QUAR-VLA: Vision-Language-Action Model for Quadruped Robots

Pengxiang Ding[12] Han Zhao[12] Wenjie Zhang[3] Wenxuan Song[4]
Min Zhang[12] Siteng Huang[12] Ningxi Yang[2] Donglin Wang[1⋆]

[1]Zhejiang University [2]Westlake University
[3]Beijing Normal University [4]Monash University
{dingpengxiang}@westlake.edu.cn

Abstract. The important manifestation of robot intelligence is the ability to naturally interact and autonomously make decisions. Traditional quadruped robot learning typically handles language interaction and visual autonomous perception separately, which, while simplifying system design, also limits the synergy between different information streams. This separation poses challenges in achieving seamless autonomous reasoning, decision-making, and action execution. To address these limitations, a novel paradigm, named **V**ision-**L**anguage-**A**ction tasks for **QUA**druped **R**obots (**QUAR-VLA**), has been introduced in this paper. This approach tightly integrates visual information and instructions to generate executable actions, effectively merging perception, planning, and decision-making. The central idea is to elevate the overall intelligence of the robot. Within this framework, a notable challenge lies in aligning fine-grained instructions with visual perception information. This emphasizes the complexity involved in ensuring that the robot accurately interprets and acts upon detailed instructions in harmony with its visual observations. Consequently, we propose **QUA**druped **R**obotic **T**ransformer (**QUART**), a VLA model to integrate visual information and instructions from diverse modalities as input and generates executable actions for real-world robots and present **QUA**druped **R**obot **D**ataset (**QUARD**), a large-scale multi-task dataset including perception, navigation and advanced capability like whole-body manipulation tasks for training **QUART** model. Our extensive evaluation shows that our approach leads to performant robotic policies and enables **QUART** to obtain a range of generalization capabilities.

Keywords: Robotics · Quadruped Robot Learning · Vision-Language-Action Model

1 Introduction

Quadruped robots, characterized by their excellent traversability on complex terrains and agile movements, have garnered significant attention in the field of robotics [14]. Researchers have extensively employed these robots to explore tasks encompassing autonomous navigation and manipulation [16, 17, 36].

⋆ Corresponding author

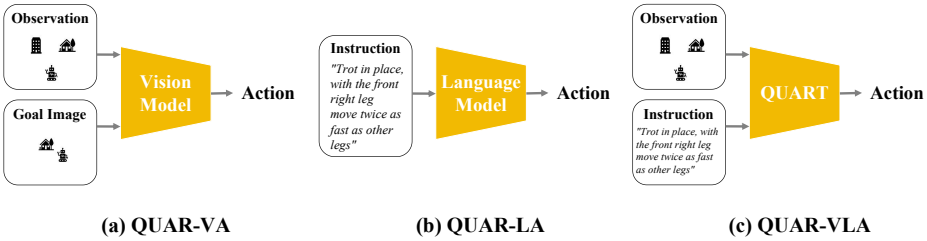

Fig. 1: Comparison of QUAR-VA, QUAR-LA, and QUAR-VLA. QUAR-VA solely utilizes coarse-grained vision information, lacking explicit instructions for handling diverse tasks. In contrast, **QUAR-LA** exclusively relies on language information and lacks of vision information for autonomy. Therefore, **QUART-VLA** combines both vision information and language instructions as inputs, enabling autonomous problem-solving across a range of tasks, revealing distinct input modalities and task capabilities.

Broadly speaking, the quadruped tasks consist of two major specifications: **V**ision-**A**ction tasks for **QUA**druped **R**obots (**QUAR-VA**) and **L**anguage-**A**ction tasks for **QUA**druped **R**obots (**QUAR-LA**). As depicted in Fig. 1, in **QUAR-VA** approaches [31], quadruped robots {receive} perception images captured from a first-person perspective and instruction images obtained from a third-person perspective to guide their actions. However, such a task specification often relies on a single (coarse-grained) goal image instruction, making it difficult to apply in many real-world combination tasks, *i.e.* requiring combining multiple sub-instructions. In contrast, employing language as instructions [36], **QUAR-LA** formulation allows for executing more fine-grained and diverse tasks, suitable for expressing combinational commands (*i.e. "before / then"*), complex spatial relationships (*i.e. "move to the left"*), and many commonsense command priors (*i.e. "move fast"*). Nevertheless, **QUAR-LA** approaches, which lack the integration of visual modality, hinder the robots' ability to perceive the environment, thereby impeding their autonomous navigation capabilities [36]. To enable quadruped robots to autonomously navigate and manipulate various tasks, in this paper, we propose a new paradigm: **V**ision-**L**anguage-**A**ction tasks for **QUA**druped **R**obots (**QUAR-VLA**), integrating visual information and instructions from diverse modalities as input and generating executable actions for real-world robots.

This task primarily encompasses two challenges. Firstly, there is a lack of large-scale datasets in the research community for quadruped robots performing diverse tasks. Pretraining models require abundant trajectory data from a variety of tasks. However, collecting substantial amounts of real-world data often necessitates human involvement and significant time investment in manual instruction operations.

