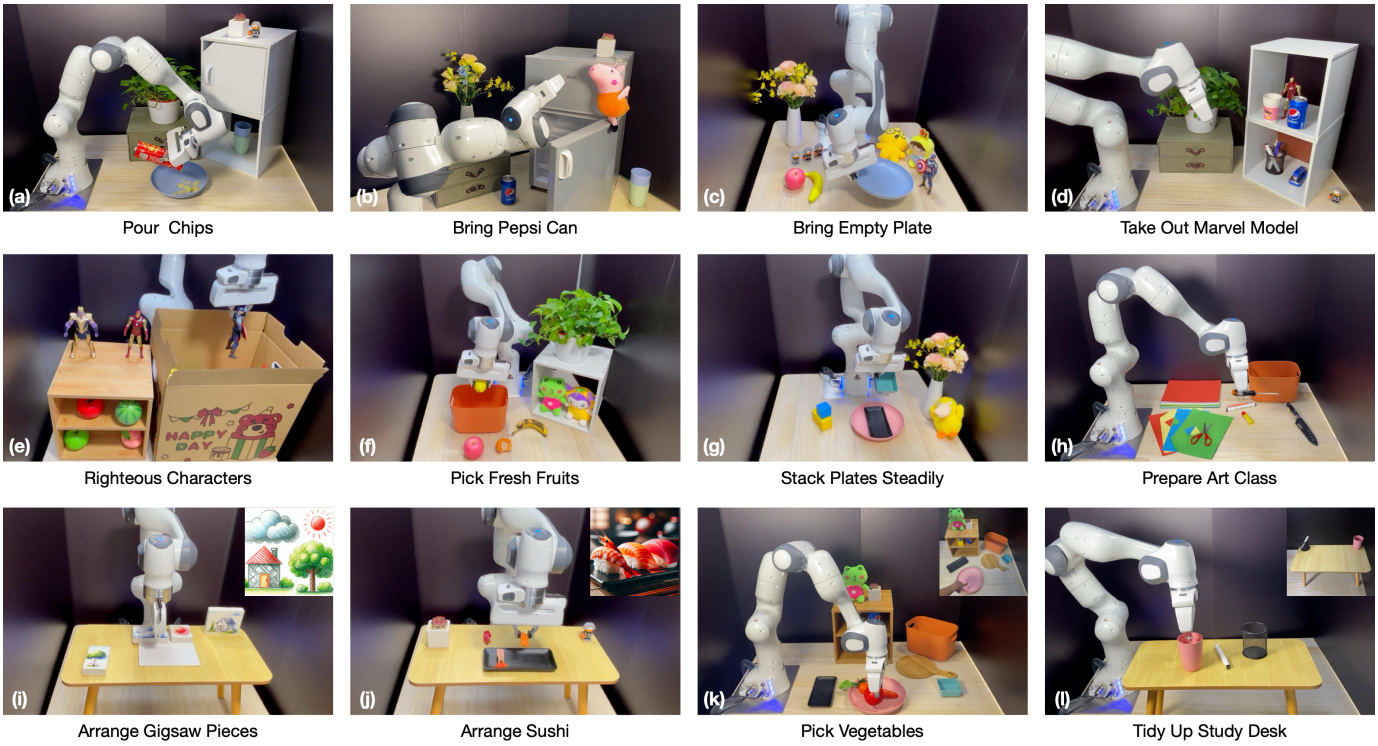

Fig. 1: We present **VILA**, a simple and effective method for long-horizon robotic task planning. By integrating vision directly into the reasoning process, VILA can leverage the wealth of commonsense knowledge grounded in the visual world. This results in remarkable performance in tasks that demand an understanding of spatial layouts (top row), object attributes (middle row), and tasks with multimodal goals (bottom row).

***Abstract*—In this study, we are interested in imbuing robots with the capability of physically-grounded task planning. Recent advancements have shown that large language models (LLMs) possess extensive knowledge useful in robotic tasks, especially in reasoning and planning. However, LLMs are constrained by their lack of world grounding and dependence on external affordance models to perceive environmental information, which cannot jointly reason with LLMs. We argue that a task planner should be an inherently grounded, unified multimodal system. To this end, we introduce Robotic Vision-Language Planning (VILA), a novel approach for long-horizon robotic planning that leverages vision-language models (VLMs) to generate a sequence of actionable steps. VILA directly integrates perceptual data into its reasoning and planning process, enabling a profound understanding of commonsense knowledge in the visual world, including spatial layouts and object attributes. It also supports flexible multimodal goal specification and naturally incorporates visual feedback. Our extensive evaluation, conducted in both real-robot and simulated environments, demonstrates VILA's superiority over existing LLM-based planners, highlighting its effectiveness in a wide array of open-world manipulation tasks.**

I. INTRODUCTION

Scene-aware task planning is a pivotal facet of human intelligence [83, 75]. When presented with a simple language instruction, humans demonstrate a spectrum of complex behaviors depending on the context. Take the instruction "get a can of coke," for example. If a coke can is visible, a person will immediately pick it up. If not, they will search locations like the refrigerator or storage cabinets. This adaptability reflects humans' deep understanding of the scene and exten-

* The first two authors contributed equally.
† Correspondence to: Yang Gao <gaoyangiiis@tsinghua.edu.cn>.

sive common sense, enabling them to interpret instructions contextually. In this paper, we explore how we can create an embodied agent, such as a robot, that emulates this human-like adaptability and exhibits long-horizon task planning in varying scenes.

In recent years, large language models (LLMs) [9, 62, 13, 6] have showcased their remarkable capabilities in encoding extensive semantic knowledge about the world [65, 42, 29]. This has sparked a growing interest in leveraging LLMs for generating step-by-step plans for complex, long-horizon tasks [2, 37, 38]. However, a critical limitation of LLMs is their lack of world grounding — they cannot perceive and reason about the physical state of robots and their environments, including object shapes, physical properties, and real-world constraints.

