New Explorations in Electrical Engineering (NECE)
Department of Electrical Engineering,
Faculty of Engineering
Universiti Malaya, MALAYSIA

Vol. 2, No. 1 (2026)
pp. 21–30

Evaluating Visual Generalization Boundaries in Low-Parameter Vision-Language-Action Models During Rigid Robotic Manipulation

Jac Tze Tang[a], Pauline Shan Qing Yeoh[a,b*] and Hong Vin Koay[c]
[a] *Department of Biomedical Engineering, Universiti Malaya, 50603 Kuala Lumpur, Malaysia.*
E-mail: jactze@gmail.com
[b] *School of Engineering, Monash University, 47500 Bandar Sunway, Selangor Darul Ehsan, Malaysia*
*E-mails:*pauline.yeoh@monash.edu*
[c] *Independent Researcher*
E-mail: koayhongvin@gmail.com

Abstract. Open-source Vision-Language-Action (VLA) models enable rapid physical deployment in robotics, but their visual generalization limitations remain underexplored. This study evaluates the SmolVLA model when confronted with out-of-distribution visual variables. An SO-101 robotic arm, equipped with a vertical parallel pincer, was trained via imitation learning on over 100 teleoperated demonstrations to grasp and displace a baseline 3D-printed black cube, conditioned on the language instruction "pick the cube." To systematically assess generalization limits, we evaluated four categories: baseline, color, scale, and geometry variation. Performance degrades under out-of-distribution shifts, while most evaluated conditions improve after further fine-tuning. The success rates for the red and yellow cubes gradually improved with fine-tuning but remained lower than those observed under the other evaluated out-of-distribution conditions. Initial fine-tuning favored the smaller cube (90% success), but prolonged training reversed this trend, indicating that object scale may impact long-term policy generalization. The failure modes observed in the colour variation condition may reflect a model limitation in translating learned manipulation trajectories to cubes with differing pixel-level visual appearances. These results provide insight into the limited out-of-distribution generalization of the SmolVLA model, highlighting the need for targeted data augmentation to ensure the reliability of assistive robotic systems in real-world environments.

Keywords. Vision-Language-Action Model, Imitation Learning, Robotic Manipulation, Visual Generalization

1. Introduction

The emergence of Vision-Language-Action (VLA) models has transformed the landscape of robotic manipulation (Kawaharazuka et al., 2025; X. Li et al., 2026; Z. Li et al., 2026; Ma et al., 2026). VLA models are end-to-end neural architectures that directly map visual observation and natural language instruction to a continuous stream of physical motor actions (Yi et al., 2026). By fusing large language models with visual encoders, VLAs enable robots to interpret natural language instructions and translate

NECE. All rights reserved

them directly into physical motor commands (Kawaharazuka et al., 2025; Kim et al., 2024; Wang et al., 2026). The general VLA architecture is illustrated in Fig.1a.