Secondly, different from navigation tasks [31] for general mobile robots and manipulation tasks for fixed-base systems [7, 9, 12, 34], building a VLA model

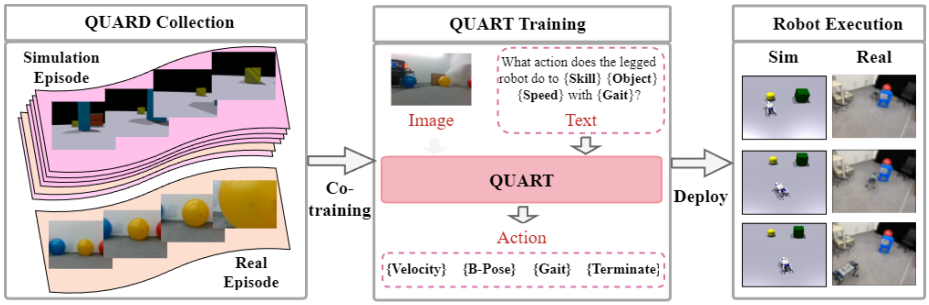

Fig. 2: Overview of QUAR-VLA. Our tasks encompass a diverse range of perception, navigation, and other advanced capability. The Vision-Language-Action (VLA) model first undergoes training with a huge amount of simulation data (246K episodes) and a small amount of real-world data (3K episodes). In the inference phase, images and texts undergo tokenization, after which **QUART** generates 12-dimensional action tokens. These tokens are subsequently detokenized into valid robot actions and deployed on a physical quadruped robot. This methodology effectively extends the learned capabilities from a simulated environment to real-world applications.

to solve complex quadruped robot tasks is considerably more challenging due to their agile locomotion behaviors. The action space needs to be properly defined to strike a balance between movement flexibility and computational efficiency. The action generated by the model should not be too simplistic, akin to the planner base velocities output by 2-D navigation modules, nor should they require a high execution frequency like the low-level motion policy that directly controls the joint motors.

To address these two problems, we collect a large-scale multi-task dataset **QUA**druped **R**obot **D**ataset (**QUARD**). It includes multiple tasks such as perception, navigation, and advanced capabilities like object avoidance. To the best of our knowledge, this is the first quadruped robot dataset that incorporates a significant amount of vision, language instruction, and robot command data. As collecting data on real robots is expensive and inefficient, we primarily rely on data generated in simulation, which exhibits significant differences in visual, sensor, and system dynamics. We also introduce **QUA**druped **R**obotic **T**ransformer (**QUART**), a VLA model for training **QUARD**. **QUART** takes images of the robot's first-view camera and the natural language instruction as inputs and generates rich control commands that include controlling the robot's base velocity, posture, and gait parameters. **QUART** leverages a pre-trained large-scale visual language model and fine-tunes it on our dataset to enable the generation of executable commands for quadruped robots. To address the sim-to-real gap caused by the data disparity, we construct a co-training pipeline to effectively distill the knowledge of simulation data for real-scene deployment. Our extensive evaluation shows that our approach leads to performant robotic policies and enables **QUART** to obtain a range of generalization capabilities.

Our contributions are as follows: 1) To the best of our knowledge, we first propose a new paradigm: **QUAR-VLA**, integrating visual and language instructions to executable actions for more autonomous and diverse quadruped robot tasks. 2) We present a large-scale multi-task dataset, **QUARD**, and a Vision-Language-Action model, **QUART** to solve the **QUAR-VLA** tasks. 3) Our extensive evaluation shows that our approach leads to performant robotic policies and enables **QUART** to obtain a range of generalization capabilities.

2 Related Work

Quadruped Robot Learning. Previous research [2,4,10,13,16,17,19,26,30,36, 39] has delved into single-model policies. As to vision-based methods (**QUAR-VA**), Kareer *et al*. [16] utilizes vision perception to predict a privileged terrain map, enabling the robot to perform specific locomotion based on the terrain characteristics. Karnan [17] leverages ego-motion estimates obtained through vision perception, allowing the robot to understand its relative position and direction to avoid collisions and navigate toward a goal location. While some works have succeeded in these tasks through vision perception, relying solely on vision limits the capability to handle only a specific type of task. In response to this limitation, Tang *et al*. [36] have started exploring language-conditioned tasks (**QUAR-LA**). However, the absence of vision perception in their approach results in a lack of autonomy for the robot to interact with its environment. Therefore, this paper proposes a new **QUAR-VLA** paradigm to integrate visual information and instructions to generate executable actions, effectively ensuring that the robot accurately interprets and acts upon detailed instructions in harmony with its visual observations.

Vision-language-action models. Visual-language-action models [1, 5, 20, 24, 28, 33, 35]integrates visual information and instructions to generate executable actions, effectively merging perception, planning, and decision-making and elevating the overall intelligence of the robot. As a result, vision-language-action models have been a popular area of research in robotics. In the realm of visual-language-action models, Shridhar *et al*. [33] have made significant strides by leveraging CLIP to encode text inputs, thereby enhancing the vision model's semantic comprehension and action execution capabilities. Further advancing the field, Brohan *et al*. [8] and Li *et al*. [20] have innovatively applied large language models to craft manipulation policies specifically for robotic applications. This paper delves into the domain of embodied agents, focusing on the quadruped robot as a case study. It explores the challenges and opportunities in enabling quadruped robots to autonomously navigate and perform a variety of tasks as directed by human instructions, contributing to the broader discourse on robotic autonomy and intelligence.

Robot Dataset. The realm of robotic learning has been at the forefront of advancing open-source datasets tailored explicitly for robot learning [6, 15, 17, 32, 36,38,40]. Within the domain of quadrupedal robots, previous works on datasets primarily focused on algorithms and leg controllers. Eckert *et al.* [11] introduced

a grading system and an online dataset for the collection and distribution of agility scores. Caluwaerts *et al.* [10] sought to address the lack of a standardized and scalable environment and inductive metrics. Tang *et al.* [36] pioneered the generation of desired joint position control using LLM. Nevertheless, common limitations persist among these quadruped robot datasets: the quantity is insufficient and the tasks are limited in diversity. Hence, our data repository aims to complement these endeavors. We curate and process an extensive array of skills on various objects, across diverse scenes in both the real world and virtual environments. The dataset includes image data, robot action data, and point cloud data, providing quadruped robots with rich perceptual information to tackle increasingly intricate and demanding tasks.

