VLABench: A Large-Scale Benchmark for Language-Conditioned Robotics Manipulation with Long-Horizon Reasoning Tasks

Shiduo Zhang[1], Zhe Xu[1], Peiju Liu[1]*, Xiaopeng Yu[1]*, Yuan Li[1], Qinghui Gao[1], Zhaoye Fei[1], Zhangyue Yin[1], Zuxuan Wu[1], Yu-Gang Jiang[1], Xipeng Qiu[1†]

[1]Fudan University

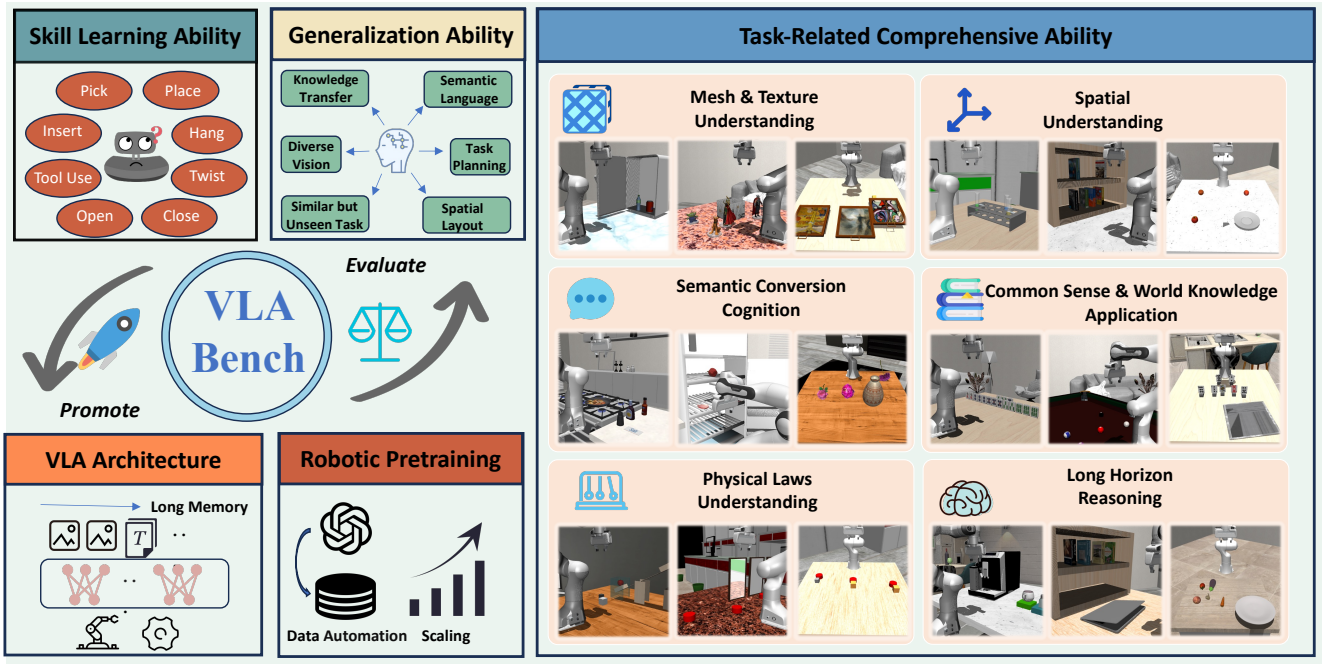

Figure 1. **Overview of VLABench.** VLABench is a large-scale language-conditioned manipulation benchmark to evaluate the comprehensive skill learning and generalization ability of action policies especially pre-trained vision-language-action models.

Abstract

General-purposed embodied agents are designed to understand the users' natural instructions or intentions and act precisely to complete universal tasks. Recently, methods based on foundation models especially Vision-Language-Action models (VLAs) have shown a substantial potential to solve language-conditioned manipulation (LCM) tasks well. However, existing benchmarks do not adequately meet the needs of VLAs and relative algorithms. To better define such general-purpose tasks in the context of LLMs and advance the research in VLAs, we present VLABench, an open-source benchmark for evaluating universal LCM task learning. VLABench provides 100 carefully designed categories of tasks, with strong randomization in each category of task and a total of 2000+ objects. VLABench stands out from previous benchmarks in four key aspects: 1) tasks requiring world knowledge and common sense transfer, 2) natural language instructions with implicit human intentions rather than templates, 3) long-horizon tasks demanding multi-step reasoning, and 4) evaluation of both action policies and language model capabilities. The benchmark assesses multiple competencies including understanding of mesh&texture, spatial relationship, semantic instruction, physical laws, knowledge transfer and reasoning, etc. To support the downstream finetuning, we provide high-quality training data collected via an automated framework incorporating heuristic skills and prior information. The experimental results indicate that both the current state-of-the-art pretrained VLAs and the workflow based on VLMs face challenges in our tasks. [1][2].

[1]Codes and more videos are available at https://vlabench.github.io/

[2]Corresponding to: sdzhang23@m.fudan.edu.cn, xpqiu@fudan.edu.cn

1. Introduction

Language-conditioned manipulation represents a fundamental challenge in embodied AI and a stepping stone toward Artificial General Intelligence [1, 5, 14]. Such tasks require agents to master multiple capabilities: interpreting natural language instructions, understanding complex environments, making decisions, formulating plans, and executing precise actions. The rapid advancement of Large Language Models (LLMs) and Vision-Language Models (VLMs) [1, 15] has revolutionized the field with their impressive general abilities in semantic understanding, coding, planning, and reasoning. The strong generalization capabilities has inspired two main approaches in language-conditioned manipulation: pre-training vision-language-action models using large-scale robotics data, as demonstrated by RT-2 and Palm-E [5, 14, 50], and integrating foundation models into agent workflows, like VoxPoser and Copa [24, 25], which combine LLM/VLM outputs with grasp prediction [16, 17] and motion planning algorithms [28].

