Supplementary material: 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

1 Abstract

In this supplementary material, we will provide the following supplementary information:

Extended Data Collection Details: We will present additional details regarding the data collection process, including information about the specific environments used and the criteria for determining successful data acquisition for different tasks.

Enhanced Experimental Results: We will provide more comprehensive information about the experimental setup, including a comparative analysis of the results obtained in both seen and unseen environments. Furthermore, we will compare the performance of different model architectures to provide a more thorough evaluation.

Expanded Deployment Results: To facilitate real-world deployment, we will present additional experimental results showcasing the model's performance across a broader range of tasks and scenarios.

Extended Visualizations: We will include supplementary visual results that highlight various failure cases observed in both real-world and simulation settings, offering a more comprehensive understanding of the model's limitations and areas for improvement.

2 Details of data collection

Environment: We prioritize the richness of tasks. In terms of environmental setup, our current tasks are all carried out in relatively simple scenarios without complex visual background information for data collection and model evaluation. At the same time, all experiments are conducted on flat terrains, and we have not yet conducted experiments on complex terrains. Our next goal is to use simulation environments with higher visual fidelity, and add some backgrounds

⋆ Corresponding author

that are more in line with real-world scenarios and more varied terrains for data collection.

Planning Algorithm: In simulation scenarios, we adopted a large-scale parallel simulation environment for the need of rapid automated data collection. During data collection, we used traditional path search algorithms to implement the robot's route selection. In the early stage of data collection, we adopted the D* algorithm. After evaluating the quality of the early data, we found that the paths planned by D* tend to generate curves with larger turning angles. This could potentially lead to the robot frequently losing sight of the target object, thus affecting the quality of the strategy obtained by imitation learning. Therefore, we changed the navigation algorithm to A*. Taking the go_avoid task as an example, we directly access the positions of the robot, the target object, and obstacles through the simulator. The initial centroid coordinates of the robot on the x-y plane are used as the starting point of the path, and the centroid coordinates of the target object are used as the endpoint. We then add the areas of obstacles on the map. After the path planning is completed, a PD controller is used to convert the path into current velocity and direction. Since we set the speed of the path collection to three levels (fast, normal, slow), the obtained speed needs to be processed to roughly satisfy the speed range set by the levels.

Excution Details: In control systems, the combination of high-frequency and low-frequency control is a strategic approach to achieve finer control and optimize system performance. This integration can be described by a simple mathematical relationship: $N = f_{high}/f_{low}$, where f_{high} represents the high-frequency control rate, f_low represents the low-frequency control rate, and N is the ratio between the two. In this equation, N indicates the number of times the high-frequency control needs to be executed within a given low-frequency period. The advantage of this combination is that it allows the system to achieve rapid responses and finer adjustments while maintaining the stability and efficiency of the low-frequency control. High-frequency control is typically used to address the rapid changes in system states, while low-frequency control focuses more on long-term system stability and energy management.

3 More experiments

3.1 Detailed results on seen tasks

Firstly, we supplement two additional frameworks: **VLA(LLaVa)** (which uses LLaVa as the backbone for multi-modal large models) and **Aligned VLM+P (Transformer)** (policy head is a decoder-only transformer) to validate the rationality of our VLA model framework design; Next, we provide the results trained on a single task to verify the performance advantages on multi-task; Then, we present more granular results for each task.

Two additional frameworks:

1. Experiment Setting:

Table 1: Detailed performance on different architecture.

		Distinguish	Go to	Go avoid	Go through	Crawl	Unload
VLM+P(MLP)	CLIP [3]	0.44	0.43	0.45	0.19	0	0
VLM+P(MLP)	VC-1 [2]	0.46	0.43	0.45	0.31	0	0
VLM+P(Transformer)	RT-1 [1]	0.22	0.15	0.4	-	0	0
VLA(Fuyu)	**QUART**	**0.66**	**0.60**	**0.53**	**0.41**	**0.32**	**0.12**
VLA(LLaVa)	**QUART**	**0.66**	**0.52**	**0.40**	**0.32**	**0.18**	**0.16**

1) **VLM baseline**: In the main paper, we reproduced previous baselines based on VLM by utilizing features extracted from VLM in conjunction with a policy head to generate actions. However, the policy head at that time was merely in the form of an MLP, called **VLM+P(MLP)**, and we did not explore the differences between various types of policy heads, particularly the decoder-only transformer (**VLM+P(Transformer)**).Therefore, we have additionally supplemented this study with different policy-head layers to ascertain whether the limitations of the VLM+Policy approach are related to the design of the policy head.