To overcome this challenge, a prevalent approach involves employing external affordance models [27], such as open-vocab detectors [57] and value functions [2], to provide real-world grounding for LLMs [2, 40]. However, these modules often fail to convey the truly necessary task-dependent information in complex environments, as they serve as one-directional channels transmitting perceptual information to LLMs. In this scenario, the LLM is like a blind person, while the affordance model serves as a sighted guide. On the one hand, the blind person relies solely on their imagination and the guide's limited narrative to comprehend the world; on the other hand, the sighted guide may not accurately comprehend the blind person's purpose. This combination often leads to unfeasible or unsafe action plans in the absence of precise, task-relevant visual information. For instance, a robot tasked with taking out a Marvel model from a shelf (see Figure 1 (d)) may overlook obstacles like the paper cup and coke can, leading to collisions. Consider another example of preparing art class (see Figure 1 (h)), scissors can be perceived as sharp and hazardous objects, or as essential tools for handicrafts. This distinction is challenging for the vision module due to the lack of specific task information. These examples highlight the limitations of LLM-based planners in capturing intricate spatial layouts and fine-grained object attributes, underscoring the necessity for active joint reasoning between vision and language.

The recent advancements in vision-language models (VLMs), exemplified by GPT-4V(ision) [61, 88], have significantly broadened the horizons of research. VLMs synergize perception and language processing into a unified system, enabling direct incorporation of perceptual information into the language model's reasoning [53, 14, 5, 97]. Building upon these developments, we introduce Robotic **Vi**sion-**La**nguage Planning (**ViLa**) — a simple, effective, and scalable method for long-horizon robotic planning. ViLa distinguishes itself from previous LLM-based planning methods by eschewing independent affordance models and instead directly prompting VLMs to generate a sequence of actionable steps based on visual observations of the environment and high-level language instructions. ViLa exhibits the following key properties absent in LLM-based planning methods:

- **Profound Understanding of Commonsense Knowledge Grounded in the Visual World.** ViLa excels in complex tasks that demand an understanding of spatial layouts (e.g., `Take Out Marvel Model`) or object attributes (e.g., `Stack Plates Steadily`). This kind of commonsense knowledge pervades nearly every task of interest in robotics, but previous LLM-based planners consistently fall short in this regard.
- **Versatile Goal Specifiaction.** ViLa supports flexible multimodal goal specification approaches. It is capable of utilizing not just language instrctions but also diverse forms of goal images, and even a blend of both language and images, to define objectives effectively.
- **Visual Feedback.** ViLa effectively utilizes visual feedback in an intuitive and natural way, enabling robust closed-loop planning in dynamic environments.

We conduct a systematic evaluation of ViLa across 16 real-world, everyday manipulation tasks, which involve a diverse range of open-set instructions and objects. ViLa consistently outperforms LLM-based planners, such as SayCan [2] and Grounded Decoding [40], by a significant margin. To facilitate a more exhaustive and rigorous comparison, we extend our evaluation to include 16 simulated tasks based on the RAVENS environment [93], wherein ViLa continues to show marked enhancements. All these outcomes provide compelling evidence that ViLa possesses the potential to serve as a universal task planning method for general-purpose robotic systems.

II. Related Work

Vision-Language Models. The striking advancements made by scaling up large language models (LLMs) [9, 62, 13, 79, 31] have sparked a surge of interest in similarly expanding large vision-language models (VLMs) [23, 61, 88]. The prevalent approach to construct VLMs involves employing a cross-modal connector to align the features of pre-trained visual encoders with the input embedding space of the LLMs [3, 53, 52, 10, 47, 36, 90, 97, 5, 81]. The ability of VLMs to understand both images and text renders them highly adaptable for a range of applications, including visual question answering [4, 92], image captioning [1, 34], and optical character recognition [48]. In contrast to these uses, our study takes a different path. We concentrate on harnessing the rich world knowledge and the visually grounded attribute of VLMs to address complex long-horizon planning challenges in robotics.

Pre-Trained Foundation Models For Robotics. Recent advancements in applying large pre-trained foundation models to robotics can be classified into three categories:

(1) Pre-Trained Vision Models: A wealth of prior approaches employ vision models pre-trained on large-scale image datasets [28, 15] to generate visual representations for visuomotor control tasks [64, 85, 67, 59, 55, 95]. Nonetheless, a robotic system encompasses more than just a perception module; it includes a control policy as well. Relying solely on visual representations that capture high-level semantics may not ensure the control policy's generalizability or the system's overall effectiveness [35, 91, 30].

(2) Pre-Trained Language Models: Another research avenue explores the use of large language models (LLMs) for robotic tasks, particularly in reasoning and planning [37, 2, 73, 84, 74, 51, 80, 50, 16, 68]. However, to ground these language models in physical environments, auxiliary modules such as affordance models [40], perception APIs [49], and textual scene descriptions [38, 94] are essential. In contrast, our work emphasizes generating plans without depending on these auxiliary models for grounding. This approach allows for the seamless integration of perceptual information directly into the reasoning and planning process.

(3) Pre-Trained Vision-Language Models: Numerous studies have explored the application of vision-language models (VLMs) in robotics [39, 19, 70, 96, 24]. Notably, RT-2 [8] demonstrates the integration of VLMs in low-level robotic control. In contrast, our research is primarily centered on high-level robotic planning. Although PaLM-E [18] shares similarities with our approach, it necessitates training on a substantial mixture of robotics and general visual-language data [11, 54]. This approach implies that introducing a robot to a new environment necessitates the collection of additional data and subsequent retraining of the model. In stark contrast, our VILA stands out as an open-world, zero-shot model. It is capable of performing a broad spectrum of everyday manipulation tasks without additional training data and in-context examples in the prompt.

Task and Motion Planning. Task and Motion Planning (TAMP) [43, 26] stands as a critical framework in solving long-horizon planning tasks, integrating low-level continuous motion planning [46] with high-level discrete task planning [22, 69, 60]. While traditional research in this domain has predominantly centered on symbolic planning [22, 60] or optimization-based methods [77, 78], the advent of machine learning [87, 20, 25, 41] and LLMs [16, 12, 72, 86] is revolutionizing this arena. In our work, we leverage VLMs to comprehend the robot environment and interpret high-level instructions. By incorporating commonsense knowledge that is intrinsically grounded in the visual world, our approach excels in handling complex tasks beyond the reach of previous LLM-based planning methods.

III. METHOD

We first provide the formulation of the planning problem in Sec. III-A. Subsequently, we present how VILA utilize vision-language models as robot planners (Sec. III-B). Finally, we describe unique properties of VILA that contribute to its advantages (Sec. III-C).