3 Method

This section provides a detailed exposition of our proposed methodology. Initially, we present the definition of our proposed **QUAR-VLA** in Section 3.1. Following that, Section 3.2 outlines the collecting process of **QUARD**. Lastly, we delve into the overarching architecture of **QUART** in Section 3.3.

3.1 Problem Setup

The objective of **QUAR-VLA** is to construct a vision-language-action model learned from large-scale demonstration data and generate actions for closed-loop robot control.

An overview of our **QUAR-VLA** is shown in Fig. 2. Our goal is to train a conditional policy **QUART** that can interpret RGB image(s), denoted $s \in \mathcal{S}$, together with a task instruction $w \in \mathcal{W}$, which correspond to a language string. The policy is a mapping from images and instructions to actions, and can be written as $\mu : \mathcal{S} \times \mathcal{W} \rightarrow \mathcal{A}$, where the action space $\mathcal{A}$ consists of the 11-dimensional high-level commands as well as a terminate signal.

$$\left[v_x, v_y, \omega_z, \theta_1, \theta_2, \theta_3, f, h_z, \phi, s_y, h_z^f, t\right] \tag{1}$$

Here, v_x, v_y, and ω_z represent the velocities along the x-axis, y-axis, and z-axis respectively. θ_1, θ_2, and θ_3 indicate the gait pattern, f denotes the frequency, h_z represents the height of the robot, ϕ denotes the pitch angle, s_y corresponds to the foot width, h_z^f represents the foot height, and t indicates the termination signal of the action.

Notably, in this study, we employ the discretization method proposed by [9] to discretize all continuous dimensions into 256 uniformly sized bins. This choice of discretization is motivated by the desire to reduce the complexity of action search and improve the stability and convergence of algorithms.

3.2 Large-scale Quadruped Robot Datasets

To enable an imitation learning system to generalize to new tasks with zero demonstrations of said task, we must be able to easily collect a diverse dataset,

Table 1: Tasks Definition. The "Type" means different capabilities of robots. The "Level" devides the difficulty into 3 levels. The "Skill" means different skill/task categories. The "Episode" signifies the number of experiments conducted for each task, which also corresponds to the number of trajectories. The "Description" is the illustration of the tasks.

Type	Level	Skill	Episode	Description
Perception	Easy	Distinguish *Letter* (sim)	10K	Identify the correct one from multiple boxes with different printed letters
Basic Navigation	Medium	Go to *Object* (sim)	72K	Navigate to the object and stop in front of it
	Medium	Go to Object (real)	3K	Navigate to the object and stop in front of it
Advanced Capability	Hard	Go through *Tunnel* (sim)	48K	**Spatial Navigation**: Go through the correct tunnel from two tunnels with different colors and shapes
	Hard	Go to *Object* and avoid the obstacle (sim)	63K	**Obstacle Avoidance**: Navigate to the object without colliding with the obstacle
	Hard	Crawl under *Bar* (sim)	1K	**Environment Adaptation**: Crawl under a bar with a low height
	Hard	Unload *Object* into *Receptacle* (sim)	52K	**Object Manipulation**: Move with a ball on the back and unload it into a receptacle
		Total	249K	The total number of episodes

provide corrective feedback, and evaluate many tasks at scale. Therefore, we collect a large-scale multi-task dataset, **QUA**druped **R**obot **D**ataset (**QUARD**), which includes multiple tasks such as perception, basic navigation, and advanced capability like object manipulation.

Task Definition. As is shown in Table 1, there are seven tasks distributed across three levels of difficulty: easy, medium, and hard. The easy and medium tasks are designed such that the robot can accomplish them using fundamental skills. For instance, the "Go to" task is a basic action integral to all navigation tasks, while the "Distinguish" task serves as the foundational action for all perception tasks. Hard tasks necessitate the robot to execute tasks using more advanced skills. A case in point is the "Unload object" task, which involves the robot distinguishing the target container, navigating to the target position, adjusting body position, and considering potential collisions to successfully unload the object. This process integrates perception, navigation, and whole-body manipulation. The complexity of tasks is reflected in the average trajectory length, with more challenging tasks requiring a greater number of steps and consequently taking longer to complete. The specific trajectory lengths and task distribution for each task are detailed in Fig. 3 and Fig. 4.

System Setup. The robot used to collect the trajectory data is WR-2, which is a quadruped robot with 12 joints. WR-2 has around 25cm in standing height and 40cm body length. The command output is sent to the low-level command tracking controller (pre-trained command-conditioned policy in [23]) to generate the actual joint action of the robot. The simulation data was collected in Nvidia's Isaac Gym [22], a powerful simulator that allows us to collect massive robot trajectories in parallel. The real data was collected by manually manipulating a quadruped robot in the lab setting. The perception data is provided by a RealSense d435 camera installed in the front of the robot.

Data Collection. As to simulation data, A* and D* algorithms are used to plot optimal paths for a dog navigating through various objects and obstacles. The

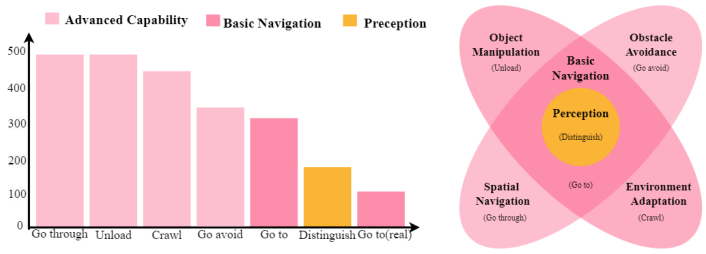

Fig. 3: The left figure illustrates the trajectory lengths corresponding to different tasks and the right figure illustrates the relationships between tasks. As the difficulty of the skill increases, it can be observed that the average trajectory length gradually increases. The right figure demonstrates the relationship between the types of tasks: perception as foundational capability; basic navigation ability is built on perception; object manipulation, obstacle avoidance, spatial navigation, and environment adaption extend from perception and navigation.

