Research Project

Kilian Preuss

September 2025

1 LLMs for Robotics

There has been a growing body of research focusing on leveraging the general knowledge of LLMs to reason about robotic tasks. This has led to the emergence of Vision-Language-Action (VLA) models. Initially, LLMs were adapted with a vision encoder to enable reasoning over visual input, which gave rise to Vision-Language Models (VLMs). VLAs are VLMs that are typically fine-tuned on robot-specific question–answering tasks, such as spatial reasoning Q&A. In a second stage, these models are further modified to stop generating text tokens and instead produce actions. For this latter stage, there exist numerous methods to adapt a robot-informed VLM into a VLA. I divided these methods into three main families.

1.1 Vision-Language Models for High-Level Planning

The earliest research using LLMs for robotic tasks employed them as high-level planners, while relying on another type of model for low-level control.

In SayCan [1], the LLM was prompted only with text (without image information) and had to provide the most useful high-level skill to perform in order to complete the task. Hard-coded skills were then used to execute the corresponding actions.

PaLM-E [4] was the first paper to also encode images and robot state information for the LLM to process. The LLM then output high-level plans and sequences of skills. These skills were subsequently passed to a low-level policy that mapped each step into actual robot actions.

1.2 Vision-Language-Action Models for End-to-End Control

During the early exploration of using Vision-Language Models 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. In a second stage, the model was trained on robotic trajectories (image–action pairs) in order to directly output actions.

One method consisted of discretizing the action space into bins to tokenize them [2, 17], and then training the model to predict the correct action token given a natural language description of the task and images of the current state. Other methods, instead of discretizing the actions, passed the embeddings generated by the VLM into an MLP projector that predicted actions in continuous space [7, 13].

Another approach was to fine-tune the VLM on a large set of embodied reasoning tasks and then use it for zero-shot robot control via code generation, as well as few-shot control via in-context learning [12].

The main limitation of these approaches is that Vision-Language Models typically have a very large number of parameters. When deployed on a robot and used directly as the control policy, they cannot run at sufficiently high frequencies (typically only 1–3 Hz).

1.3 VLMs with Diffusion Experts

Another approach is to use a two-system architecture: a slow, large VLM that runs at a lower frequency, and a smaller, lighter expert that can run at a higher frequency. The VLM is used to create rich embeddings of the natural language task and the images. These embeddings are then fed to the action expert, which uses them to generate informed actions. In this way, the problem of low inference frequency is mitigated while still leveraging large language models. The action expert is usually a diffusion model [16, 9, 11].

There are also other methods that do not only use the final embedding, but instead merge the diffusion expert more directly and elegantly into the main VLA [14, 10].

2 Open X-Embodiment

The Open X-Embodiment dataset [15] is a collection of 60 different datasets gathered from 34 research labs, including 527 distinct skills and 22 different embodiments. It is the largest dataset of robotic trajectories (camera–action pairs) annotated with natural language instructions.

The dataset contains camera viewpoints both from a distance and mounted on the gripper.

It has been used to train many VLA models [17, 13, 14, 10].

Among the embodiments included in the dataset is the XArm, which we plan to use in this project. In fact, it is the embodiment with the largest number of trajectories.

3 Work Plan

Vision-Language-Action models (VLAs) demonstrate state-of-the-art generalization, both to new tasks, new environments, and to different embodiments.

The first part of this research project is to evaluate whether VLAs can generalize well enough to be used directly, out of the box, on the robot in our laboratory.

In the second stage, we will attempt to develop an effective method to efficiently adapt VLAs to scenarios that are too far out of distribution, where they otherwise fail when used out of the box.

Regarding the first stage, we will use VLAs that have been trained on data from the same embodiment as our robot. OpenVLA [17] is such a VLA, as it was trained on a subset of the Open-X-Embodiment dataset, which included the XArm robot. CogACT [16] was also trained on a subset of Open-X-Embodiment that included the XArm. Finally, Pi Zero [14] was trained on the entire Open-X-Embodiment dataset, as well as internal data, and thus was already exposed to the XArm.

We will first evaluate the VLA in simulation, specifically in Isaac Sim [8]. This simulator provides state-of-the-art photorealistic rendering, which is important since VLAs rely heavily on RGB camera input. Using such a simulation helps minimize the sim-to-real gap.

The environments we plan to use for evaluation are ...

The tasks we plan to use for evaluation are ...

We will evaluate the VLA using task completion and success rate as metrics, as is common in related work [11, 17, 14].

For the second stage, there is already significant research on post-training adaptation methods aimed at improving VLA performance. For example, Optimized Fine-Tuning (OFT) [5] adapts the action decoder to improve performance on a new embodiment. Similarly, TGRPO and SimpleVLA-RL [3, 6] adapt the GRPO algorithm, originally developed for NLP, to the

VLA setting for robotic tasks. Thus, I need to read more about the different methods in order to be able to propose a new one.

References

[1] Michael Ahn, Anthony Brohan, Yevgen Chebotar, Juan Cortes, , et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, , et al. Rt-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems (RSS)*, 2023.

[3] Zengjue Chen, Runliang Niu, He Kong, Qi Wang, Qianli Xing, and Zipei Fan. Tgrpo: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization. In *arXiv preprint arXiv:2506.08440*, 2025.

[4] Daniel Driess, Fei Xia, Mehdi S. M. Sajjadi, , et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[5] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[6] Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, Dehui Wang, Dingxiang Luo, Yuchen Fan, Youbang Sun, Jia Zeng, Jiangmiao Pang, Shanghang Zhang, Yu Wang, Yao Mu, Bowen Zhou, and Ning Ding. Simplevla-rl: Scaling vla training via reinforcement learning. *arXiv preprint arXiv:2509.09674*, 2025.

[7] Kevin Lin, Nikhil Patel, Alex Fang, , et al. Tinyvla: Towards efficient vision-language-action models. *arXiv preprint arXiv:2407.06440*, 2024.

[8] Viktor Makoviychuk, Alexey Makoviychuk, Felix Widmaier, , et al. Nvidia isaac sim: A scalable robotics simulation platform. *arXiv preprint arXiv:2108.10470*, 2021.

[9] Takuma Miki, Siddharth Karamcheti, , et al. Groot: Learning generalist robots through internet-scale simulation. *arXiv preprint arXiv:2307.05742*, 2023.

[10] Ilija Radosavovic, , et al. Smolvla: Small vision-language-action models. *arXiv preprint arXiv:2408.04296*, 2024.

[11] DexVLA Team. Dexvla: Generalizable dexterous manipulation via vla models. *arXiv preprint arXiv:2407.08419*, 2024.

[12] Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across modalities. *arXiv preprint arXiv:2403.05530*, 2024.

[13] Octo Team et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[14] OpenAI Robotics Team. Pi0: A vision-language-action model for generalist robots. *arXiv preprint arXiv:2407.16875*, 2024.

[15] Annie Xie, Anthony Brohan, Noah Brown, , et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[16] Annie Xie, Kuan Fang, and Sergey Levine. Cognitive action models: Grounding large language models in robot actions. In *Conference on Robot Learning (CoRL)*, 2023.

[17] Bolun Zhao, Anthony Brohan, Yuchen Li, , et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.