A. Problem Statement

Our robotic system takes a visual observation $\mathbf{x}_t$ of the environment and a high-level language instruction $\mathcal{L}$ (e.g. "*stack these containers of different colors steadily*") that describes a manipulation task. We assume that the visual observation $\mathbf{x}_t$ serves as an accurate representation of world state. The language instruction $\mathcal{L}$ can be arbitrarily long-horizon or under-specified (i.e., requires contextual understanding). The central problem investigated in this work is to generate a sequence of text actions, represented as $\ell_1, \ell_2, \cdots, \ell_T$. Each text action ℓ_t is a short-horizon language instruction (e.g. "*pick up blue container*") that specifies a sub-task/primitive skill $\pi_{\ell_t} \in \Pi$. Note that our contributions do not focus on the acquisition of these skills Π; rather, we assume that all the necessary skills are already available. These skills can take the form of predefined script policies or may have been acquired through various learning methods, including reinforcement learning (RL) [76] and behavior cloning (BC) [66].

Algorithm 1 VILA

Require: Initial visual observation $\mathbf{x}_1$, a high level instruction $\mathcal{L}$ and a set of skills Π.

$t = 1$, $\ell_1 = \emptyset$
while $\ell_{t-1} \neq$ "done" **do**
  $p_{1:N} = \text{VLM}(\mathbf{x}_t, \mathcal{L}, \ell_1, ..., \ell_{t-1})$ ▷ Get plan steps
  $\ell_t = p_1$ ▷ Select the first step
  Execute skill $\pi_{\ell_t}(\mathbf{x}_t)$, updating observation $\mathbf{x}_{t+1}$
  $t = t + 1$
end while

B. Vision-Language Models as Robot Planners

To generate feasible plans, high-level robot planning must be grounded in the physical world. While LLMs possess a wealth of structured world knowledge, their exclusive reliance on language input necessitates external components, such as affordance models, to complete the grounding process. However, these external affordance models (e.g., value functions of RL policies [2, 44], object detection models [57], and action detection models [71]) are manually designed as independent channels, operating separately from LLMs, rather than being integrated into an end-to-end system. Moreover, their role is solely transmitting high-dimensional visual perceptual information to LLMs, lacking the capability for joint reasoning. This separation of vision and language modalities results in the vision module's inability to provide comprehensive, task-relevant visual information, thereby hindering the LLM from planning based on accurate task-related visual insights.

Recent advances in vision-language models (VLMs) offer a solution. VLMs demonstrate unprecedented ability in understanding and reasoning across both images and language [53, 14, 5, 97]. Crucially, the extensive world knowledge encapsulated in VLMs is inherently grounded in the visual data they process. Therefore, we advocate for directly employing VLMs that synergizes vision and language capabilities to decompose a high-level instruction into a sequence of low-level skills.

We refer to our method as Robotic **Vi**sion-**La**nguage Planning (**VILA**). Concretely, given current visual observation $\mathbf{x}_t$ of environment and a high-level language goal $\mathcal{L}$, VILA operates by prompting the VLMs to yield a step-by-step plan $p_{1:N}$. We enable closed-loop execution by selecting the first step as the text action $\ell_t = p_1$. Once the text action ℓ_t is selected, the corresponding policy π_{ℓ_t} is executed by the robot

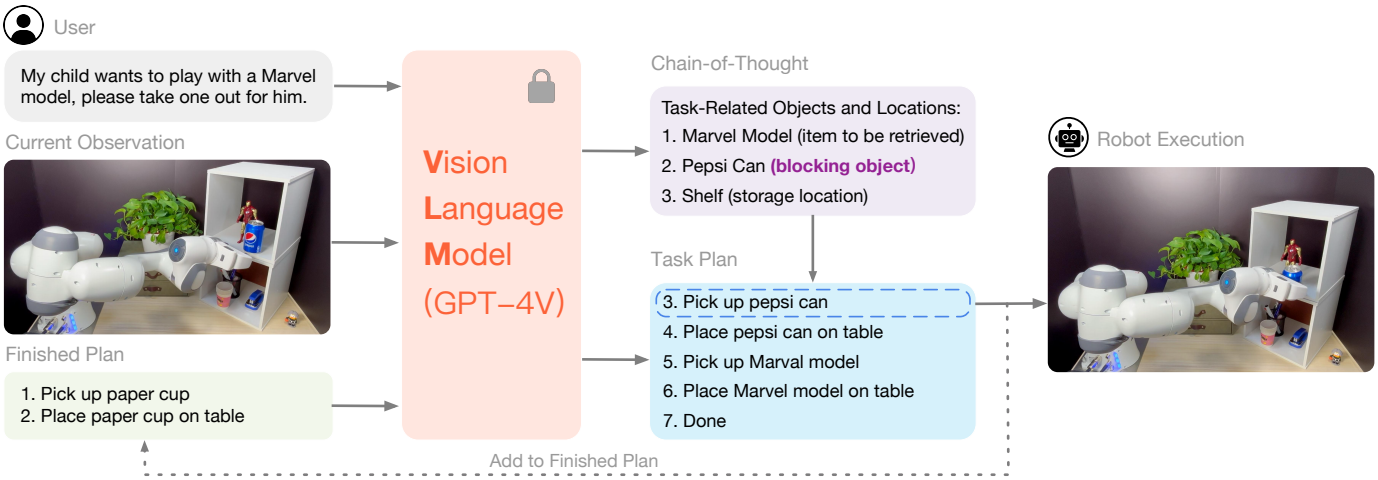

Fig. 2: **Overview of VILA.** Given a language instruction and current visual observation, we leverage a VLM to comprehend the environment scene through chain-of-though reasoning, subsequently generating a step-by-step plan. The first step of this plan is then executed by a primitive policy. Finally, the text action that has been executed is added to the finished plan, enabling a closed-loop planning method in dynamic environments.

and the VLM query is amended to include ℓ_t and the process is run again until a termination token (e.g., "done") is reached. The entire process is shown in Figure 2 and described in Algorithm 1.

In our study, we utilizes GPT-4V(ision) [61, 88] as the VLM. GPT-4V, trained on vast internet-scale data, exhibits exceptional versatilities and extremely strong generalization capabilities. These attributes make it particularly adept at handling open-world scenarios presented in our paper. Furthermore, we find that VILA, powered by GPT-4V, is capable of solving a variety of challenging planning problems, even when operating in a *zero-shot* mode (i.e., without requiring any in-context examples). This significantly reduces the prompt engineering efforts required in previous approaches [2, 37, 40].

C. Intriguing Properties of VILA

In this section, we delve deeper into VILA, shedding light on its advantages and differentiations from previous planning methods.