A* algorithm seeks the most cost-effective path, while the D* algorithm adapts to changes in real time. A Proportional-Derivative controller then converts these paths into target velocities for the dog, ensuring smooth movement. The combination of both algorithms provides a flexible and efficient method for path planning in various environments. For real data, The real data set is obtained in a laboratory environment using remote control.

Consistency constraints. To maintain consistency between simulation and reality, we've established constraints for data collection. The robot starts at the origin, with the target randomly positioned within [2.7,3.3] meters on the x-axis and [0.9,1.1] meters on the y-axis. In obstacle scenarios, obstacles are placed 1.5 meters from the target's x-coordinate, with the same y-coordinate. Task termination criteria vary. For "go to object", "go to the object and avoid obstacle", and "crawl under bar" tasks, success is when the robot is less than one meter from the target. Other task success criteria are: 1) "Unload object": successful when the object is in the target container. 2) "Go through tunnel" and "Crawl under bar": successful when the robot reaches a specific position behind the tunnel or bar. 3) "Distinguish letter": successful when the robot correctly orients towards the visual target.

More Statistics. The data collection setup adheres to a predefined language template as depicted in Fig. 4, where the task, target object, speed, and gait are explicitly defined. To enhance diversity and generalization capabilities, the dataset includes a variety of common indoor furniture and outdoor facilities, in addition to basic shapes. The basic shapes come in four color variations: green, red, blue, and yellow. Examples of indoor furniture include bookshelves, ovens, and vases, while outdoor facilities encompass trashcans and benches. Recognizing

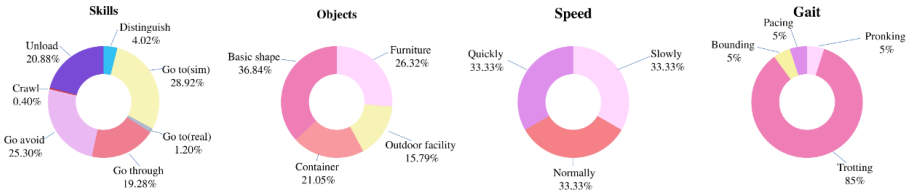

Fig. 4: Statistic analysis of **QUARD**. 1) Simulation data accounts for a larger proportion compared to real data. 2) Trotting occupied most of the share of gait. 3) Most tasks own similar episodes while Distinguish Letter and Go to Object (in reality) owns much less. 4) 3 types of speed occupied 1/3 separately.

the distinct nature of motion robots as compared to manipulation robots, we also require robots to perform the same task with varying gaits and speeds to adapt to different instructions and environments. The corresponding ratios of gaits and speeds are presented in Fig. 4. More details about the datasets can be found in the supplementary material.

3.3 Vision-Language-Action Model

In comparison to VLA models [7, 9, 12, 31, 34] in other field, our research stands out due to the abundance of kinematics information available in this domain. This allows the robot to not only excel at goal-oriented navigation tasks but also perform a diverse range of whole-body manipulation operations with enhanced flexibility in gaits and body control. Next, we will present our **QUA**druped **R**obotic **T**ransformers **(QUART)**.

QUART relies on a pre-trained vision-language-model [3] to associate tokens from the model's existing tokenization with the discrete action bins. It is worth noting that training MLLMs to override existing tokens with action tokens is a form of symbol tuning [37], which has been shown to work well in prior work. For [3], integers up to 1000 each have a unique token, so we simply associate the action bins with the token representing the corresponding integer. Notably, **QUART** model takes a single image s and a natural language instruction w as input, which are first converted into corresponding tokens t through a tokenizer $\tau(t|s, w)$ and fed into a decoder-only transformer module to obtain discretized action tokens $p(a_d|t)$.

The policy **QUART** could be shown as follow:

$$\text{QUART}(a_d|s, w) = p(a_d|t)\tau(t|s, w) \tag{2}$$

where w, s are the input images and language instruction and τ represents the tokenizer and p indicates the vision-language model to output action a_d.

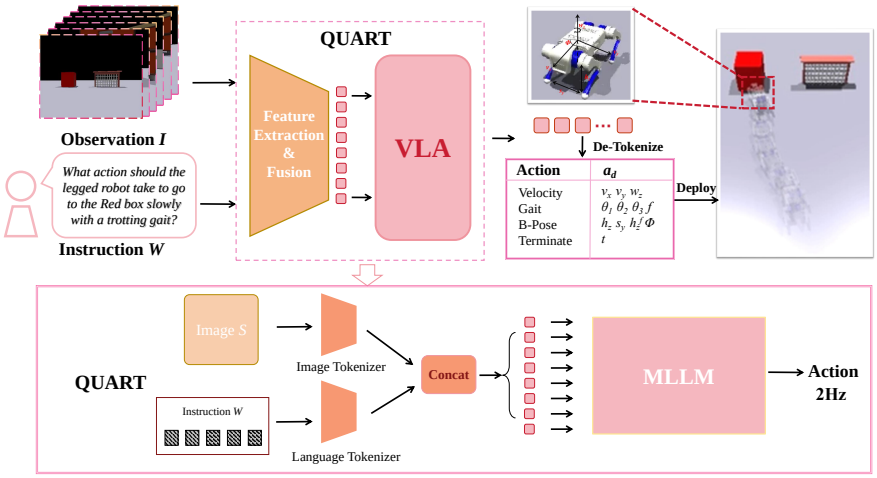

Fig. 5: Architecture of QUART. It is designed to leverage the scene comprehension capability of a pretrained MLLM. It receives visual information as observation, and outputs an action representing the actual action taken by the robot based on text-form instructions, and de-tokenizes it into specific action values. **QUART** can generate a complete action sequence at a processing rate of 2Hz in actual scenarios, and hand it over to the underlying low-level strategy for execution.