While real-world robotics experiments provide valuable insights, their complexity and environmental variability often challenge reproducibility. Simulation-based evaluation has emerged as a fair and practical alternative. Existing benchmarks like RLBench, Calvin, and LIBERO [27, 38, 45] offer diverse task sets but fall short in addressing the unique requirements of foundation model-based methods. Tasks designed to align with the capabilities of foundation model-based algorithms should encompass nuanced semantic understanding of user intent, the integration of common-sense knowledge, and a robust ability to interpret diverse visual scenes, as well as require sophisticated multi-step reasoning. Such tasks demand a sophisticated integration of multimodal understanding to effectively interpret and respond to complex, real-world contexts. For example, one of the tasks in RT-2 [5] is "move the Coke can to Taylor Swift", while another task in CoPA [24] is "make me a cup of pour-over coffee". The first task challenges the robot to use common sense to identify Taylor, a capability of knowledge transfer that previous policies struggled to achieve. The second task further intensifies the difficulty, requiring the robot to decompose the task into subtasks and execute the steps to operate a coffee machine—a long–horizon challenge that has previously been difficult for a single policy to accomplish.

To better define the types of language-conditioned manipulation tasks suited for foundation models and provide a standardized evaluation suite to advance robotics research, we introduce VLABench. VLABench is an open-source benchmark specifically designed for methods utilizing foundation models. The tasks in VLABench are carefully divided into several dimensions to evaluate models across various aspects, including 1) Mastery of **common sense and world knowledge**, 2) Understanding of **mesh and texture**, 3) Comprehension of **semantically rich instructions**, 4) **Spatial** understanding, 5) Grasp of **physical rules**, and 6) **Reasoning** ability. For benchmarking purposes, VLABench offers 100 task categories with comprehensive evaluations across various methods. With a diverse collection of over 2,000 3D objects and scenes, VLABench creates a wide range of visual contexts and tasks. It enables the assessment of generalization capabilities through learning across multiple skills, providing thorough evaluations spanning visual, linguistic, planning, knowledge transfer, and action dimensions.

To ensure fair comparison and evaluation, we develop an automated data collection framework to construct standardized datasets for each task, supporting model training and fine-tuning. Using this dataset, we conduct extensive experiments to evaluate and benchmark three distinct types of approaches: pre-trained VLA, workflows integrating foundation models, and vision-language models (VLMs). The experimental results indicate that existing VLA methods perform poorly on our tasks and don't exhibit the level of generalization abilities or the "emergent" phenomena observed in large models [61]. We summarize contributions as follows:

- We propose VLABench, the first benchmark designed to comprehensively evaluate the capabilities of VLAs and VLMs in robotics manipulation tasks, covering multiple dimensions such as skills, vision, language, task execution, common sense, and reasoning.
- We define 100 novel LCM tasks tailored to the capabilities of foundation models within a standardized evaluation framework. These tasks require a deep understanding of semantics, vision, space, and physical laws, as well as the ability to plan long-horizon tasks and transfer world knowledge and common sense into task execution.
- We provide a scalable data construction framework and a standardized evaluation dataset. This automated data construction approach facilitates future research on pre-training robotics data.
- Our experiments demonstrate that current pre-trained VLAs have yet to exhibit the strong generalization capabilities observed in LLMs, and existing SOTA VLMs also show limitations in embodied scenarios.

2. Related Works

Benchmarks and Datasets. Numerous benchmarks such as RLBench and LIBERO [27, 35, 38, 45, 66] have been proposed to evaluate language-conditioned manipulation policies. A comparison of these benchmarks is provided in Table 1. However, most of these focus on skill learning and fail to sufficiently address long-horizon planning capabilities. Meanwhile, some benchmarks [52, 55, 63] address room-scale mobile manipulation tasks that require

Benchmarks	SemLang	LogiReason	Knowledge	DL	DR	N-task	Cate-obj	N-obj	AI-Gen	MultiCam	PCD	Cross Emb	Auto Traj
Alfred[52]	✗	✗	✗	-	✗	7	-	3578	✗	✗	✗	✓	✗
Rlbench[27]	✗	✗	✗	28.86	✓	100	28	168	✗	✓	✗	✗	✓
Calvin[45]	✗	✗	✗	40.00	✗	34	5	30	✗	✓	✗	✗	✗
ManiSkill[21, 46, 56]	✗	✗	✗	-	✓	20	100	2600	✗	✓	✓	✓	✓
LIBERO[38]	✗	✗	✗	17.96	✓	130	51	75	✗	✗	✗	✗	✗
RoboCASA[47]	✗	✗	✗	-	✓	100	153	2509	✓	✓	✗	✓	✓
ARNOLD[20]	✗	✗	✗	-	-	8	-	40	✓	✓	✓	✗	✓
SimplerEnv[35]	✗	✗	✗	27.66	✓	8	-	17	✓	✓	✓	✗	✓
Behavior-1K[32]	✗	✗	✗	-	-	1000	2211	9331	✗	✗	✗	✓	✗
Habitat 2.0[55]	✗	✗	✗	-	-	3	46	169	✗	✗	✗	✗	✗
VLABench	✓	✓	✓	75.96	✓	100	163	2164	✓	✓	✓	✓	✓