Comprehension of Common Sense in the Visual World. Previous studies primarily focus on leveraging the knowledge of LLMs for high-level planning [2, 37], centering on language while often overlooking the crucial role of vision. Images and languages, as distinct types of signals, offer unique nature: languages are human-generated and semantically rich, yet they are limited in their ability to represent comprehensive information. In contrast, images are natural signals imbued with low-level fine-grained features, a single image can capture the entirety of a scene's information. This disparity is especially pertinent when the complex environment is challenging to encapsulate in simple language. Directly integrating images into the reasoning and planning process, such as in the case of VILA, allows for a more intuitive understanding of commonsense knowledge grounded in the physical world. Specifically, this understanding manifests in two key aspects:

1) Spatial Layout Understanding: Describing complex geometric configurations, particularly spatial localization, object relationships, and environmental constraints, can be challenging with just simple language. Consider a cluttered scene where object A obscures object B. To reach object B, one must first reposition object A. Relying solely on verbal language descriptions to convey these nuanced relationships between objects is inadequate. Moreover, consider a situation where the desired object is inside a container (like a cabinet or refrigerator). In that case, if an external affordance model (like object detection model) is utilized, since the desired object is not visible, the affordance model would predict a zero probability of successful retrieval, leading to task failure. However, by directly incorporating vision into the reasoning process, VILA can deduce that the sought object, hidden from view, is likely inside the container. This realization necessitates opening the container as a preliminary step to accomplish the task.

2) Object Attribute Understanding: An object is defined by multiple attributes, including its shape, color, material, function, etc. However, the expressive capacity of natural language is limited, making it a somewhat cumbersome medium for conveying these attributes comprehensively. Furthermore, note that an object's attributes is intricately tied to the specific tasks at hand. For example, scissors might be deemed hazardous for children, but they become essential tools during a paper-cutting art class. Previous approaches employ a standalone affordance model to identify object attributes, but this method can only convey a limited subset of attributes in a unidirectional manner. Therefore, active joint reasoning between image and language emerges as a crucial necessity when our tasks demand a thorough understanding of an object's attributes.

Versatile Goal Specification. In many complex, long-term tasks, using a goal image to represent the desired outcome is

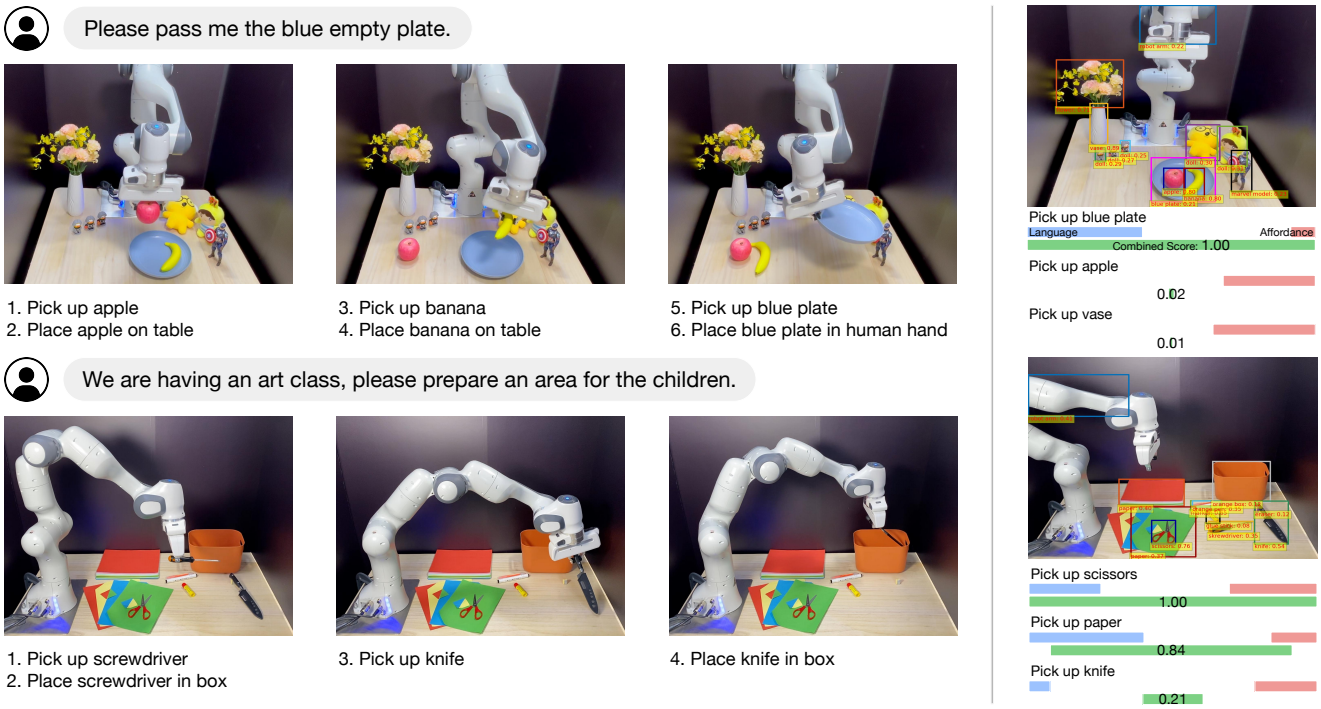

Fig. 3: **Illustration of the execution of ViLa (left) and the decision-making process of SayCan (right)**. In the `Bring Empty Plate` task, the robot must first relocate the apple and banana from the blue plate. However, SayCan's initial step is to directly pick up the blue plate. In the `Prepare Art Class` task, while the scissor is supposed to remain on the table, SayCan erroneously picks up the scissor and places it in a box.

often more effective than relying solely on verbal instructions. For example, to direct a robot to tidy a desk, providing a photo of the desk arranged as desired can be more efficient. Likewise, for food plating tasks, a robot can replicate the arrangement from an image. Such tasks, previously unattainable with LLM-based planning methods, are now remarkably straightforward with ViLa. Specifically, ViLa can not only accepts current visual observation $\mathbf{x}_n$ and language instructions $\mathcal{L}$ as inputs but also incorporates a goal image $\mathbf{x}_g$. This feature sets it apart from many existing goal-conditioned RL/IL algorithms [58, 21, 17], as it does not require the goal and visual observation images to originate from the same domain. The goal image merely needs to convey the essential elements of the task, offering flexibility in its form – it could range from an internet photo to a child's drawing, or even an image showing a target location indicated by a pointing finger. This versatility greatly enhances the system's practicality. Additionally, the ability to combine images and language in describing task goals introduces an additional layer of flexibility and diversity in our goal specification approach.