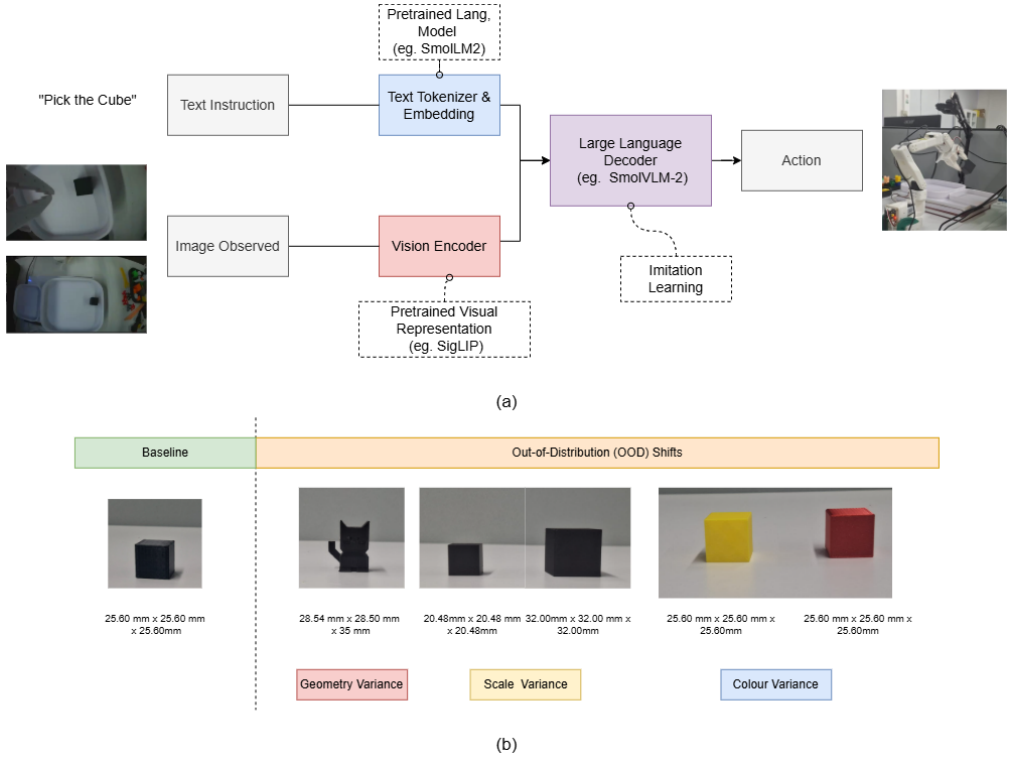

Figure 1. (a) General Vison-Language-Action architecture b) Object used for baseline and out-of-distribution shift evaluation.

While existing pretrained, massive billion-parameter VLA models have demonstrated remarkable zero-shot reasoning (Kim et al., 2024; X. Li et al., 2026; Shukor et al., 2025), heavy reliance on cloud infrastructure and high computational cost limits practical deployment (Sejnova et al., 2024; Takagi et al., 2026; Wen et al., 2025; Zitkovich et al., 2023). Consequently, current research is pivoting toward efficient or lightweight VLAs (Kawaharazuka et al., 2025; Yu et al., 2025), such as SmolVLA (Shukor et al., 2025), TinyVLA(Wen et al., 2025) and AnoleVLA(Takagi et al., 2026), that are more accessible for real-world deployment. This shift is democratizing robotics, allowing for the deployment of affordable, open-source hardware like the SO-101 arm (Cadene et al., 2026).

Advancements in VLA models have demonstrated strong potential for enabling generalist robotic manipulation (Goyal et al., 2025; Z. Li et al., 2026; Sridhar et al., 2025). However, despite the mechanical and computational advantages of low-parameter VLAs, model performance relies on imitation learning, which introduces a critical limitation, which is the generalization gap (Z. Li et al., 2026). This limitation is frequently manifested through distribution shift and visual brittleness, where models fail to maintain robust performance under visually altered environments or previously unseen object characteristics (Kim et al., 2024). For instance, a model trained exclusively on a black cube may experience substantial performance degradation when evaluated using objects with different colors, textures, scales, or geometries, such as a yellow cube or a black cat-shaped object (Fig.1b). Such failures become particularly pronounced when VLAs are trained on narrow sets of teleoperated demonstrations, resulting in

heightened sensitivity to out-of-distribution (OOD) visual variations (Black et al., 2025; Zitkovich et al., 2023).

For assistive robotic arms to achieve practical real-world deployment, they must demonstrate robust operational capability in varying environments without failing due to minor visual perturbations. This requirement becomes particularly critical for low-parameter VLA models, whose computational efficiency often comes at the expense of robust generalization capability (Takagi et al., 2026; Zitkovich et al., 2023). Hence, a research question arises: To what extent do low-parameter VLA models exhibit performance degradation under specific out-of-distribution visual shifts, including variations in color, scale, and geometry, during rigid-object robotic manipulation tasks? Specifically, when the model is trained exclusively to grasp a baseline object, it remains unclear how its zero-shot generalization performance is affected by appearance-level versus structure-level distribution shifts when confronted with different object instances.