Table 1. Comparison of popular benchmarks in robot learning. **SemLang**: semantically non-template rich language instructions. **LogiReason**: task logic and relevant information reasoning. **Knowledge**: tasks require the application of common sense and world knowledge. **DL**: quantitative task difficulty level. **DR**: task domain randomization. **N-task**: total number of tasks. **Cate-obj**: categories of assets used in the simulation. **N-obj**: total number of objects in the asset library. **AI-Gen**: use of generative AI models for asset library creation. **MultiCam**: use of multiple cameras. **PCD**: support point cloud data in 3D methods. **Cross Emb**: support for cross-embodiment. **Auto Traj**: supporting automated data collection.

long memory and reasoning. Yet, these interactions typically occur through interfaces, rather than direct physical manipulation, limiting the transferability of learned policies to real-world scenarios. Additionally, while efforts [21, 32, 38, 47] have made strides in task format, difficulty, and scale, these benchmarks have largely overlooked the guiding role of language in tasks, often relying on template instructions. VLABench is the first to introduce features such as natural human interaction, implicit goal-oriented semantics, and requirements based on common sense into robot manipulation tasks, as shown in Figure 20. In terms of generalization evaluation, previous works [27, 38, 45] typically assess models at the instance level within the same category, which limits their ability to evaluate generalization across diverse object categories or different tasks within the same skill set. In contrast, VLABench is the first benchmark to evaluate generalization capabilities across a wide range of tasks, object types, and task categories, providing a more comprehensive assessment of model versatility.

Large-scale datasets have been built in both real and simulation [6, 47, 50, 60] for large-scale imitation learning. However, real-world data faces challenges related to scalability, making it difficult to gather sufficient data at scale [4]. Simulated datasets, while more scalable, often suffer from limited diversity in scenarios and tasks [27, 45], and still require teleoperation [20, 38] for data collection. VLABench addresses these limitations by offering a broader range of tasks that are more closely aligned with real-world conditions, covering diverse aspects of vision, language, tasks, and skills. Furthermore, it introduces an efficient and robust process for the automated generation of simulated data, significantly enhancing task diversity and scalability.

Pretrained Vision-Language-Action Models. The recent rise of multimodal models [1, 13, 40, 65] and the organization of operational datasets [49, 60], have led to the integration of vision-language-action models (VLAs) [4, 5, 14, 30] into language-conditioned manipulation tasks. While the term VLA generally refers to models that combine visual and language inputs for policy learning, we focus specifically on approaches leveraging pre-trained models. Several works [4, 14, 30] have applied further training to pre-trained vision-language models (VLMs) for language-conditioned manipulation. These models demonstrate impressive generalization to unseen objects and tasks, yet their control precision is somewhat limited by the discretization of actions [51]. To address this limitation, some approaches have explored using diffusion models [9, 51] as policy networks or using diffusion decoders [3, 34, 62]. Pre-trained models based on diffusion models [37, 43] have shown promising advancements in improving continuous space distribution learning. VLABench includes a selection of these representative methods for comprehensive evaluation.

Framework Utilizing Foundation Models. Pre-trained language models [6, 15, 53] and vision-language models [1, 42] have demonstrated strong generalization and versatility. Some researchers [24, 25, 39] combine the general perception and cognitive abilities of these pre-trained models with traditional planning and control algorithms to create agent workflows. These frameworks allow robots to perform complex zero-shot manipulation tasks without requiring additional training. To harness the capabilities of foundation models for manipulation, some works [25, 26, 36] utilize the code comprehension and generation abilities of large language models alongside motion planning optimization algorithms to tackle fundamental manipulation tasks. Additionally, some methods [23, 24] leverage large models to decompose long-horizon tasks into subtasks, then integrate perception and trajectory generation modules to construct the entire manipulation pipeline. However, most zero-shot methods of this kind heavily rely on prompt design [25], the accuracy of each module, and even the specific parameters of the models being invoked [24]. While these methods demonstrate strong generalization, they often face challenges with accuracy. VLABench provides a

zero-shot evaluation framework to assess the performance of such workflows and offers insights into their effectiveness.

3. VLABench

3.1. Task Description

VLABench is composed of 60 primitive and 40 composite tasks, categorized by task difficulty levels. These tasks are designed to encompass a rich variety of skills while covering ample visual and language semantic information. For skill learning, 100 tasks in VLABench cover a wide range including 1) Pick&place, 2) Open&close door, 3) Open&close drawer, 4) Hang objects on the wall, 5) Use tool e.g. Hammer nail, 6) Press button, 7) Insert, 8) Pour, 9) Twist, and 10) Explore. In addition, VLABench places greater emphasis on real-life scenarios and essential daily tasks, representing more interactive language instructions, a wider variety of task settings, the integration of common sense and societal knowledge, and long-horizon tasks requiring logical planning, as shown in Figure 2. Notably, VLABench adopts a stricter definition of task generalization, which will be elaborated on in the Section 3.3. The whole task list can be found in supplementary material.

Primitive Tasks. Primitive tasks are divided into five dimensions, each corresponding to the assessment of a specific ability dimension.