Visual Feedback. The embodied environments are inherently dynamic, making closed-loop feedback essential for robots. In an effort to incorporate environment feedback into planning methods that rely solely on LLMs, Huang et al. [38] investigate converting all feedback to natural language. However, this approach proves to be cumbersome and ineffective because most of the feedback is initially observed visually. Converting visual feedback into language not only adds complexity to the system but also risks losing valuable information. We believe that providing visual feedback directly is a more intuitive and natural approach, as demonstrated in ViLa. Within ViLa, the VLM serves both as a scene descriptor to recognize object states and as a success detector to determine if the environment satisfies the success conditions defined by the instructions. By reasoning over visual feedback, ViLa enables robots to make corrections or replan in response to changes in the environment or when a skill fails.

IV. Experiments and Analysis

In this section, we first carry out extensive experiments in a real-world system to evaluate ViLa's capability in planning everyday manipulation tasks (Sec. IV-A). Subsequently, we conduct a detailed quantitative comparison of ViLa against baseline methods within a simulated tabletop environment (Sec. IV-B).

A. Real-World Manipulation Tasks

Experimental Setup.

1) Hardware: We set up a real-world tabletop environment. We use a Franka Emika Panda robot (a 7-DoF arm) and a 1-DoF parallel jaw gripper. For perception, we use a Logitech Brio color camera mounted on a tripod, at an angle, pointing towards the tabletop. To ensure consistency in our experiments, we maintain a fixed camera view for all tasks, but for visual aesthetics, we record video demos at different views.

Task	SayCan	GD	ViLa
Pour Chips	20%	40%	80%
Bring Pepsi Can	40%	30%	90%
Bring Empty Plate	0%	0%	100%
Take Out Marvel Model	0%	10%	70%
Righteous Characters	0%	10%	80%
Pick Fresh Fruits	20%	30%	80%
Stack Plates Steadily	20%	10%	70%
Prepare Art Class	0%	30%	70%
Total	13%	20%	80%

TABLE I: **Quantitative evaluation results in tasks requiring rich commonsense knowledge.** ViLa demonstrates superior performance in tasks necessitating a understanding of spatial layouts (top half) and object attributes (bottom half).

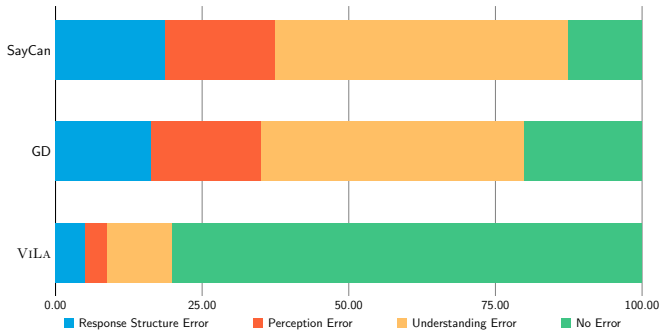

Fig. 4: **Error breakdown of ViLa and baselines.** By leveraging commonsense knowledge grounded in the visual world, ViLa significantly reduces understanding error.

2) Tasks and Evaluation: We design 16 long-horizon manipulation tasks to assess ViLa's performance in three domains: comprehension of commonsense knowledge in the visual world (8 tasks), flexibility in goal specification (4 tasks), and utilization of visual feedback (4 tasks). Figure 1 illustrates a selection of 12 tasks, drawn from the first two domains. For each task, we evaluate all methods across the 10 different variations of the environment, including changes in scene configuration and lighting conditions, etc. For comprehensive details of each task, please see the Appendix.

3) VLM and Prompting: We use GPT-4V from OpenAI API as our VLM. Unlike previous approaches [2, 40], we do not include any in-context examples in the prompt, but only use high-level language instructions and some simple constraints that the robot needs to meet (i.e., strict *zero-shot*). The full prompt is shown in the Appendix.

4) Primitive Skills: We use five categories of primitive skills that lend themselves to complex behaviors through composition and planning. These include "pick up `object`", "place `object` in/on `object`", "open `object`", "close `object`", and "pour `object` into/onto `object`". We concentrate on high-level, temporally extended planning rather than acquiring low-level primitive skills, which is orthogonal to our study. Therefore, we employ script policies as the primitive skills for both the baselines and ViLa. Additional details of low-level primitive skills are in the Appendix.

5) Baselines: We compare with SayCan [2] and Grounded Decoding (GD) [40], which both ground LLMs with external affordance models. Implementing these baselines necessitates accessing output token probabilities from LLMs. However, since OpenAI API currently does not return these probabilities, we employ the open-source Llama 2 70B [79] as an alternative. For the affordance models, we utilize the open-vocabulary detector OWL-ViT [57, 56], following Huang et al [40].

ViLa can understand commonsense knowledge in the visual world. In Table I, we compare the planning success rates on tasks that require understanding of spatial layouts and object attributes (see Figure 1 (a-h) for illustrations of the tasks). ViLa stands out with an average success rate of 80% across 8 tasks, significantly surpassing the performances of SayCan and GD, which achieve success rates of only 13% and 20%, respectively. Particularly in intricate and challenging tasks such as `Take Out Marvel Model` (it's crucial to avoid the cup and coke can) and `Righteous Characters` [1], SayCan and GD's success rates are close to *zero*. These tasks all necessitate the integration of images into the reasoning and planning processes and a deep understanding of commonsense knowledge in the visual world. Furthermore, the tasks outlined in Table I are representative of typical real-world scenarios and are not specifically tailored for ViLa. The across-the-board exceptional performance of ViLa not only highlights its superior generalizability but also underscores its potential as a universal planner for open-world tasks.

Figure 3 shows two environment rollouts comparing ViLa with SayCan. In the first `Bring Empty Plate` task, ViLa identifies the need to relocate the apple and banana from the blue plate before picking it up. In contrast, SayCan recognizes the items (apple, banana, blue plate) but lacks awareness of their spatial relationship, leading it to attempt picking up the blue plate directly. This highlights the significance of comprehending complex geometric configurations and environmental constraints visually. In another scenario involving the preparation of a safe area for a children's art class (`Prepare Art Class`), ViLa discerns that only the screwdriver and fruit knife are hazardous, sparing the scissors necessary for the class, based on the contextual clue of paper cuttings on the table. However, SayCan misclassifies the scissors as dangerous, showing that a comprehensive, global visual understanding is crucial to accurately assess object attributes. The videos of experiment rollouts can be found on the project website: robot-vila.github.io.