2) **VLA architecture**: Furthermore, in the paper, we experimented with a multi-modal large model based on the fuyu-8b model(**VLA(Fuyu)**) to verify whether the effectiveness of VLA is related to the choice of multi-modal large model foundations. To this end, we have also presented experiments based on the LLaVa(**VLA(LLaVa)**) to validate the benefits of a VLA approach grounded in multi-modal large models.

2. Experiment Analysis:

1) VLM + Policy(MLP): In the case of CLIP and VC-1, the visual and textual features have been aligned, enabling the models to comprehend and execute simple tasks. They perform reasonably well on tasks such as "go to," "go through," and "go avoid," which do not involve manipulation of the robot's body. The primary reason for this adequate performance is that these tasks only require changes in velocity along the x-axis and yaw orientat.

2) VLM + Policy(Transformer): As is shown in Table 1, we have also referred to the RT-1 paradigm, employing different policy heads to ascertain whether the limitations are inherent to the MLP architecture. We can observe that even when the policy is switched to a Decoder-only Transformer, the trend of the RT-1 method remains consistent with the previous VLM+P (MLP) approach. Only tasks that involve simple distinction and those related to the velocity of the aircraft's center of mass have success rates; tasks such as crawl and unload are still not achievable. This demonstrates that the choice of different policy heads does not affect the performance of the VLM+policy paradigm.

4)Vision + Language + Action (VLA): Within the VLA framework, we have utilized the entire decoder-only VLM backbone. However, we directly map action instructions to the language space. During inference, each dimension of our action (e.g., leg width) engages in joint reasoning with previously inferred information (e.g., robot height). This approach allows for the implicit learning of associations between different action dimensions, thereby effectively grasping

Table 2: Multi-task performance vs Single-task performance.

	Distinguish	Go to	Go avoid	Go through	Crawl	Unload
CLIP-Multi [3]	0.44	**0.43**	**0.45**	**0.19**	0	0
CLIP-Single [3]	**0.52**	0.34	0.37	0.04	0	0
VC-1-Multi [2]	0.46	**0.43**	**0.45**	0.31	0	0
VC-1-Single [2]	**0.70**	0.37	0.40	**0.34**	0	0
QUART-Multi	**0.66**	**0.60**	**0.53**	**0.41**	**0.32**	**0.12**
QUART-Single	0.30	0.36	0.19	0.30	0.25	0.08

the coordinated relationships between multiple parts of the robot and performing well on more complex tasks (e.g., crawling).

To investigate whether different multi-modal large models (MMLMs) affect our model's performance, we introduce two variants: Fuyu-8B(used in main paper) and LLaVa-7B. The primary distinction between the two lies in the fact that the former encodes the original image directly, while the latter employs the widely used visual CLIP feature extraction module. From the results, we can see that there is not much difference in performance corresponding to which base model is used. This indicates the importance of the VLA paradigm.

Multi-task vs Single-task Performance: To validate the performance of our multi-task learning approach, we conducted separate training on each individual task to ascertain the benefits that multi-task learning confers on the interrelated tasks. As is shown in Table 2, It has been observed that, for both single and multi-task scenarios, the performance of multi-task training has yielded superior results in all but the simplest tasks(distinguish). This indicates that the paradigm of joint training in multi-task settings has enabled the learning of commonalities between different tasks, thereby underscoring the necessity of multi-task co-training.

Detailed Performance: As is shown in Table 3, we present detailed results for the seen tasks.