- **Mesh&Texture Understanding.** This type of task requires the model to recognize different meshes and understand various texture features. Take the *SelectToy* task shown in Figure 2 (a) as an example, the robot is directly required to place a specific toy e.g. Aquaman into a receptacle. The model must possess strong visual capabilities to accurately recognize such complex meshes and textures.
- **Spatial Understanding.** Spatial understanding tasks involve various spatial relationships, such as the n-th left-/right position, inside/outside of a receptacle, the m-th row and n-th column, near/far, and beside a specific object, representing relative positional relationships. Figure 2 (b) shows one case in task in *PullBook*. Such a complex relative positional relationship imposes extremely high demands on the model's multimodal understanding [8].
- **Common Sense & World Knowledge.** Tasks relative to common sense/world knowledge require the agent to transfer the knowledge gained in the pre-train stage to solve the problem. The task shown in Figure 2 (c) requires the agent not only to recognize different types of flowers from visual information but also to leverage world knowledge to determine that "the tulip is the national flower of the Netherlands".
- **Semantic Understanding.** This type of task emphasizes the complexity, subtlety, and natural interactivity of language instructions. Task objectives are often implicitly conveyed through a natural conversation. To perform well in *GetDrink* task as Figure 2 (d) shown, the agent must capture the implied request from a lengthy instruction: to take out a chilled cola from the refrigerator.
- **Physical Law.** This type of task expects the robot to integrate visual information and take correct actions based on physical principles and real-time observation. In *UseSeesaw* task in Figure 2 (e), the robot is commanded to grasp an object that can not be achieved directly. The agent must recognize the need to apply the principle of leverage by using sufficient weight to lift the target object on the other end.

Composite Tasks. Composite tasks in VLABench involve the combination of multiple skills, long-term task planning, and multi-step logical reasoning from instructions, scenes, and even game rules. Figure 2 (f) showcases a variety of challenging complex tasks. Composite tasks have a significantly longer trajectory horizon, with an average episode length exceeding 500 timesteps—considerably more than the average of 120 timesteps for primitive tasks. In Figure 2 (f1), the agent must not only correctly identify all poker cards from visual information and use world knowledge of poker rules to select the best hand, but also flip face-down cards to acquire complete information. This type of task, requiring the agent to consciously fulfill prerequisite conditions, has never been modeled before. Composite tasks also require extracting the user's implicit needs from natural dialogue. The case in Figure 2 (f2) needs the agent to place the Python textbook on the table and open the laptop, without direct instruction. More cases are shown in Figure 20.

3.2. Task Difficulty Level

VLABench stands out among language-conditioned manipulation benchmarks not only for its multi-dimensional task design and evaluation but also for its hierarchical difficulty levels. The difficulty level (DL) is quantified using three metrics: task variants [27, 44], task sub-steps, and reasoning steps. These metrics quantify the overall task difficulty level from the perspectives of generalization difficulty, normalized episode horizon length, and intelligence. The detailed computation of DL is provided in Section 6.3. As shown in Table 1, VLABench's difficulty level surpasses all previous benchmarks and introduces new challenges that a VLA model should overcome.

3.3. Benchmark

Evaluation. VLABench organizes evaluations into three categories: assessments of pretrained or fine-tuned vision-language-action (VLA) models, heuristic workflows that integrate foundation models with various algorithms, and multi-dimensional evaluations of vision-language models.

