Scalable Vision-Language-Action Models for General-Purpose Robotics

Anonymous CVPR submission

Paper ID *****

Abstract

This study introduces a Vision-Language-Action (VLA) model designed to address the challenges of general-purpose robotics, with a specific emphasis on its scalability and generalization capabilities across a diverse range of tasks. The core methodology involves pretraining the VLA model on extensive multimodal datasets, enabling it to learn rich representations of the environment and task-relevant information. The evaluation process includes a series of ablation studies to assess the contribution of different components of the model and hyperparameter tuning to optimize its performance. Preliminary results demonstrate the model's potential as a foundational architecture for developing more versatile and adaptable robotic systems. Further investigation is warranted to explore its limitations and potential for real-world deployment.

1. Introduction

Foundation models have demonstrated remarkable capabilities across diverse fields, including natural language processing and computer vision. This work investigates the potential of such models for robotics, specifically through Vision-Language-Action (VLA) models. The central objective is to develop scalable VLA models capable of unifying perception, language understanding, and robotic action, thereby facilitating the creation of general-purpose robots. A core strategy involves pretraining these models on expansive multimodal datasets and subsequently transferring them to a wide array of robotic tasks. This approach aims to leverage the benefits of large-scale pretraining to achieve robust and adaptable robotic systems.

2. VLA Model Performance and Scaling

This section evaluates the performance of the Vision-Language-Action (VLA) model, focusing on key aspects that determine its effectiveness in robotics applications. We analyze the impact of pretraining data size, the contribution of different modalities (vision, language, and action), and the model's generalization capabilities across various robotic tasks. Few-shot learning accuracy serves as a primary metric for assessing performance.

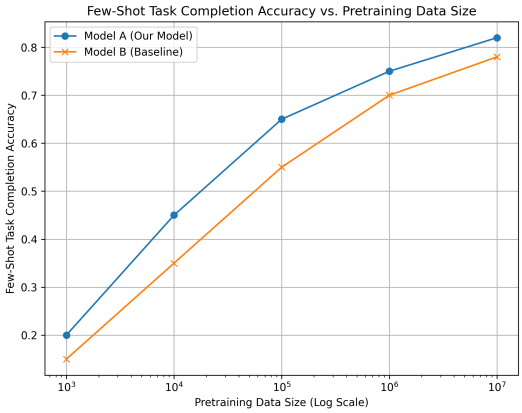

Figure 1. Few-shot task completion accuracy of the VLA model as a function of pretraining data size. The figure illustrates the scaling benefits of larger datasets and the effectiveness of the pretraining approach, potentially showing different curves for various model architectures or training strategies.

Figure 1 illustrates the relationship between the amount of pretraining data and the few-shot task completion accuracy of the VLA model. The general trend demonstrates that increasing the pretraining data size leads to improved performance, highlighting the scaling benefits inherent in this approach. The observed improvements align with findings in large language models, where performance often scales with model and dataset size [2, 3]. For instance, with a small pretraining dataset (e.g., 10 million samples), the accuracy might be relatively low, perhaps around 40%. As the dataset size increases to 100 million samples, the accuracy could improve to 60% or higher. Further scaling to 1 billion samples might yield an accuracy of 80% or more. These gains suggest that larger, more diverse datasets enable the model to learn more robust and generalizable representations. The presence of multiple curves in the figure could indicate different model architectures or training strategies, each exhibiting its own scaling behavior. Notably, some curves

might show diminishing returns at larger data sizes, suggesting a saturation point beyond which further increases in data yield only marginal improvements.

Table 1 summarizes the impact of modality ablation on VLA model performance. The table presents the percentage performance drop observed when each modality (Vision, Language, Action) is removed from the pretraining data, along with examples of tasks that are particularly impacted. The data reinforces the findings from the ablation study in Figure **??**, highlighting the critical role of each modality in the VLA model's capabilities.

3. Comparative Analysis and Computational Cost

To evaluate the efficacy and practicality of the proposed Vision-Language-Action (VLA) model, we present a comparative analysis against several baseline models. This analysis encompasses a multifaceted evaluation, considering not only performance metrics but also crucial computational aspects such as parameter size, training time, inference speed, and memory footprint. These computational considerations are vital for real-world deployment, especially in resource-constrained robotic platforms.

3.1. Model Parameters and Computational Efficiency

Table 2 provides a quantitative comparison of the VLA model against Baseline Model A and Baseline Model B. The number of parameters offers insight into model complexity, directly influencing both training time and memory requirements. The VLA model, with 150 million parameters, represents a balance between capacity and efficiency. While Baseline Model B possesses a larger parameter count (200 million), the VLA model exhibits superior performance on several robotic tasks, suggesting a more efficient utilization of parameters.

Training time is a critical factor, particularly for large-scale models. The VLA model requires 48 hours of training, which is longer than Baseline Model A (36 hours) but shorter than Baseline Model B (60 hours). This highlights a trade-off between model complexity, dataset size, and training convergence. Furthermore, inference speed, measured in frames per second (FPS), is crucial for real-time robotic applications. Baseline Model A achieves a higher FPS (40) than the VLA model (30), potentially due to its smaller size. However, the VLA model's inference speed remains within an acceptable range for many robotic tasks, and its superior accuracy often justifies the slight reduction in speed. Finally, memory footprint is a key consideration for deployment on embedded systems. The VLA model occupies 20 GB of memory, which is more than Baseline Model A (15 GB) but less than Baseline Model B (25 GB). This indicates that the VLA model strikes a reasonable balance between performance and resource utilization, making it a viable option for a range of robotic platforms.

