EVALUATING AND FINE-TUNING VISION-LANGUAGE-ACTION MODELS FOR ROBOTIC CONTROL IN NOVEL ENVIRONMENTS

Kilian Preuss[1]
[1]*Computer Science Department, IMT Atlantique, Brest, France*

Summary Vision-Language-Action (VLA) models, built upon pretrained Vision-Language Models (VLMs) and trained on large-scale robotics datasets, have demonstrated strong task, environment, and semantic generalization capabilities. However, it remains unclear how to efficiently adapt them to new environments, embodiments, and tasks during the post-training phase.
In this work, we investigate the use of Parameter-Efficient Fine-Tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA), originally developed for natural language processing, vision–language models, or vision tasks, to adapt VLA models. We evaluate their effectiveness in terms of both task performance and fine-tuning computational cost, and compare them to naïve full fine-tuning. To this end, we will conduct experiments in the MuJoCo simulator using the LIBERO benchmark, a widely adopted framework for evaluating generalization and adaptation in robotic learning.

Keywords: Vision-Language-Action (VLA) Models, Parameter-Efficient Fine-Tuning (PEFT), Low-Rank Adaptation (LoRA), Robotics, Transfer Learning

INTRODUCTION

There has been growing interest in leveraging the high-level reasoning and instruction-following capabilities of large language models (LLMs) for robotic decision-making. Recent work has extended LLMs with visual encoders to create Vision–Language Models (VLMs) [1], [2], and subsequently fine-tuned them on robot-centric multimodal datasets to obtain Vision–Language–Action (VLA) models [3], [4], [5]. These systems achieve high success rates on a wide variety of manipulation and navigation tasks.
However, despite their impressive generalization capabilities, current VLA models remain difficult to adapt to new environments, embodiments, or task distributions. Full fine-tuning is computationally expensive and memory-intensive, making post-training adaptation impractical in many robotic settings.
Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA [6], [7] have emerged as a promising family of techniques for adapting large models while updating only a small number of parameters. While these approaches have been extensively studied in NLP and in vision, their applicability to VLA models for robotic control remains largely unexplored.
In this work, we evaluate the effectiveness of several PEFT methods for adapting a pretrained VLA model to new robotic tasks. We assess their parameter efficiency, and performance on the LIBERO benchmark, providing the first comparative study of PEFT strategies in the context of VLA adaptation.

RELATED WORK

Vision-Language-Action models for robotics

The earliest research using Large Language Models (LLMs) for robotic tasks employed them as high-level planners, while relying on another type of model for low-level control [8], [9].
Subsequently, Vision-Language Models (VLMs) were adapted for end-to-end control in robotics. The main approach was to fine-tune a VLM on Internet-scale vision-language tasks, such as spatial reasoning question–answering, and then train it in a second stage on robotic trajectories (image–action pairs) to directly output actions, thereby creating a Vision-Language-Action (VLA) model [5], [10], [11], [12], [13].
Another line of work employs a two-system architecture: a slow, large VLM running at a lower frequency that provides high-level understanding, paired with a smaller, lightweight expert running at a higher frequency to produce low-level actions [4], [14], [15], [16], [17].

Parameter-Efficient Fine-Tuning (PEFT)

Due to the large number of parameters in modern foundation models, full fine-tuning, where all weights are updated, is computationally expensive and memory-intensive. To address this limitation, a family of techniques known as Parameter-Efficient Fine-Tuning (PEFT) [6] has emerged. These methods update only a small subset of the model parameters while keeping the majority frozen, and have become widely adopted for adapting LLMs.
LoRA [7] is one of the most commonly used PEFT approaches. It introduces trainable low-rank matrices into linear layers, enabling efficient adaptation with a small number of parameters. LoRA has been extensively applied in NLP for fine-tuning LLMs, making it a natural candidate for evaluating its performance when adapting VLA models.

* Corresponding author. E-mail: kilian.preuss@imt-atlantique.net

LayerNorm Tuning [18] updates only the scale and shift parameters of LayerNorm layers. Originally proposed for efficiently adapting LLMs to vision-language tasks, it is an appealing choice for VLAs, which rely heavily on robust visual and spatial understanding.

IA^3 [19] adapts the model by rescaling internal activations through learned vectors applied to the attention queries, keys, values, and feed-forward layers. It is particularly relevant in our context, as it was developed for few-shot learning, a setting similar to LIBERO, where only a small number of demonstration trajectories are available.

OFT (Orthogonal Fine-Tuning) [20] constrains the parameter updates to an orthogonal transformation of selected weight matrices. Initially developed for improving image generation conditioned on natural language, OFT is relevant here because VLA models also generate actions conditioned on natural language instructions.

METHOD

Pretrained VLA Model

In this work, we use the widely adopted open-source OpenVLA [5] model. OpenVLA is trained on the Open X-Embodiment dataset [21], a large-scale collection of robotic trajectories covering diverse embodiments, tasks, and environments.

Architecturally, OpenVLA builds on the Prismatic-7B Vision–Language Model (VLM) [22], which itself is based on the LLaMA 2 7B [23] backbone and incorporates both SigLIP and DINOv2 [24], [25] as vision encoders.