- **Generalization Ability of VLAs.** For pre-trained VLA

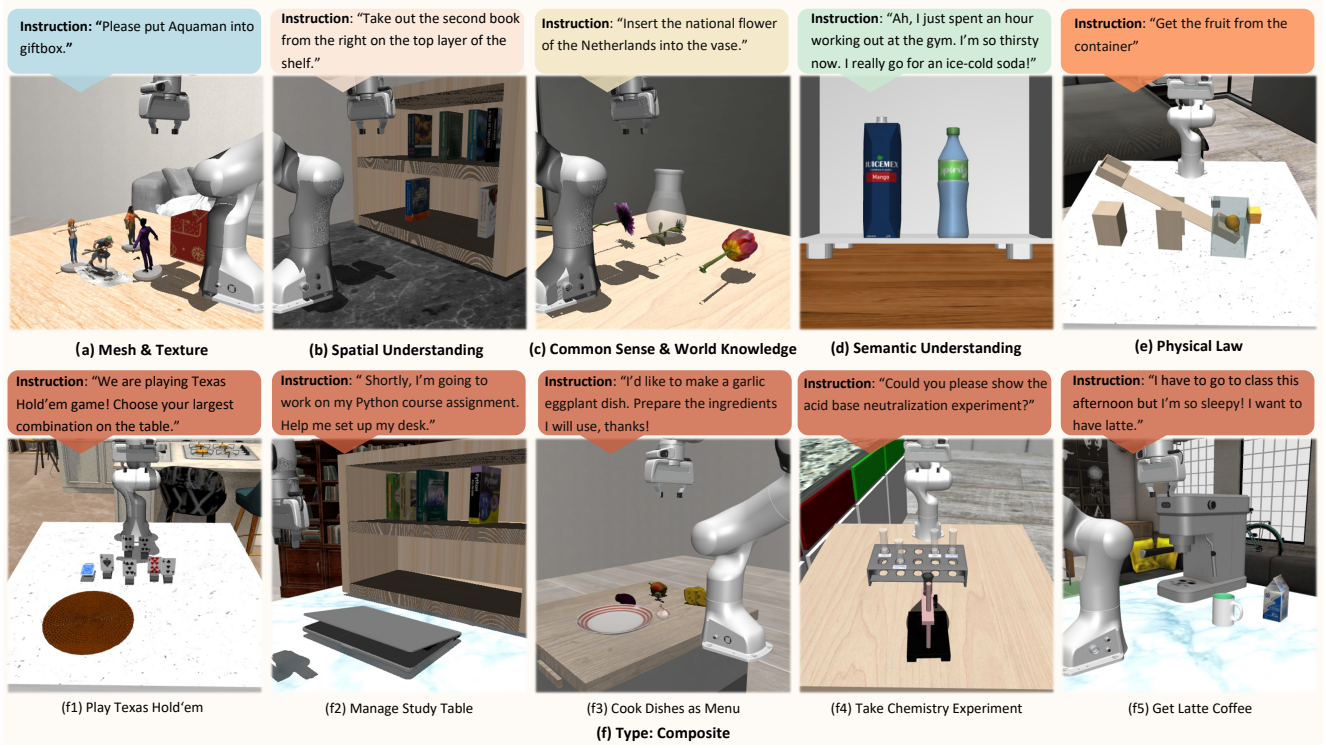

Figure 2. **Task examples in each dimension.** The first row showcases examples of primitive tasks from Section 3.1, while the second row presents examples of composite tasks.

models, the evaluation focuses on the model's generalization ability across multiple dimensions. In the standard evaluation, we have set up six evaluation tracks.

- Track 1: Rapid downstream task transferability and skill acquisition ability. The evaluation episodes use objects seen in the training set, with minimal differences.
- Track 2: Cross-category visual generalization. The evaluation episodes use unseen categories of objects as the target entity for evaluation.
- Track 3: Common sense application. Building upon Track 1, replace parts of the instructions that directly reference the target object with content requiring common sense reasoning in evaluation episodes.
- Track 4: Semantically rich instruction adaptation. Building upon Track 1, replace the instructions with unseen and more complex ones.
- Track 5: Cross-domain behavior transferability. The evaluation tasks are replaced with ones that differ from those in the training set, but require similar actions.
- Track 6: Long-horizon task learning. Training and evaluation using more challenging composite tasks require additional skills and reasoning.

- **Zero-shot Transfer Ability of Heuristic Workflow.** Training-free workflow methods are evaluated under sets of tasks in all dimensions mentioned in Section 3.1. We also extend the evaluation to cover various skills and long-horizon tasks to assess the overall capability and execution robustness of the workflow. All agents are evaluated in a zero-shot setting.
- **Comprehensive Evaluation of VLMs' Capabilities.** Similar to heuristic workflows, the evaluation of VLMs is also comprehensive. Since VLMs lack intrinsic action capabilities, we organized a skill library and integrated it into a domain-specific language (DSL) [48, 54], leveraging annotated asset information as prior knowledge. This DSL functions as a straightforward API that VLMs can call, enabling efficient interaction. The evaluation pipeline is shown as Figure 3 and the details will be discussed in Section 9.3.

Metric. In addition to the success rate (SR), considering the long-horizon nature and high difficulty level of our tasks, we introduce the intention score (IS) and progress score (PS) for more granular evaluation. Intention score is a useful soft metric measuring the correctness trend in selecting the target object for the operation especially when sub-tasks are not yet completed. The discrete version metric is:

$$IS = \begin{cases} 1, & \text{if } distance_{min} < \epsilon \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $distance_{min}$ is the minimum distance between the end-effector's position and the target entity, and ϵ is empiri-

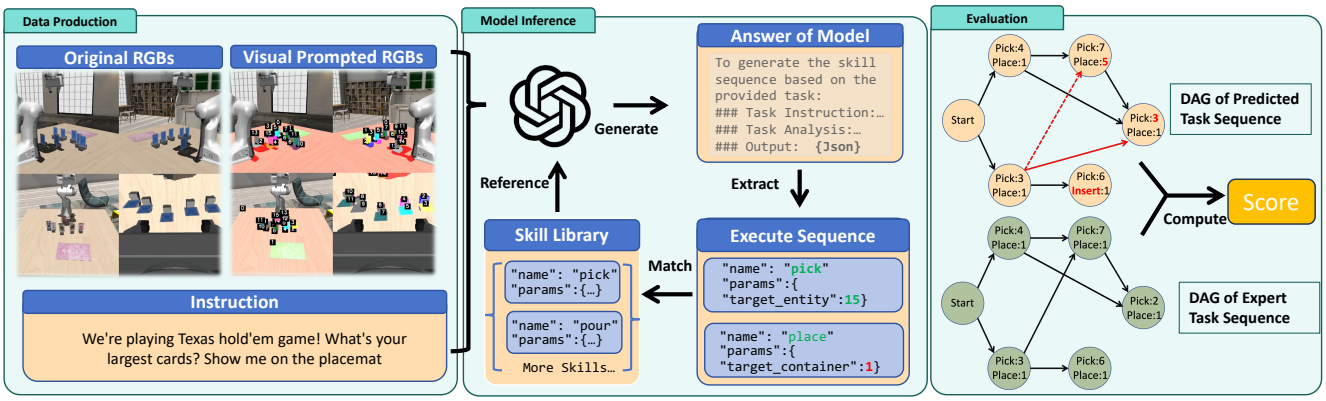

Figure 3. **Evaluation pipeline for VLMs**. **Step 1:** Sample the required four-view images, as well as those segmented with numerical information, from the simulation. Meanwhile, save the corresponding instructions and operation sequences. **Step 2:** Input the original and annotated images and the instructions into the model. Then, obtain the model's output and extract the generated operation sequence. **Step 3:** The two operation sequences are evaluated using four metrics, which are weighted and summed to produce the final score. The red part represents the error output of the model, where the red solid arrows represent the dependencies generated by the error, and red dashed arrows represent the dependencies lost by the error.

cally set to 0.15 meters in most cases after experimentation. The progress score refers to the completion level of sub-tasks in a long-horizon task and serves as a softer process supervision metric compared to the success rate.

$$PS = N_{done}/N_{sub} \tag{2}$$

where N_{done} is the number of completed sub-tasks, and N_{sub} is the total number of task sub-tasks.