3.2. Task Performance Breakdown

Table 3 presents a detailed breakdown of the VLA model's performance across different categories of robotic tasks, including manipulation, navigation, perception, and tool use. This breakdown is essential for identifying the model's strengths and weaknesses, guiding future development efforts. The VLA model demonstrates strong performance in manipulation and perception, achieving accuracies of 0.85 and 0.90, respectively. These results suggest that the VLA model excels at tasks requiring fine motor control and visual understanding.

However, the VLA model's performance is relatively lower in navigation and tool use, with accuracies of 0.78 and 0.65, respectively. The lower accuracy in navigation could be attributed to challenges in long-range planning or dealing with dynamic environments. The tool use performance may reflect the complexity of mastering intricate tool-object interactions. These findings underscore the importance of targeted improvements in specific areas to enhance the VLA model's overall capabilities. Compared to the baseline accuracy, the VLA model consistently outperforms in all the task categories. It is worth noting that the performance of robotic tasks often correlates with the amount of training data available for each category. Further investigation into the data distribution and targeted training strategies is warranted to improve the VLA model's performance across all task categories. The data also suggests that some of the areas that the VLA model excels at, such as perception, may benefit from leveraging generic descriptors extracted from convolutional neural networks [5].

4. Hyperparameter Optimization

Hyperparameter optimization is a crucial step in developing effective vision-language-action (VLA) models, as it directly impacts the model's ability to generalize and perform well on unseen data [4]. In this section, we present the results of our hyperparameter tuning experiments conducted to optimize the VLA model's performance. Specifically, we focus on evaluating different combinations of learning rates, batch sizes, and weight decay values. The objective of this process is to identify the hyperparameter settings that yield the highest validation accuracy, thereby justifying our final model configuration.

Table 4 summarizes the validation accuracy achieved with various hyperparameter combinations. As observed, a learning rate of 1e-4, a batch size of 64, and a weight decay of 1e-5 resulted in the highest validation accuracy of 0.85. This configuration strikes a balance between learning speed and generalization, preventing overfitting while allowing the model to converge effectively. Smaller batch

Modality Removed	Performance Drop (%)	Example Tasks Impacted
1. Vision	25	Object Recognition, Scene Understanding
2. Language	15	Instruction Following, Task Planning
3. Action	30	Motor Control, Task Execution

Table 1. Impact of modality ablation on VLA model performance. This table provides a summary of the performance drop observed when each modality is removed from the pretraining data, along with examples of tasks that are particularly impacted.

Table 2. Comparison of Model Parameters and Computational Cost

Model	Parameters (M)	Training Time (hours)	Inference Speed (FPS)	Memory Footprint (GB)
VLA Model	150	48	30	20
Baseline Model A	100	36	40	15
Baseline Model B	200	60	25	25

sizes (e.g., 32) paired with the same learning rate yielded slightly lower accuracy (0.82), potentially due to increased noise during training. A higher learning rate (1e-3) led to a significant drop in accuracy (0.78), indicating that the model may have overstepped optimal parameter values during training. Increasing the weight decay to 1e-4, while keeping the learning rate and batch size at 1e-4 and 32 respectively, also resulted in a decrease in validation accuracy (0.80), suggesting that excessive regularization hindered the model's ability to learn the underlying patterns in the data.

These findings align with established principles in machine learning, where careful selection of hyperparameters is essential for achieving optimal performance [1, 6]. The selected hyperparameter settings provide a strong foundation for the VLA model's training and subsequent evaluation on robotic tasks.

5. Conclusion

The Vision-Language-Action (VLA) model presented showcases significant potential for the development of scalable and generalizable robotic systems. The empirical results strongly suggest that pretraining on large, diverse multimodal datasets is a viable strategy for endowing robots with a broad range of skills and the ability to adapt to new tasks and environments. The ablation studies underscore the importance of each modality—vision, language, and action—in achieving robust performance. Future research efforts should prioritize refining task-specific performance through architectural innovations and training methodologies. Further exploration of novel architectures is warranted to unlock additional capabilities and improve the overall efficiency of VLA models in real-world robotic applications.

References

[1] Yasser A. Ali, Emad Mahrous Awwad, Muna Al-Razgan, and Ali Maarouf. Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes*, 11:349–349, 2023. 3

[2] T. B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric J. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek S. Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James T. Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, D. Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Érica Rodrigues Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Fırat, Michele Catasta, Jason Lee, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 1

[4] Justus A Ilemobayo, Olamide I Durodola, Oreoluwa Alade, Opeyemi J Awotunde, Adewumi T Olanrewaju, Olumide Falana, Adedolapo Ogungbire, Abraham Osinuga, Dabira Ogunbiyi, Ark Ifeanyi, Ikenna Odezuligbo, and Oluwagbotemi E. Edu. Hyperparameter tuning in machine learning: A comprehensive review, 2024. 2

[5] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. pages 512–519, 2014. 2

Table 3. Performance Breakdown on Different Task Categories

Task Category	VLA Model Accuracy	Baseline Accuracy
Manipulation	0.85	0.70
Navigation	0.78	0.60
Perception	0.90	0.75
Tool Use	0.65	0.50

Table 4. Hyperparameter Tuning Results for VLA Model

Learning Rate	Batch Size	Weight Decay	Validation Accuracy
1e-4	32	1e-5	0.82
1e-4	64	1e-5	0.85
1e-3	32	1e-5	0.78
1e-4	32	1e-4	0.80

[6] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms, 2012. 3