Action Detokenize. To directly convert models' output to valid robot actions for downstream control, we need detokenize the discrete action token a_d into continuous representation a_c (except for the discrete termination command).

$$a_c = \text{Detokenize}(a_d) \tag{3}$$

Loss. We use a standard categorical cross-entropy objective and causal masking that was utilized in prior Transformer-based controllers [18, 29].

Inference Speed. In contrast to many applications of large models, such as natural language or image generation, one of the unique requirements for a model that needs to run on real robots in real-time is fast and consistent inference speed. It is important to highlight that our research specifically emphasizes the command control of quadruped robots. In contrast to low-level motor control, the command control outputs alleviate the strict control frequency requirements, enabling smooth integration of larger models and unlocking improved reasoning capabilities. For **QUART**, the inference time could get 2Hz. More details about the deployment could be found in the supplementary material.

4 Experiments

We concentrate our experiments on the multi-task ability and generalization. We try to address the following questions: 1. How effective is the vision-language-

action architecture for multi-task quadruped task compared to previous VLM baselines? 2. How well do these models generalize to unseen semantic attributes like object shape, color, and unseen verbal information?

4.1 Implementation Details

Training Details. We train a specific instantiation of **QUART**, derived from 8B pre-trained VLM models [3] for superior performance. And we use learning rate 2e-5 and batch size 256 to fine-tune the model for 100K gradient steps. Both models are trained with the next token prediction objective, which corresponds to the behavior cloning loss in robot learning.
Evaluation Details. We follow the standard robot evaluation metrics [7, 9], success rate (SR), to evaluate the overall performance. The signal we used for robot control is the 11-dimensional command information with the action space outlined in Section 3.1 and 1 stop terminal command means the end of action. More details for each experiment can be seen in the supplementary material.
Baselines. Considering the absence of VLA models work on quadruped robots at present, we have taken the following baselines into account for a fair comparison: **CLIP** [27], **R3M** [25], and **VC-1** [21].

1. **CLIP** [27]uniquely encodes both textual and visual data, fusing these embeddings to create feature sets that integrate textual and visual information.

2. **R3M** [25]is a visual representation model derived from the Ego4D dataset, which acts as a unified, static perception module for policy learning applications.

3. **VC-1** [21]pushes the boundaries of visual representation learning by amalgamating data from multiple datasets and assessing performance across a diverse range of tasks via the implementation of CortexBench.

Since the aforementioned models do not include language conditioning, we include this aspect by separately embedding the language command, allowing us to compare it to our method. Specifically, we concatenate the resulting language embedding tokens with the image tokens generated by the vision model and pass the combined token sequences through a policy head to produce action outputs.

4.2 Overall Performance

Multi-task Performance. To assess the overall capabilities of VLA models on multiple tasks, we conduct evaluations using instructions randomly selected from the training set. It's important to note that this evaluation introduces variations in the placement of objects and other setup factors (such as robot position). This variation demands the system's ability to generalize effectively to realistic environmental variability. In total, over 1500 episodes are tested in this evaluation, comprising 425 episodes for going to objects, 500 for going to objects without colliding with the obstacle. 150 for going through the tunnel, 100 for unloading objects, 100 for distinguishing objects, and 75 for crawling under the bar.

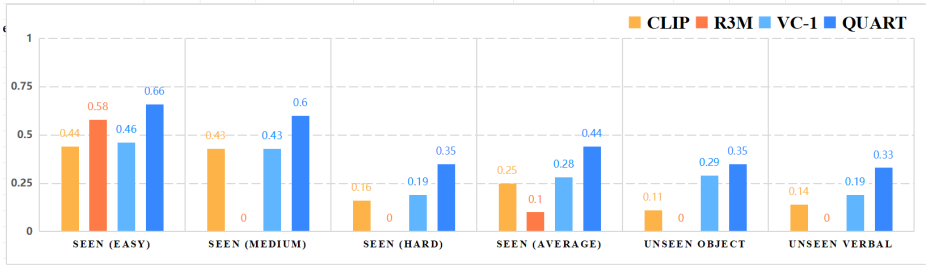

Fig. 6: Tasks Successful Rate. R3M lacks alignment with text, which means that although it has some recognition capabilities, it struggles to understand actions through instructions. CLIP and the VC-1 possess the ability to align text and images, enabling them to generate and execute actions based on instructions and observations.As shown above, while R3M may have a slight advantage in easy seen tasks (distinguish), it performs poorly in more complex tasks requiring action execution, significantly lagging behind CLIP and VC-1. **QUART** fully leverages the advantages of large models, enabling semantic understanding of images and alignment with textual information.

Table 2: Overall performance. QUART has achieved success rates far exceeding those of the baselines in tasks of all difficulty levels, especially in the most challenging crawl and unload tasks, where the baselines have no record of success.