3.4. Dataset Construction

Trajectory Generation. As human teleoperation is time-consuming and not scalable [38, 47], we developed an efficient, scalable automated data collection pipeline based on our custom skill library. Inspired by [20], our data collection framework leverages the prior information including point clouds of the environment, annotated entities' grasp-points, target entity at the current step, etc. The data collection framework includes multiple task-specific motion planners. These motion planners call upon the skills in the skill library based on the current task progress and determine parameters by incorporating prior information. Subsequently, the selected skills generate trajectories using RRT [29], with quaternion interpolation achieved through spherical linear interpolation. The final trajectory is smoothed using a Bezier curve to optimize path quality. To enhance sample efficiency, reject sampling and failure-triggered early termination are applied. We demonstrate the superiority of our data framework through comparative experiments. Detailed analysis can be found in Section 8.3.

Data Diversity and Quality. During the data construction process, we introduced diverse task variants and domain randomization across different episodes of the same task to ensure the diversity of the training data, as discussed in Sections 6.1 and 6.4. Following the approach of previous benchmarks built on Mujoco [38, 47], the dataset is stored in the same format, with similar visual rendering quality and trajectory accuracy. We further analyze the data quality in Section 8.5.

Instruction Augmentation. We use GPT-4o [1] to generate descriptions that incorporate target-specific characteristics and interactive instructions that encompass a variety of contexts and intentions. Section 8.4 provides more details of the instruction generation.

4. Experiments

Following Section 3.3, we conducted experiments centered on pre-trained VLA models, workflows incorporating multiple algorithmic modules, and various VLMs. The remainder of this section provides a detailed description of the experimental setup.

4.1. Generalization Ability of VLAs

Pretrained VLAs are expected to possess robust generalization and versatility similar to LLMs. Experiments about are set to address the following research questions:

Q1: Do pre-trained VLAs exhibit stronger general abilities with unseen categories of objects?

Q2: Can pre-trained VLAs transfer their general knowledge and behavioral abilities to similar but unseen tasks?

Q3: Can pre-trained VLAs understand natural user interactions and implicit goal requirements?

Q4: Do pre-trained VLAs have the potential to transfer their world knowledge to related tasks?

Q5: Can existing VLA architectures accurately support the completion of long-horizon tasks?

Experiment Setup. To investigate the questions outlined

Model	Track 1			Track 2			Track 3			Track 4			Track 5			Track 6		
	IS	PS	SR	IS	PS	SR	IS	PS	SR	IS	PS	SR	IS	PS	SR	IS	PS	SR
Octo[57]	32.0	5.6	2.0	20.4	2.0	0.0	22.0	2.8	1.2	26.8	4.0	1.2	13.6	0.0	0.0	17.2	3.0	0.0
OpenVLA[30]	50.0	13.2	6.4	35.6	7.0	3.2	38.0	8.0	3.2	38.0	9.2	2.8	16.0	0.0	0.0	28.8	9.6	0.4
RDT-1B[43]	50.0	14.3	6.0	39.2	8.2	3.2	36.0	7.2	4.0	35.6	8.0	4.4	20.4	0.0	0.0	21.6	9.2	0.0
π_0-Base[3]	59.6	**29.3**	**13.2**	**49.6**	**21.4**	**6.0**	**60.4**	**20.0**	**5.2**	**60.4**	**19.3**	3.6	22.0	0.0	0.0	**34.4**	12.8	**2.0**
π_0-Fast[3]	**59.9**	25.5	9.2	47.6	16.5	2.4	50.0	15.4	3.6	52.0	17.2	**6.8**	25.6	0.0	0.0	34.0	**14.4**	1.6

Table 2. **Overall experiment result of 6 evaluation tracks of fine-tuned VLAs.** The detailed result of each task is reported in Table 9.

above, we fine-tuned various pre-trained VLA architectures [3, 30, 43, 57] on our high-quality dataset, and evaluated them on different tracks mentioned in Section 3.3 to assess their performance across a range of tasks and conditions. Considering the high task difficulty, we selected five primitive tasks for experiments in tracks 1 to 5, with 200 data samples per task, and trained the models following a multitask training approach, following the setting in previous work [3, 43]. We then assessed the models across 50 episodes for each task, resulting in a total of 250 trials per track. For track 6, we selected five composite tasks, keeping the other settings consistent. Section 9.1 provides a complete list of tasks in different tracks and details of the model fine-tuning process.

Result and Analysis. Table 2 presents the evaluation results across multiple tracks. These models perform poorly on most of tasks, as they are trained on a more diverse and limited set of fine-tuning data, with a higher difficulty level. Through the experimental results of each track and comparisons across different tracks, our analysis is as follows:

- The primary cause of poor performance is the insufficient ability for rapid transfer in downstream tasks. Episodes in Tracks 1 and 6 closely resemble the training set, primarily assessing the model's basic fitting ability. Despite this, the models still perform poorly on certain tasks. We believe the main reason for the failure of most generalization experiments is the inaccuracy in basic action generation, especially Octo.
- Pre-trained VLA models exhibit some ability to handle unseen object categories. By comparing the results of Track 1 and Track 2, we observe that although all models experience a performance decline in Track 2, they are still able to complete a small subset of tasks. This is likely due to the benefits of large-scale robotics data pretraining or even earlier VLM pretraining.
- Performance decline with unseen semantic language. By combining the experimental results of Track 3 and Track 4, we find that the model's language understanding ability has deteriorated by a larger margin. Among these, the performance decline is more pronounced in models based on non-VLM architectures: RDT's progress score decreases by 45.2% while OpenVLA and π_0 only decreases by 31.3% and 34.2%, respectively, in Track 4.
- Nearly absent cross-domain transferability. We found that all the models fails to perform similar but unseen tasks, due to significant visual and linguistic input differences. These policies exhibit uncertainty in behavior when faced with unfamiliar tasks, resulting in IS that are similar.