In Figure 4, we present a failure breakdown analysis. "Response structure error" here refers to errors of LLMs and VLMs in generating plan steps that fall outside our predefined

[1]Choose righteous characters from three Marvel models, while referring to the model only by its color. Details of this task are in Appendix.

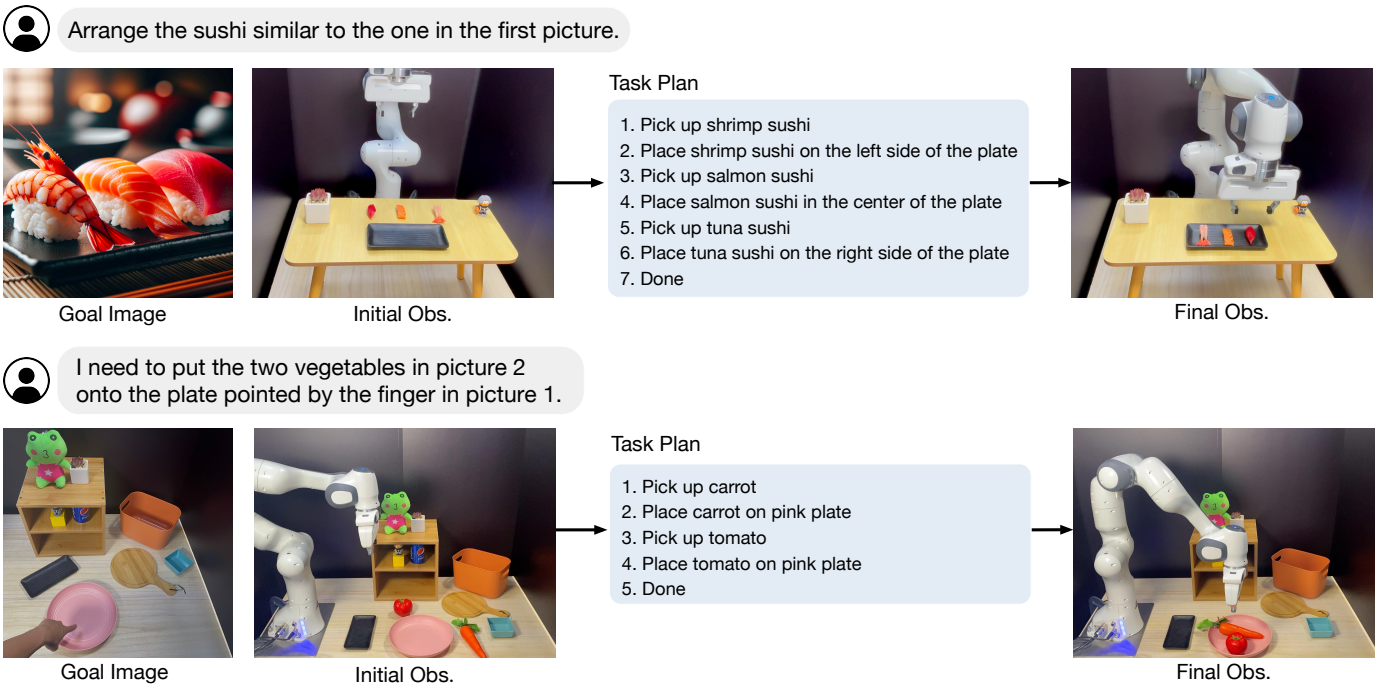

Fig. 5: **Illustration of the execution of ViLa on image goal-conditioned tasks.** In the `Arrange Sushi` task, ViLa generates a plan to arrange sushi based on a reference image. In the `Pick Vegetables` task, the scenario involves a table set with a pink plate, a black sushi plate, a pizza plate, and a green snack plate. Here, ViLa deduces from pointing finger in the goal image that the vegetables should be placed on the pink plate.

Task	Goal Type	Succ. %
Arrange Sushi	Real Image	80%
Arrange Gigsaw Pieces	Drawing	100%
Pick Vegetables	Pointing Finger	100%
Tidy Up Study Desk	Image + Language	60%

TABLE II: **Quantitative evaluation results of ViLa in tasks featuring multimodal goals.**

set of primitive skills. In the case of baselines, "perception error" denotes failures within the open-vocab detector [56]. While VLMs lack a separate perception module, their output, as observed in the chain-of-thought process [82], occasionally fails to recognize some objects. The dominant error in baseline models is "understanding error", which involves errors in understanding the complex spatial layouts and object attributes in the physical world, such as occlusions and context-specific attributes. ViLa significantly reduces the "understanding error" by seamlessly integrating vision and language reasoning, thereby resulting in the lowest overall error. Furthermore, we suggest that careful prompt engineering (i.e., providing examples in the prompt) [9, 63] could steer VLM outputs towards admissible primitive skills, thereby reducing "response structure error".

ViLa supports flexible multimodal goal specification. We introduce a suite of 4 tasks, each with distinct goal types, as illustrated in Figure 1 (i-l). The quantitative results are shown in Table II, where ViLa demonstrates strong capabilities across all tasks. Utilizing the internet-scale knowledge imbued in GPT-4V, ViLa exhibits the remarkable ability to understand a variety of goal images. This includes interpreting vibrant children's drawings for puzzle completion, preparing a sushi platter by referencing a photograph of the dish (illustrated in Figure 5 top row), and even accurately identifying the intended arrangement of vegetables as indicated by a human finger (refer to Figure 5 bottom row). Additionally, we explore goal specification through a combination of image and language instructions. For instance, in the `Tidy Up Study Desk` task, we not only provide an image of a neatly organized desk as the target but also verbally direct the swapping of two specific objects on the desk. Leveraging its dual-capacity in vision and language reasoning, ViLa consistently achieves success in this task as well.