3.2 Detailed results on unseen tasks

In the context of unseen tasks, we conducted experiments to assess the model's performance on novel objects and unseen language instructions. The novel objects are categorized into three types: objects of the same category but with different shapes; objects of the same shape but with different colors; and entirely different objects. The unseen language instructions involve paraphrasing existing descriptions with synonymous terms to test the robustness of the model's performance.

Detailed Performance on unseen objects: In the experiments, we use the yellow, red, green and blue as four base color in the seen tasks, and use the gold, pink, orange, purple as the unseen color. For each objects appear in the seen tasks, we all test another object which is the same type but with different shape. We also test on objects which do not appear in the seen tasks: pillow, computer and window. As is shown in Table 4, we present detailed results for the unseen

Table 3: Detailed results on seen tasks.

	Distinguish letter c	Distinguish letter d	Go to cooker
CLIP [3]	0.36	0.52	0.36
VC-1 [2]	0.48	0.44	0.36
QUART	**0.76**	**0.56**	**0.72**
	Go to ball	Go to cube	Go to oven
CLIP [3]	0.56	0.24	**0.56**
VC-1 [2]	**0.64**	0.40	0.32
QUART	0.60	**0.64**	0.44
	Go avoid cooker	Go avoid drawers	Go avoid fan
CLIP [3]	0.44	**0.52**	0.28
VC-1 [2]	0.44	**0.52**	0.20
QUART	**0.44**	0.48	**0.36**
	Go avoid sofa	Go through triangle tunnel	Go through rectangle tunnel
CLIP [3]	0.56	0.04	0.24
VC-1 [2]	0.64	**0.40**	0.28
QUART	**0.84**	0.24	**0.47**
	Crawl gate	Unload traybox	Average
CLIP [3]	0	0	0.25
VC-1 [2]	0	0	0.28
QUART	**0.32**	**0.04**	**0.44**

Table 4: Detailed results on unseen objects.

	Go to	Go avoid	Go through	Crawl	Unload
CLIP [3]	0.4	0.46	0.19	0.04	0
VC-1 [2]	0.38	0.41	0.36	0	0
QUART	**0.4**	**0.73**	**0.41**	**0.35**	**0.01**

Table 5: Detailed results on unseen verbal information.

	Identify Letter	Navigate to target
CLIP [3]	0.40	0.44
VC-1 [2]	0.28	0.48
QUART	0.40	**0.52**
	Move under barrier	Deposit object into container
CLIP [3]	0.12	0
VC-1 [2]	0	0
QUART	**0.28**	**0.04**

objects. We can see that our method have more generalization ability in unseen objects.

Detailed Performance on unseen verbal instruction: Here are alternative expressions for the tasks while maintaining the same meaning. Within the parentheses are the instructions for the seen tasks, followed by the modified instructions. 1. (Distinguish Letter) Identify Letter 2. (Go to the object) Navigate to target 3. (Crawl under the barrier) Move under barrier 4. (Unload the ob-

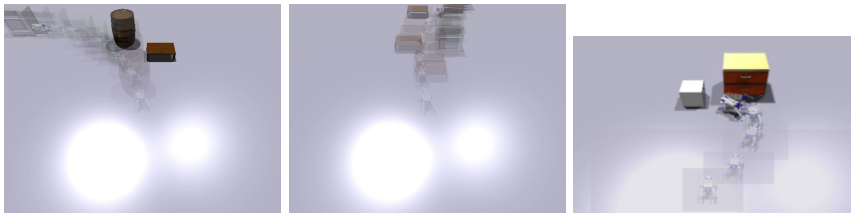

Fig. 1: Mission *go to the left corner of the object*. The left picture is produced by model CLIP. The middle picture is produced by model VC-1. The right picture is produced by **QUART**.