To enable robotic control, OpenVLA converts continuous robot actions into discrete 256-bin action tokens. Each action dimension is uniformly quantized, and the least-used LLaMA-2 vocabulary tokens are replaced with these action-specific symbols. During inference, the model predicts action tokens autoregressively and decodes them back into continuous Δposition/Δrotation/gripper values by mapping each token to its corresponding quantization bin.

PEFT Methods Considered

In our experiments, we apply LoRA and IA^3 [7], [19] to the query, key, and value projection matrices of each transformer block, as well as to all feed-forward layers. OFT [20] is applied only to the query, key, and value projection matrices. For both LoRA and OFT, we set the low-rank dimension to $r = 32$. For LayerNorm Tuning [18], we fine-tune all LayerNorm scale and shift parameters.

Benchmark: LIBERO

LIBERO [26] is a benchmark for lifelong robot learning, providing 130 manipulation tasks grouped into four suites: spatial, object, goal, and long-horizon tasks. For each environment, it offers a training set of human-teleoperated robotic trajectories, consisting of RGB camera observations paired with continuous control actions. After fine-tuning the model on this small training set, performance is evaluated in the MuJoCo simulator using the per-task success rate on a predefined set of tasks.

RESULTS

To compare the different fine-tuning methods, we restrict our evaluation to the LIBERO Spatial suite and report the success rate, defined as the percentage of tasks successfully completed, together with the proportion of trainable parameters in Table 1.

Method	**Spatial SR (%)**	**Trainable Params (%)**
OpenVLA + LoRA	83.2	1.4483
OpenVLA + LayerNorm FT	0.00	0.0030
OpenVLA + OFT	0.00	0.8820
OpenVLA + IA3	0.00	0.0151

Table 1. Performance and parameter efficiency of PEFT methods applied to OpenVLA on the LIBERO Spatial suite.

We observed that most fine-tuning methods fail to achieve a non-zero success rate. This outcome highlights the sparsity of the success metric, which alone does not fully capture the extent to which each model has learned the task. When examining the rollout videos more closely, we observe that OFT and IA^3 still manage to partially follow the given prompt. In contrast, LayerNorm Tuning fails to follow any instructions. The resulting robot motions are erratic and unstable, reflecting the limited expressive capacity of updating only LayerNorm parameters.

CONCLUSION

We evaluated and compared several PEFT methods on the LIBERO benchmark. Our results show that, in our experimental settings, LoRA appears to be the most effective method for fine-tuning VLAs on novel tasks. OFT and IA^3 learn some aspects of the instructed behavior but still fail to execute the complete task. In contrast, LayerNorm Tuning lacks the expressive capacity required to capture the complexity of the tasks.

References

[1] A. Radford et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

[2] J. Li et al., "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, PMLR, 2023, pp. 19 730–19 742.

[3] B. Zitkovich et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*, PMLR, 2023, pp. 2165–2183.

[4] K. Black et al., "π: A vision-language-action flow model for general robot control," *CoRR*, 2024.

[5] M. J. Kim et al., "Openvla: An open-source vision-language-action model," in *8th Annual Conference on Robot Learning*.

[6] Z. Han et al., "Parameter-efficient fine-tuning for large models: A comprehensive survey," *CoRR*, 2024.

[7] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*.

[8] M. Ahn et al., "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*, 2022. [Online]. Available: `https://api.semanticscholar.org/CorpusID:247939706`.

[9] D. Driess et al., "Palm-e: An embodied multimodal language model," in *International Conference on Machine Learning*, 2023. [Online]. Available: `https://api.semanticscholar.org/CorpusID:257364842`.

[10] A. Brohan et al., "Rt-1: Robotics transformer for real-world control at scale," *Robotics: Science and Systems XIX*, 2023.

[11] J. Wen et al., "Tinyvla: Toward fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 10, pp. 3988–3995, 2024. [Online]. Available: `https://api.semanticscholar.org/CorpusID:272753287`.

[12] O. Mees et al., "Octo: An open-source generalist robot policy," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[13] M. Reid et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *CoRR*, 2024.

[14] Q. Li et al., "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," *CoRR*, 2024.

[15] J. Bjorck et al., "Gr00t n1: An open foundation model for generalist humanoid robots," *CoRR*, 2025.

[16] J. Wen et al., "Dexvla: Vision-language model with plug-in diffusion expert for general robot control," *arXiv preprint arXiv:2502.05855*, 2025.

[17] M. Shukor et al., "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[18] B. Zhao et al., "Tuning layernorm in attention: Towards efficient multi-modal llm finetuning," in *The Twelfth International Conference on Learning Representations*.

[19] H. Liu et al., "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950–1965, 2022.

[20] Z. Qiu et al., "Controlling text-to-image diffusion by orthogonal finetuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 79 320–79 362, 2023.

[21] A. O'Neill et al., "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 6892–6903.

[22] S. Karamcheti et al., "Prismatic vlms: Investigating the design space of visually-conditioned language models," in *Forty-first International Conference on Machine Learning*.

[23] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[24] X. Zhai et al., "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[25] M. Oquab et al., "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*,

[26] B. Liu et al., "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.