ViLa can leverage visual feedback naturally. We design 4 tasks that require real-time visual feedback for successful execution. In the `Stack Blocks` task, we inject Gaussian noise into the joint position controller, which increases the likelihood of failure in the primitive policy. For the `Pack Chip Bags` task, task progress is reverted by an experimenter who takes out previously packed chip bags from the box. In the `Find Stapler` task, the stapler's location varies among three potential places: the top drawer, the bottom drawer, or the

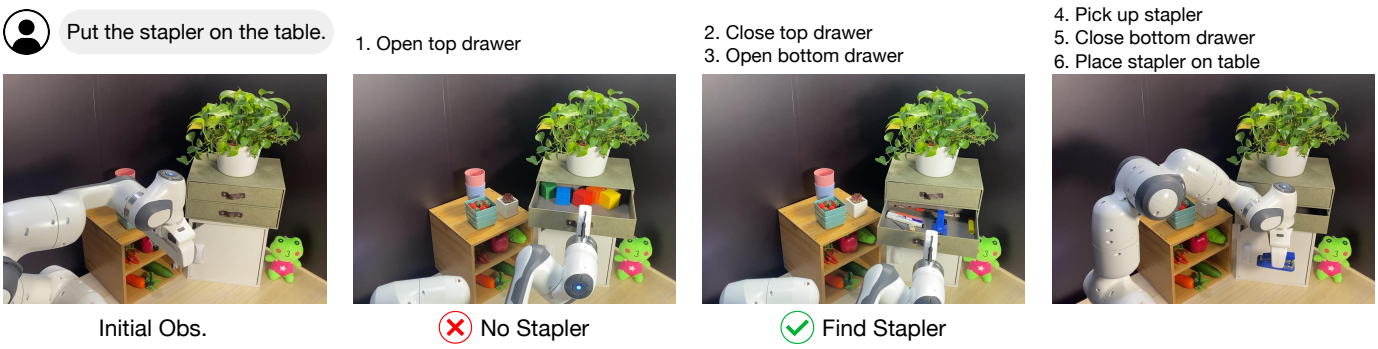

Fig. 6: **Illustration of the execution of ViLa on the `Find Stapler` task.** By incorporating visual feedback and replanning at every step, ViLa is able to continue exploring the bottom drawer when it does not find the stapler in the top drawer, thereby successfully locating the stapler.

Task	Open-Loop	w/ Feedback
Stack Blocks	20%	90%
Pack Chip Bags	0%	100%
Find Stapler	30%	90%
Human-Robot Interaction	20%	80%

TABLE III: **Open-loop ViLa vs. closed-loop ViLa.** By leveraging visual feedback, closed-loop ViLa substantially outperforms the open-loop variant.

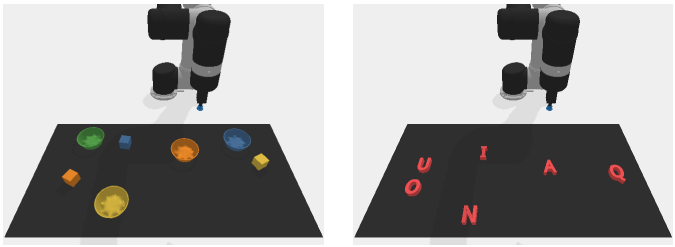

Fig. 7: **Simulated environment based on RAVENS.** We design 16 distinct tasks, which are grouped into two categories: `Blocks & Bowls` (left) and `Letters` (right).

cabinet. The `Human-Robot Interaction` task requires the robot to pause until a person retrieves the cola it has picked up. We evaluate the performance of ViLa against an open-loop variant that formulates a plan based solely on the initial observation. The quantitative results, presented in Table III, reveal that the open-loop variant struggles with these dynamic tasks that demand continuous replanning, while the closed-loop ViLa significantly outperforms it. ViLa is not only able to effectively recover from external disturbances but can also adapt its strategy based on real-time visual observations. A case in point, depicted in Figure 6, is when ViLa, not finding the stapler in the top drawer, proceeds to check the bottom drawer, successfully locates the stapler, and completes the task.

	CLIPort		LLM		GD	ViLa
Tasks	Short	Long	Llama 2	GPT-4		
Seen Tasks						
Blocks & Bowls	3.3%	68.3%	1.7%	0%	18.3%	78.3%
Letters	0%	40.0%	25.0%	25.0%	51.7%	88.3%
Unseen Tasks						
Blocks & Bowls	6.0%	6.0%	20.0%	22.0%	23.0%	81.0%
Letters	1.0%	0%	15.0%	15.0%	42.0%	82.0%

TABLE IV: **Average success rate in simulated environment.** See the Appendix for a detailed breakdown. ViLa consistently outperforms baselines across seen and unseen tasks.

B. Simulated Tabletop Rearrangement

Experimental Setup. We conduct experiments on simulated tabletop rearrangement tasks to provide a more rigorous and fair comparison with baseline methods. Following the setting in Grounded Decoding [40], we develop 16 tasks based on the RAVENS environment [93]. These tasks are categorized into two groups: a seen group, consisting of 6 tasks used for few-shot prompting or as training for supervised baselines, and an unseen group of 10 tasks. Each task requires a UR5 robot to rearrange the objects on the table in some desired configuration, specified by high-level language instructions. The tasks are further classified into two types (see Figure 7): *(i)* `Blocks & Bowls` (8 tasks), which focus on rearranging or combining blocks and bowls (e.g., "put all the blocks in the bowls with matching colors"). *(ii)* `Letters` (8 tasks), which involve rearranging alphabetical letters (e.g., "put the letters on the tables in alphabetical order").

Our comparison encompasses three baseline categories: *(i)* CLIPort, a language-conditioned imitation learning agent that directly take in the high-level language instructions without a planner. We consider two variants: "Short", trained on single-step pick-and-place instructions, and "Long", trained on high-level instructions. *(ii)* An LLM-based planner that does not

relay on any grounding/affordance model. We evaluate Llama 2 and GPT-4. *(iii)* Grounded Decoding (GD), which integrates an LLM with an affordance model for enhanced planning. Here, Llama 2 is used as the LLM. For the `Blocks & Bowls`, affordances are derived from CLIPort's predicted logits, while for `Letters`, we use ground-truth affordance values obtained from simulation. We use script policies as the primitive skills for LLM-based planner, GD and our VILA.

Analysis. The results are presented in Table IV, where each method is evaluated over 20 episodes per task within each category. We observe that CLIPort-based methods have a limited capacity for generalizing to novel, unseen tasks. Given that GD requires access to the output token probabilities of LLMs, we employ Llama 2 instead of GPT-4 for GD. As depicted in Table IV, both Llama 2 and GPT-4 exhibit comparable performances across all tasks, ensuring a fair comparison between GD and VILA (utilizing GPT-4V). While GD surpasses other LLM-based planning methods by leveraging an external affordance model, it significantly lags behind VILA. This finding further highlights the benefits of synergistic reasoning between vision and language for high-level robotic planning.