To investigate this limitation, this paper evaluates the performance of the fine-tuned SmolVLA (Shukor et al., 2025), a 450M-parameter pre-trained VLA model, under controlled visual perturbation scenarios. Specifically, the model is fine-tuned solely on demonstrations involving the manipulation of a baseline black cube and subsequently evaluated under three isolated OOD conditions: color shift, scale shift, and geometric shift. The experimental protocol includes testing on cubes with different colors, enlarged and reduced black cubes, and a cat-shaped object to isolate the OOD conditions on the policy performance. The primary motivation of this work is to critically assess the practical reliability and deployment readiness of low-parameter VLA models for assistive robotic manipulation tasks. By quantitatively characterizing failure mechanisms under controlled visual distribution shifts, this study contributes to a targeted ablation-based analysis toward understanding the visual generalization boundaries of lightweight VLA systems.

2. Methodology

1. Experimental setup

A 6-degree-of-freedom (6-DOF) SO-101 Lerobot Follower and Leader Arm (Cadene et al., 2026) equipped with vertical parallel pincers, and a standing camera was used for data collection and inference. All experiments were implemented with NVIDIA RTX5090 Graphical Processing Unit and Intel Core i9 processor. 3D printed items, trays and cubes of various colors are printed using BambuLabs P1S 3D Printer. The overall experimental framework is presented in Fig. 2.

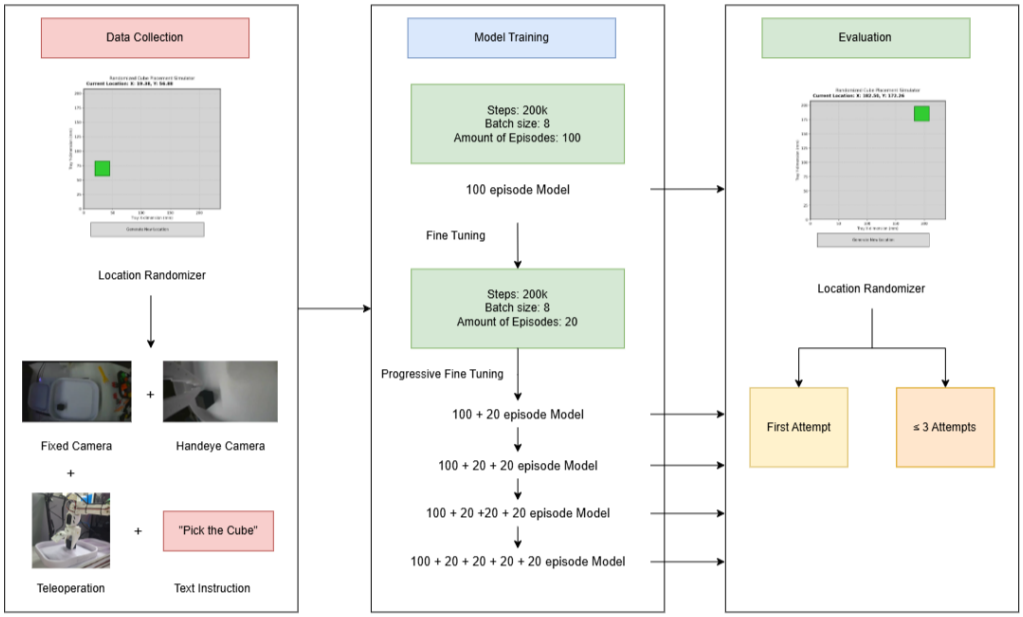

Figure 2. Overall framework diagram.

2. *Data collection and training*

Prior to recording each demonstration episode, a custom Python-based location randomizer (via Matplotlib) was used to generate the exact coordinates for the cube on the 3D-printed base tray (pick location) and its destination on the collector tray (drop location). This allows consistent random cube placements and al-lows precise repeatability of a recorded episode with undesirable motor movement and video footage.

Demonstration data were subsequently collected using the LeRobot leader–follower teleoperation framework. During each episode, a human operator teleoperates the Follower Arm by manipulating the Leader arm to pick and drop the black cube into the collection tray, then returning to the resting position. The Leader arm was operated with only the live video footage captured by the handeye camera on the SO-101 Follower and a fixed camera, both with 640x360p resolution to pick up the black cube. The task command is “Pick the cube” for all demonstration episodes to encourage generalized task interpretation during model evaluation. The collected dataset consists of 180 episodes with 640 × 360 resolution fixed camera and handeye camera feed.