We further analyze that data scaling and pre-training contribute positively to the model's performance Through basic ablation experiments. We also discuss in detail the potential issues with current VLAs, such as multimodal data co-training and model architecture designs. These discussions are presented in Section 10.1. We also discuss the gap between real and simulated environments in Section 10.2.

4.2. Zero-shot Ability of Agent

For our evaluation of foundation model-based algorithms, we reviewed two state-of-the-art frameworks, Voxposer [25] and CoPA [24], and the comparison results are shown in Figure 4. Given Voxposer's dependence on large language models (LLMs), we assessed Voxposer's performance with and without visual perception capabilities. While Voxposer performed adequately on basic tasks and achieved the Progress Scores of 30–40, its reliance on LLM-driven motion planning often led to grasping failures due to limited information for effective grasp planning, especially when interpreting rotation in non-visual contexts, resulting in low overall scores.

Interestingly, the foundational LLM alone maintained relatively stable scores in semantic understanding and reasoning tasks without visual input. However, adding visual perception slightly reduced performance in these areas while significantly improving spatial reasoning, where LLM-only setups struggle with spatial accuracy due to lack of spatial information.

The lack of closed-loop feedback limits these models' ability to perform physical reasoning tasks, particularly those involving dynamic interactions, leading to lower scores in this dimension. Both models struggle with high-complexity tasks, succeeding mainly in entity recognition but performing poorly in long horizon task reasonable breaking down. This finding underscores the requirements for advancing foundational model-based frameworks to address complex reasoning. In fact, although the methods mentioned above emphasize their zero-shot capabilities and generalization to new scenarios, their modular design often limits the upper bound of their performance. A more detailed discussion of this can be found in Section 10.3.

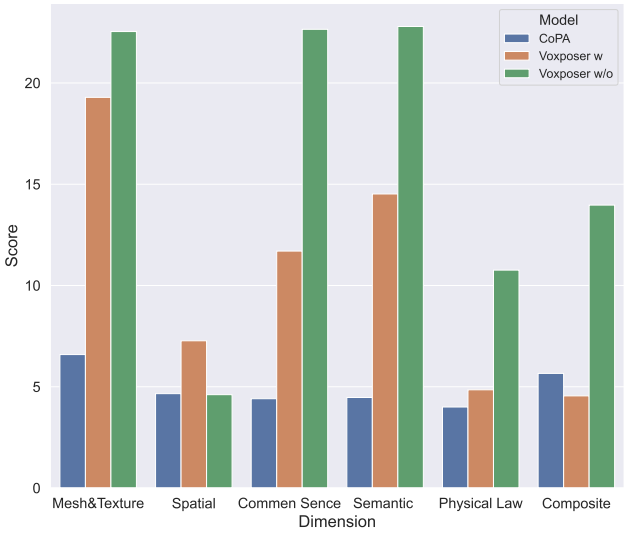

Figure 4. Evaluation progress score for Voxposer and CoPA. *Voxposer w/o* refers to the version without visual perception, where ground truth labels are directly provided for object selection. *Voxposer w* uses GPT-4V as the visual perception module.

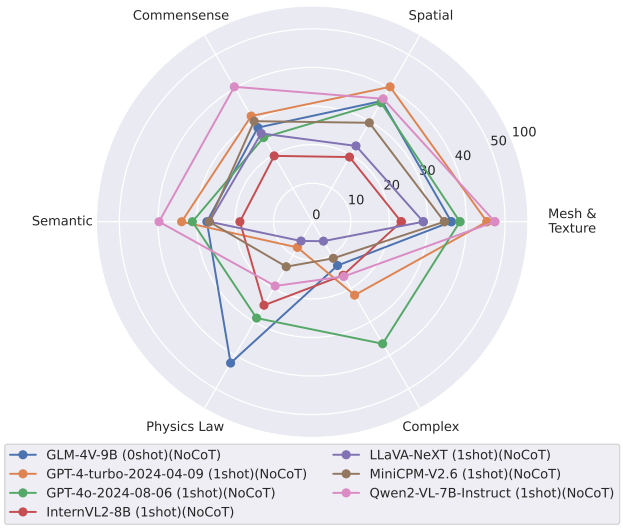

Figure 5. Radar charts depicting the performance of all VLM models across six dimensions. The reason why only GLM-4V-9B is evaluated in a zero-shot setting is that it does not support multi-graph inference, which is required for the other models.

4.3. Comprehensive Ability of VLMs

We referred to the evaluation results of multiple series of VLMs provided by OpenCompass [11] and selected models from different families [1, 2, 7, 19, 22, 41] with strong overall performance as shown in Figure 5. We evaluate the comprehensive performance of these models with the dataset derived from naturally self-contained information within a simulated environment. This dataset consists of a complex set of tasks designed to assess the VLM's ability to perceive visual stimuli and comprehend verbal instructions.

In line with the evaluation framework outlined in the Section 3.3, we evaluate the models across six dimensions of task performance. The results for each VLM model under the 1-shot[3] setting are summarized in Figure 5.