V. Conclusion, Limitations, & Future Works

In this work, we present VILA, a novel approach for robotic planning that utilizes VLMs to decompose a high-level language instruction into a sequence of actionable steps. VILA integrates perceptual information into the reasoning and planning process, enabling the understanding of commonsense knowledge in the visual world (e.g., spatial layouts and object attributes). It also supports flexible multimodal goal specification and naturally integrates visual feedback. Our extensive evaluation, conducted in both real-world and simulated settings, demonstrate VILA's effectiveness in addressing a variety of complex, long-horizon tasks.

VILA has several limitations that future work can improve. First, we presuppose the existence of all single-step primitive skills. While obtaining robust low-level control policies remains a challenging problem, recent advancements in transferring web knowledge to robotic control [7, 8] holds promise for enabling the cultivation of a repertoire of generalizable skills. Secondly, our dependence on a black-box VLM hampers steerability and complicates the explanation of certain errors. Future developments could leverage parameter-efficient fine-tuning methods [32, 33] to customize VLMs [24]. Finally, our current approach excludes in-context examples within prompts, leading to a more versatile output format. Methods developed for prompting [45, 89] can also be used to refine output consistency.

Acknowledgments

This work is supported by the Ministry of Science and Technology of the People's Republic of China, the 2030 Innovation Megaprojects "Program on New Generation Artificial Intelligence" (Grant No. 2021AAA0150000). This work is also supported by the National Key R&D Program of China (2022ZD0161700).

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

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

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

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

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

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

[10] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[11] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[12] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. *arXiv preprint arXiv:2306.06531*, 2023.

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. Task and motion planning with large language models for object rearrangement. *arXiv preprint arXiv:2303.06247*, 2023.

[17] Yiming Ding, Carlos Florensa, Pieter Abbeel, and Mariano Phielipp. Goal-conditioned imitation learning. *Advances in neural information processing systems*, 32, 2019.

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

[19] Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-language models as success detectors. *arXiv preprint arXiv:2303.07280*, 2023.

[20] Ben Eysenbach, Russ R Salakhutdinov, and Sergey Levine. Search on the replay buffer: Bridging planning and reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

[22] Richard E Fikes and Nils J Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208, 1971.

[23] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4): 163–352, 2022.

[24] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*, 2023.

[25] Caelan Reed Garrett, Chris Paxton, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Dieter Fox. Online replanning in belief space for partially observable task and motion problems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5678–5684. IEEE, 2020.

[26] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 4:265–293, 2021.

[27] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[28] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[29] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[30] Nicklas Hansen, Zhecheng Yuan, Yanjie Ze, Tongzhou Mu, Aravind Rajeswaran, Hao Su, Huazhe Xu, and Xiaolong Wang. On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline. *arXiv preprint arXiv:2212.05749*, 2022.

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

[32] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[33] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[34] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022.

[35] Yingdong Hu, Renhao Wang, Li Erran Li, and Yang Gao. For pre-trained vision models in motor control, not all

policy learning methods are created equal. *arXiv preprint arXiv:2304.04591*, 2023.

[36] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

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

[38] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

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

[40] Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, et al. Grounded decoding: Guiding text generation with grounded models for robot control. *arXiv preprint arXiv:2303.00855*, 2023.

[41] Brian Ichter, Pierre Sermanet, and Corey Lynch. Broadly-exploring, local-policy trees for long-horizon task planning. *arXiv preprint arXiv:2010.06491*, 2020.

[42] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[43] Leslie Pack Kaelbling and Tomás Lozano-Pérez. Hierarchical task and motion planning in the now. In *2011 IEEE International Conference on Robotics and Automation*, pages 1470–1477. IEEE, 2011.

[44] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[45] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[46] Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006.

[47] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[48] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023.

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

[50] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *arXiv preprint arXiv:2303.12153*, 2023.

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

[52] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[53] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[54] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

[55] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[56] Neil Houlsby Matthias Minderer, Alexey Gritsenko. Scaling open-vocabulary object detection. *NeurIPS*, 2023.

[57] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*.

[58] Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *Advances in neural information processing systems*, 31, 2018.

[59] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[60] Dana Nau, Yue Cao, Amnon Lotem, and Hector Munoz-Avila. Shop: Simple hierarchical ordered planner. In *Proceedings of the 16th international joint conference on Artificial intelligence-Volume 2*, pages 968–973, 1999.

[61] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023.

[62] OpenAI. Gpt-4 technical report, 2023.

[63] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,

Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[64] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *International Conference on Machine Learning*, pages 17359–17371. PMLR, 2022.

[65] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[66] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

[67] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023.

[68] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

[69] Earl D Sacerdoti. *A structure for plans and behavior*. PhD thesis, Department of Computer Science, Stanford University, 1975.

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

[71] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[72] Tom Silver, Soham Dan, Kavitha Srinivas, Joshua B Tenenbaum, Leslie Pack Kaelbling, and Michael Katz. Generalized planning in pddl domains with pretrained large language models. *arXiv preprint arXiv:2305.11014*, 2023.

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

[74] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

[75] Lucille Alice Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.

[76] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[77] Marc Toussaint. Logic-geometric programming: An optimization-based approach to combined task and motion planning. In *IJCAI*, pages 1930–1936, 2015.

[78] Marc A Toussaint, Kelsey Rebecca Allen, Kevin A Smith, and Joshua B Tenenbaum. Differentiable physics and stable modes for tool-use and manipulation planning. 2018.

[79] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[80] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res*, 2:20, 2023.

[81] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

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

[83] Robert Wilensky. Planning and understanding: A computational approach to human reasoning. 1983.

[84] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*, 2023.

[85] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[86] Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128*, 2023.

[87] Danfei Xu, Suraj Nair, Yuke Zhu, Julian Gao, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Neural task programming: Learning to generalize across hierarchical tasks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3795–3802. IEEE, 2018.

[88] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023.

[89] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan.

Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[90] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[91] Zhecheng Yuan, Sizhe Yang, Pu Hua, Can Chang, Kaizhe Hu, Xiaolong Wang, and Huazhe Xu. Rl-vigen: A reinforcement learning benchmark for visual generalization. *arXiv preprint arXiv:2307.10224*, 2023.

[92] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34: 23634–23651, 2021.

[93] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pages 726–747. PMLR, 2021.

[94] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[95] Tong Zhang, Yingdong Hu, Hanchen Cui, Hang Zhao, and Yang Gao. A universal semantic-geometric representation for robotic manipulation. *arXiv preprint arXiv:2306.10474*, 2023.

[96] Xiaohan Zhang, Yan Ding, Saeid Amiri, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang. Grounding classical task planners via vision-language models. *arXiv preprint arXiv:2304.08587*, 2023.

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