	Seen						Unseen	
	Easy	Medium	Hard					
	Distinguish	Go to	Go avoid	Go through	Crawl	Unload	Object	Verbal
CLIP [27]	0.44	0.43	0.45	0.19	0	0	0.11	0.14
R3M [25]	0.58	0	0	0	0	0	0	0
VC-1 [21]	0.46	0.43	0.45	0.31	0	0	0.29	0.19
QUART	**0.66**	**0.60**	**0.53**	**0.41**	**0.32**	**0.12**	**0.35**	**0.33**

1. Comparison within VLM baselines. The experiment results reveal that R3M has poor performances on tasks except distinguish. The primary reason appears to be the lack of alignment with language semantics, which hinders its ability to comprehend tasks beyond basic discrimination. In contrast, CLIP and VC-1 demonstrate satisfactory performance on fundamental perceptual tasks, such as navigation (e.g., "go to" commands). However, their efficacy diminishes significantly when tasks require complex mechanical movements. This observation suggests that while visual language models (VLMs) can grasp abstract principles of the world, directly applying VLMs does not readily translate to the execution of intricate physical tasks.

2. VLM baselines vs QUART. As is shown in Table 2, Our model has achieved optimal performance on nearly all the baseline models. In our study, we have achieved a decoder-only VLA framework. This approach diverges from the traditional single-layer MLP policy head by leveraging the sequential nature

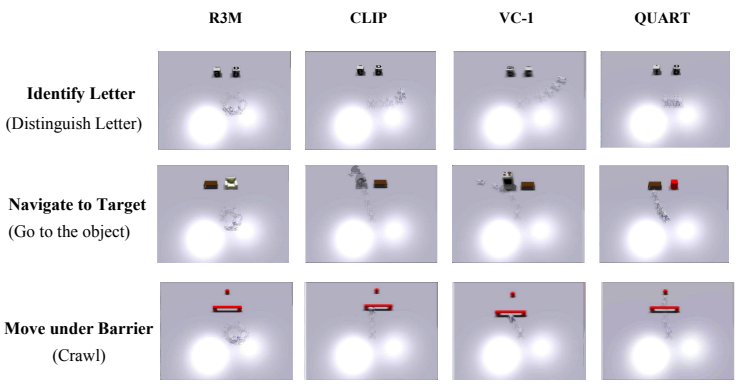

Fig. 7: Cases with unseen verbal instruction. Here are 3 verbal instructions. The words below represent the names of instructions in the training data, with the words above in bold indicating verbal instruction. When confronted with unseen instructions, the alighment between the existing language and the integration of vision and action cues within the baselines is compromised, resulting in task failure. This failure manifests in behaviors such as repetitive motion, misdirection, wrong terminate commands. Conversely, **QUART**, leveraging the language prowess inherited from large language models, adeptly achieves generalization under novel instructions, thereby effectuating the harmonization of vision, language, and action.

of action generation. This allows for the implicit learning of interdependencies between different action dimensions through the use of a transformer. Consequently, while the performance gains may be marginal in simple tasks, there is a noticeable enhancement in tasks that involve complex mechanical movements. significantly improved its perceptual capabilities by incorporating commonsense from the multi-modal large model (MLLM).

Generalization capabilities. To examine the adaptability in unanticipated scenarios, we orchestrated two primary examinations: one concentrated on unfamiliar objects, and the other on unprecedented linguistic directives. For the unfamiliar objects, we cast our gaze upon an array of circumstances: objects belonging to an identical category but exhibiting divergent textures and colors; objects from the same category but of disparate shapes; and objects with differences in shapes and textures, absent from current datasets. Turning towards the unprecedented linguistic instructions, we characterized unfamiliar instances as linguistic directives that bear the same semantics but vary in expression. For example, within the "*go to the object*" task, we gauged the directive's adaptability by employing "*navigate to target*" as the testing instruction.

1. Unseen Object. As depicted in Fig. 7, we can observe the enhanced generalization ability of our **QUART** model. For unseen objects, both **QUART** and VC-1 perform well, thanks to the advantages of the pre-trained VLM model. However, when it comes to unseen verbal descriptions, apart from **QUART**, all other models essentially failed. This is attributed to the inherent advantage pro-

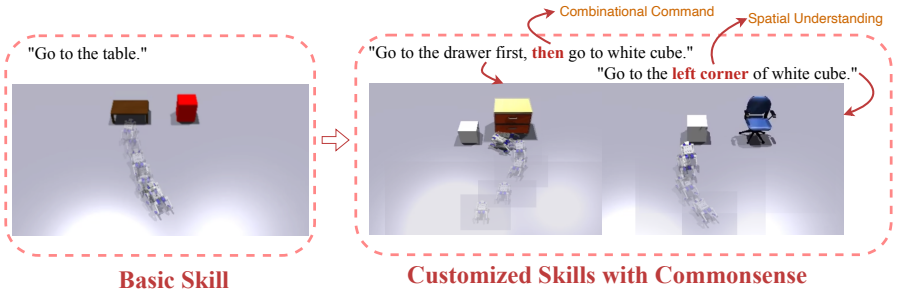

Fig. 8: QUART can extend basic skills to accommodate complex customized skills with commonsense. It shows example performances of combinational command and spatial understanding towards customized command.

vided by the pre-trained Multimodal Large Language Model (MLLM), enabling the recognition of distinct variations in object types and the comprehension of subtle semantic nuances. Consequently, the model exhibits exceptional generalization performance.

2. Unseen Verbal instruction. The training directives utilized were relatively monotonous, demonstrating that minor alterations in instruction can often perplex the model, thereby revealing an imperative for generalization across language directives. Initially, the difficulties introduced by merely replacing descriptions of skill for task generalization were investigated. Evident from Table 2 and Fig. 7 is that **QUART** surpasses other baseline methods, an innate advantage of large-scale multimodal models. In addition, to explore more intriguing linguistic directives and fully harness the universal capabilities of the Large Language Model (LLM), Fig. 8 presents instances where the model demonstrates a clear comprehension and application of combined directives constructed via "first" and "then". It was further noted that even in the absence of explicit directional information within the dataset, the model reliably interpreted linguistic directives which, in turn, influenced task execution accuracy. The ability to utilize language for precise control of quadruped robots facilitated by scene perception, as showcased in Fig. 8, is of immense significance. More intriguing examples and analyses can be found in the supplementary materials.