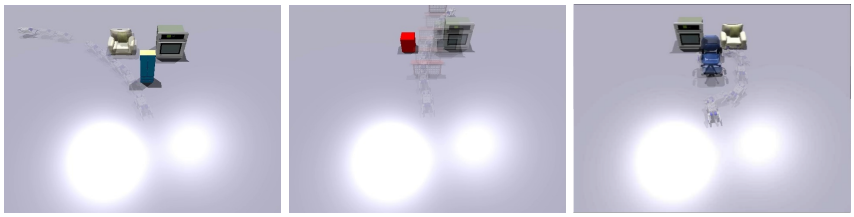

Fig. 2: Mission *go to the back of the object*. The left picture is produced by model CLIP. The middle picture is produced by model VC-1. The right picture is produced by **QUART**.

ject) Deposit Object into container As is shown is Table 5, We can observe that with instructions that carry the same semantics but different expressions, the performance of QUART significantly surpasses that of the baseline.

3.3 More results of customized skills

In the manuscript, we demonstrate the capability of our model to generalize to customized skills that were not present in the training tasks, such as complex spatial perception and the ability to combine tasks. Herein, we will present additional case studies to illustrate this skill further.

Case1: Go the left corner of the object. Figure 1
Case2: Go the the back of object. Figure 2
Case3: Go the the left and the to the right. Figure 3

In these cases, we could find our model could understand the spatial relationships and have the ability to excute combinational skills.

4 More analysis on real robot excution

We show the real robot experiments from the following 5 aspects: 1. Effectiveness in seen scenes 2. Sim2Real transfer capabilities 3. Rubustness in different localization 4. Rubustness in different workspace 5. Rubustness in unseen scenes

More results can be found in https://sites.google.com/view/quar-vla/quar-vla-eccv24.

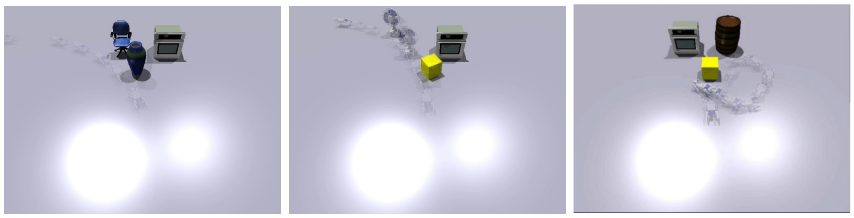

Fig. 3: Mission *go to the right and the left*. The left picture is produced by model CLIP. The middle picture is produced by model VC-1. The right picture is produced by **QUART**.

5 Limitation and future work

From the perspective of dataset composition, the current magnitude of data and the diversity of trajectories are insufficient. A potential direction for future work is to leverage GPT-based automation to generate more varied and enriched datasets. Additionally, the existing datasets lack complex terrains and long-horizon tasks; enhancing the complexity of the dataset is a crucial method for advancing quadrupedal tasks.

In terms of task formulation, the current input modalities are limited to visual and textual components. Exploring how to utilize additional modalities (e.g., LiDAR point clouds) to address issues that visual information alone cannot resolve, such as occlusion problems, is a direction worthy of investigation. Furthermore, the existing models are not yet capable of more flexible control in terms of frequency. Although the high-level command action frequency can complete some tasks, more challenging tasks, such as pole crossing, require higher frequency control to achieve higher success rates. Therefore, accelerating the base model's speed and designing a reasonable sampling mechanism for high-frequency output is an essential component.

Of course, addressing the sim2real gap is key to effectively utilizing real-world data. The co-training approach adopted in this paper is based on the premise that the sim2real gap for large models is not significant. However, how to more efficiently employ various sim2real methods, such as domain adaptation and randomization, to solve the domain gap between the real and simulated domains is also a line of thought worth exploring. Lastly, given the substantial amount of sub-optimal data present in the data collection process, how to utilize this data and enable the large model to learn valuable knowledge from failures through reinforcement learning is an important future direction.

In summary, this is the inaugural work in extending multi-modal large models to quadrupedal robots. In response to the existing challenges of quadrupedal robots, we have designed a dataset that combines extensive simulated data with a small amount of real data for quadrupedal robot VLA and developed a framework based on large models to implement this task. This work has a certain catalytic effect on the development of the robotics community, and we hope for

more suggestions to further refine this work in the future, thereby advancing the progress of mobile robotics.

References

1. 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)
2. 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)
3. 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)