Although these VLMs perform well on most multimodal tasks and even some embodied tasks [11, 33], their performance, including that of GPT-4o, falls short when faced with more complex scenarios, instructions, and more challenging tasks. We surprisingly find that the open-source model Qwen2-VL-7B-Instruct performed competitively, surpassing GPT-4o on certain dimensions. However, all models struggled with complex tasks, especially those requiring long-term task decomposition and logic reasoning. Only GPT-4o achieved a score in the reasoning dimension comparable to those in other dimensions, while the other models scored around 20. Besides, performance declines significantly when linguistic instructions transition from direct semantics to abstract meanings, as shown in the semantic dimension. Different models appear to have distinct areas of expertise, e.g. LLaVA-NeXT exhibits weaker spatial perception ability, GLM-4V-9B excels in spatial and even physical law dimensions but lag in semantic comprehension. In Section 10.4, we conducted additional ablation experiments focusing on CoT and few-shot learning, followed by further analysis. Detailed data and error analysis can be found in Table 10. Overall, while the models demonstrate promising capabilities, their understanding and planning in embodied environments remain limited, highlighting the need for further advancements.

[3]As GLM-4V-9B does not support multiple image inputs, the 0-shot method is employed.

5. Conclusion

We propose VLABench, a large-scale benchmark designed for tasks with long-horizon and multi-dimensional reasoning. Such reasoning involves many dimensions, including implicit semantic goal extracting, logical reasoning, and long-horizon planning. The most important thing we do is to provide a positive definition of the capabilities that intelligent agents withs cognitive abilities should possess and the tasks they should be able to perform, through the provision of 100 standardized tasks. Additionally, VLABench constructed a scalable automated data collection framework for future's potential larger scale pertaining and a standardized dataset for fair comparison of VLAs, both in the present and in future developments. Our diverse and multiple experiments revealed that current VLAs and VLMs face significant challenges in our tasks, and there remain substantial uncertainties in research on robotics scaling. We hope that VLABench will inspire both the future research on robotics pertaining recipe and promote more robust VLA architectures development.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 6, 8

[2] 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. 8

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 3, 7, 11

[4] 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. 3

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023. 2, 3, 10

[6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 8

[8] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 4

[9] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. 3

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 28

[11] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023. 8

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 4

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2, 3

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 3

[16] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2020. 2, 4

[17] Hao-Shu Fang, Minghao Gou, Chenxi Wang, and Cewu Lu. Robust grasping across diverse sensor qualities: The graspnet-1billion dataset. *The International Journal of Robotics Research*, 2023. 2, 4

[18] Jensen Gao, Suneel Belkhale, Sudeep Dasari, Ashwin Balakrishna, Dhruv Shah, and Dorsa Sadigh. A taxonomy for evaluating generalist robot policies. *arXiv preprint arXiv:2503.01238*, 2025. 10

[19] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 8

[20] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20483–20495, 2023. 3, 6

[21] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for

generalizable manipulation skills. In *International Conference on Learning Representations*, 2023. 3

[22] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 8

[23] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023. 3

[24] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. *arXiv preprint arXiv:2403.08248*, 2024. 2, 3, 7

[25] 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. 2, 3, 7

[26] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 3

[27] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020. 2, 3, 4, 5

[28] Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning. *The international journal of robotics research*, 30(7):846–894, 2011. 2

[29] Sertac Karaman, Matthew R Walter, Alejandro Perez, Emilio Frazzoli, and Seth Teller. Anytime motion planning using the rrt. In *2011 IEEE international conference on robotics and automation*, pages 1478–1483. IEEE, 2011. 6

[30] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 3, 7, 11

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4

[32] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023. 3, 4

[33] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *arXiv preprint arXiv:2410.07166*, 2024. 8

[34] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024. 3

[35] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024. 2, 3, 5, 11

[36] 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. 3

[37] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024. 3

[38] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 6, 5

[39] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 3

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[41] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 8

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[43] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 3, 7, 11

[44] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023. 4, 11

[45] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3): 7327–7334, 2022. 2, 3

[46] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*, 2021. 3

[47] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar,

and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems*, 2024. 3, 6, 4, 10, 11

[48] Arne Nordmann, Nico Hochgeschwender, and Sebastian Wrede. A survey on domain-specific languages in robotics. In *International conference on simulation, modeling, and programming for autonomous robots*, pages 195–206. Springer, 2014. 5

[49] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 3

[50] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 2, 3

[51] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*, 2023. 3

[52] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 2, 3

[53] Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. Moss: An open conversational large language model. *Machine Intelligence Research*, 2024. 3

[54] Craig J Sutherland and Bruce MacDonald. Robolang: a simple domain specific language to script robot interactions. In *2019 16th International Conference on Ubiquitous Robots (UR)*, pages 265–270. IEEE, 2019. 5

[55] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021. 2, 3

[56] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024. 3

[57] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 7

[58] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 3

[59] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. 3

[60] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023. 3

[61] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification. 2

[62] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024. 3

[63] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, et al. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*, 2023. 2

[64] Kevin Zakka. obj2mjcf: Cli for processing composite wavefront obj files for use in mujoco. https://github.com/kevinzakka/obj2mjcf, 2023. Version 1.0. 4

[65] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15244–15253, 2023. 3

[66] Liming Zheng, Feng Yan, Fanfan Liu, Chengjian Feng, Zhuoliang Kang, and Lin Ma. Robocas: A benchmark for robotic manipulation in complex object arrangement scenarios. *arXiv preprint arXiv:2407.06951*, 2024. 2