4.3 Qualitative Performance

Here, we offer the visualization results of seen and unseen tasks in Fig. 9 with WR-2. We also test QUART in other quadruped robots like Go2 with minor adjustments. More demos can be found in the supplementary materials.

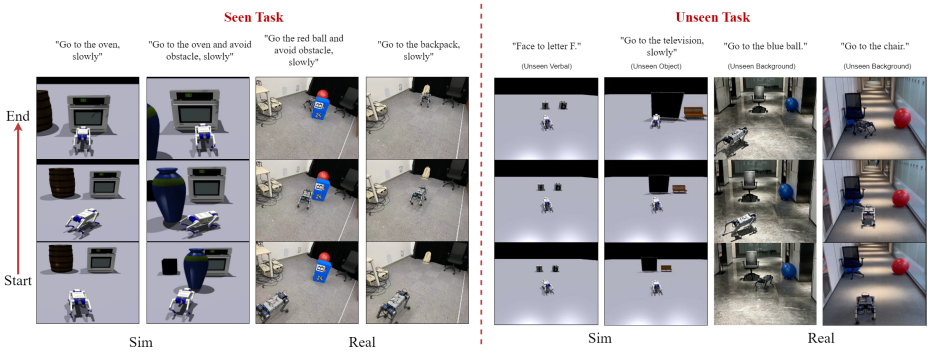

Fig. 9: Visualization of seen and unseen tasks. This figure demonstrates the excellent performance of **QUART** for seen tasks in both simulated and real-world environments. The left part shows 4 instances of seen tasks in both simulation environment and real-world. The right part shows performances on unseen tasks, which contain unseen verbal, object, and background.

5 Conclusion & Future Work

This paper emphasizes the significance of deploying Vision-Language-Action models on quadruped robots. We introduce the concept of **V**ision-**L**anguage-**A**ction tasks for **QUA**druped **R**obots **(QUAR-VLA)**, which seamlessly integrates visual information and instructions from diverse modalities to generate executable actions. Specifically, **QUAR-VLA** focuses on two main aspects of deploying VLA models for quadruped robots: defining the action space to balance flexibility and efficiency, and addressing the scarcity of large-scale training datasets. To tackle these two questions, we present the **QUART** models tailored for quadruped robots and the **QUARD** dataset, which includes diverse tasks such as navigation and manipulation. Our extensive evaluation shows that our approach leads to performant robotic policies and enables **QUART** to obtain a range of emergent capabilities. This includes generalization to novel objects, the ability to interpret commands not present in the robot training data, and the ability to perform rudimentary reasoning in response to user commands.

Furthermore, this work has these areas requiring further enhancements: **1) Automated Data Collection for Larger Datasets.** Future work can explore automated processes for data collection, involving techniques such as active learning and data augmentation can be employed to further expand the dataset size and diversity. **2) Better solution for sim2Real gap.** Integrating domain adaptation methods like domain randomization, data augmentation, and more accurate modeling of physics engines will be a better solution compared with direct co-training. **3) Improving Inference Speed for Real-time Control.** Future works will explore hardware acceleration techniques and model compression techniques to enable faster and more efficient execution of the models.

Acknowledgements

This work was supported by the National Science and Technology Innovation 2030 - Major Project (Grant No. 2022ZD0208800), and NSFC General Program (Grant No. 62176215).

References

1. Ahn, M., Dwibedi, D., Finn, C., Arenas, M.G., Gopalakrishnan, K., Hausman, K., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kirmani, S., Leal, I., Lee, E., Levine, S., Lu, Y., Leal, I., Maddineni, S., Rao, K., Sadigh, D., Sanketi, P., Sermanet, P., Vuong, Q., Welker, S., Xia, F., Xiao, T., Xu, P., Xu, S., Xu, Z.: Autort: Embodied foundation models for large scale orchestration of robotic agents (2024)
2. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from human videos as a versatile representation for robotics (2023)
3. Bavishi, R., Elsen, E., Hawthorne, C., Nye, M., Odena, A., Somani, A., Taşırlar, S.: Introducing our multimodal models (2023), `https://www.adept.ai/blog/fuyu-8b`
4. Belkhale, S., Cui, Y., Sadigh, D.: Hydra: Hybrid robot actions for imitation learning. In: Proceedings of the 7th Conference on Robot Learning (CoRL) (2023)
5. Bharadhwaj, H., Vakil, J., Sharma, M., Gupta, A., Tulsiani, S., Kumar, V.: Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking (2023)
6. Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., et al.: Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 4243–4250. IEEE (2018)
7. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M.G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.W.E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: Rt-2: Vision-language-action models transfer web knowledge to robotic control (2023)
8. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818 (2023)
9. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: Rt-1: Robotics transformer for real-world control at scale (2023)

10. Caluwaerts, K., Iscen, A., Kew, J.C., Yu, W., Zhang, T., Freeman, D., Lee, K.H., Lee, L., Saliceti, S., Zhuang, V., et al.: Barkour: Benchmarking animal-level agility with quadruped robots. arXiv preprint arXiv:2305.14654 (2023)
11. Eckert, P., Ijspeert, A.J.: Benchmarking agility for multilegged terrestrial robots. IEEE Transactions on Robotics **35**(2), 529–535 (2019)
12. Gu, J., Kirmani, S., Wohlhart, P., Lu, Y., Arenas, M.G., Rao, K., Yu, W., Fu, C., Gopalakrishnan, K., Xu, Z., Sundaresan, P., Xu, P., Su, H., Hausman, K., Finn, C., Vuong, Q., Xiao, T.: Rt-trajectory: Robotic task generalization via hindsight trajectory sketches (2023)
13. Haldar, S., Mathur, V., Yarats, D., Pinto, L.: Watch and match: Supercharging imitation with regularized optimal transport. CoRL (2022)
14. Hutter, M., Gehring, C., Jud, D., Lauber, A., Bellicoso, C.D., Tsounis, V., Hwangbo, J., Bodie, K., Fankhauser, P., Bloesch, M., Diethelm, R., Bachmann, S., Melzer, A., Hoepflinger, M.: Anymal - a highly mobile and dynamic quadrupedal robot. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 38–44 (2016). https://doi.org/10.1109/IROS.2016.7758092
15. Jiang, Y., Moseson, S., Saxena, A.: Efficient grasping from rgbd images: Learning using a new rectangle representation. In: 2011 IEEE International conference on robotics and automation. pp. 3304–3311. IEEE (2011)
16. Kareer, S., Yokoyama, N., Batra, D., Ha, S., Truong, J.: Vinl: Visual navigation and locomotion over obstacles. 2023 IEEE International Conference on Robotics and Automation (ICRA) pp. 2018–2024 (2022), https://api.semanticscholar.org/CorpusID:253117178
17. Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., Stone, P.: Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. IEEE Robotics and Automation Letters **7**(4), 11807–11814 (2022)
18. Lee, K.H., Nachum, O., Yang, M.S., Lee, L., Freeman, D., Guadarrama, S., Fischer, I., Xu, W., Jang, E., Michalewski, H., et al.: Multi-game decision transformers. Advances in Neural Information Processing Systems **35**, 27921–27936 (2022)
19. Lee, M.A., Zhu, Y., Srinivasan, K., Shah, P., Savarese, S., Fei-Fei, L., Garg, A., Bohg, J.: Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In: 2019 IEEE International Conference on Robotics and Automation (ICRA) (2019), https://arxiv.org/abs/1810.10191
20. Li, X., Liu, M., Zhang, H., Yu, C., Xu, J., Wu, H., Cheang, C., Jing, Y., Zhang, W., Liu, H., et al.: Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:2311.01378 (2023)
21. Majumdar, A., Yadav, K., Arnaud, S., Ma, Y.J., Chen, C., Silwal, S., Jain, A., Berges, V.P., Abbeel, P., Malik, J., Batra, D., Lin, Y., Maksymets, O., Rajeswaran, A., Meier, F.: Where are we in the search for an artificial visual cortex for embodied intelligence? (2023)
22. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., State, G.: Isaac gym: High performance gpu-based physics simulation for robot learning (2021)
23. Margolis, G.B., Agrawal, P.: Walk these ways: Tuning robot control for generalization with multiplicity of behavior. In: Liu, K., Kulic, D., Ichnowski, J. (eds.) Proceedings of The 6th Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 205, pp. 22–31. PMLR (14–18 Dec 2023)
24. Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3m: A universal visual representation for robot manipulation. In: Conference on Robot Learning (2022), https://api.semanticscholar.org/CorpusID:247618840

25. Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3m: A universal visual representation for robot manipulation (2022)
26. Pari, J., Shafiullah, N.M., Arunachalam, S.P., Pinto, L.: The surprising effectiveness of representation learning for visual imitation (2021)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
28. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T., et al.: A generalist agent. arXiv preprint arXiv:2205.06175 (2022)
29. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T., et al.: A generalist agent. arXiv preprint arXiv:2205.06175 (2022)
30. Schiavi, G., Wulkop, P., Rizzi, G., Ott, L., Siegwart, R., Chung, J.J.: Learning agent-aware affordances for closed-loop interaction with articulated objects. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 5916–5922 (2023). https://doi.org/10.1109/ICRA48891.2023.10160747
31. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: Vint: A foundation model for visual navigation (2023)
32. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: Proceedings Shape Modeling Applications, 2004. pp. 167–178 (2004)
33. Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation. In: Conference on Robot Learning. pp. 894–906. PMLR (2022)
34. Stone, A., Xiao, T., Lu, Y., Gopalakrishnan, K., Lee, K.H., Vuong, Q., Wohlhart, P., Kirmani, S., Zitkovich, B., Xia, F., Finn, C., Hausman, K.: Open-world object manipulation using pre-trained vision-language models (2023)
35. Szot, A., Schwarzer, M., Agrawal, H., Mazoure, B., Talbott, W., Metcalf, K., Mackraz, N., Hjelm, D., Toshev, A.: Large language models as generalizable policies for embodied tasks. arXiv preprint arXiv:2310.17722 (2023)
36. Tang, Y., Yu, W., Tan, J., Zen, H., Faust, A., Harada, T.: Saytap: Language to quadrupedal locomotion. arXiv preprint arXiv:2306.07580 (2023)
37. Wei, J., Hou, L., Lampinen, A., Chen, X., Huang, D., Tay, Y., Chen, X., Lu, Y., Zhou, D., Ma, T., et al.: Symbol tuning improves in-context learning in language models. arXiv preprint arXiv:2305.08298 (2023)
38. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
39. Yang, R., Zhang, M., Hansen, N., Xu, H., Wang, X.: Learning vision-guided quadrupedal locomotion end-to-end with cross-modal transformers. ArXiv **abs/2107.03996** (2021), https://api.semanticscholar.org/CorpusID:235765481
40. Yu, K.T., Bauza, M., Fazeli, N., Rodriguez, A.: More than a million ways to be pushed. a high-fidelity experimental dataset of planar pushing. In: 2016 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 30–37. IEEE (2016)