3. *Fine-tuning the VLA model*

In our work, the SmolVLA model was progressively fine-tuned in a sequential manner, where each stage was initialized from the previous checkpoint and further trained using additional demonstrations. First, the pretrained SmolVLA model was fine-tuned using the initial 100 demonstration episodes, corresponding to the task instruction “Pick the Cube”. The model was trained for a total of 200k steps with batch number of 8, AdamW optimizer with $\beta1 = 0.9$ and $\beta2 = 0.95$, learning rate of 0.0001 and

weight decay of 1e-10. Then, the fine-tuned model was further fine-tuned with additional sets of 20-episode datasets with the same training hyperparameters for 200k steps.

4. Evaluating the fine-tuned models

Each evaluation session consisted of 20 independent pick-and-drop tasks conducted under randomized object-placement conditions. At the start of each session, the robotic arm initialized from its predefined default pose before executing the trained policy. For every randomized object position, the policy was allowed up to three execution attempts. The success rate was used as the evaluation metric. A trial was considered successful when the robot autonomously grasped the target cube, deposited it into the designated collection tray, and returned to its default pose without external assistance within the allocated attempts. In addition to the cumulative success rate, first-attempt success rates were separately reported to evaluate single-shot manipulation robustness. All sessions were recorded using the robot's handeye and the static fixed camera to support qualitative analysis.

3. Results

Table 1, 2, 3, and 4 present the quantitative comparison between the pre-trained and fine-tuned policies across four evaluation conditions: baseline, color variation, scale variation, and geometry variation. The baseline condition refers to inference conducted using the same black cube as that used during training.

Table 1. Performance comparison of progressively fine-tuned VLA policies in baseline condition.

Policy	Success Rates (%)		Mean attempts to success
	First-attempt success	Success within ≤3 attempts	
SmolVLA (Base)			
+ Fine-tuned 100 episodes (FT-100)	55.00	80.00	1.50
+ Fine-tuned 20 episodes (FT-100 + 20)	70.00	80.00	1.19
+ Fine-tuned 20 episodes (FT-100 + 20+20)	55.00	85.00	1.53
+ Fine-tuned 20 episodes (FT-100 + 20+20+20)	65.00	85.00	1.29
+ Fine-tuned 20 episodes (FT-100 + 20+20+20+20)	50.00	95.00	1.37

Table 1 reports the performance of the progressively fine-tuned policies under the baseline condition. The results demonstrate that fine-tuning yields substantial improvements in task success over the baseline. First-attempt success reached 70.00% after a modest amount of fine-tuning (FT-100+20), demonstrating that the policy can complete the task reliably without corrective retries even with limited additional training. Notably, fine-tuning with as few as 100 episodes (FT-100) already achieved an 80.00% success rate within three attempts, suggesting that small-scale fine-tuning can produce practically useful policies for real-world applications where limited corrective interaction is acceptable. Success within three attempts improved consistently with additional fine-tuning data, rising from 80.00% at FT-100 to 95.00% at the final stage (FT-100+20+20+20+20), reflecting improved policy stability and consistency as more training data is incorporated, which aligns with observations reported in larger-scale VLA studies. This

also aligns with trends in Vision-Language Model adaptation, where increasing augmented views and few-shot samples have been shown to enhance OOD robustness (Zhu et al., 2024).

Table 2. Performance comparison of VLA policies across color variation conditions.

Policy	Conditions	Black (Baseline)	Yellow	Red
FT-100	First-attempt (%)	55.00	0.00	15.00
	≤3 attempts (%)	80.00	10.00	40.00
	Mean attempts	1.50	3.00	1.88
FT-100 + 20	First-attempt (%)	70.00	25.00	30.00
	≤3 attempts (%)	80.00	45.00	60.00
	Mean attempts	1.19	1.78	1.67
FT-100 + 20+20	First-attempt (%)	55.00	30.00	25.00
	≤3 attempts (%)	85.00	55.00	60.00
	Mean attempts	1.53	1.64	1.29
FT-100 + 20+20+20	First-attempt (%)	65.00	30.00	35.00
	≤3 attempts (%)	85.00	45.00	50.00
	Mean attempts	1.29	2.08	1.4
FT-100 + 20 + 20 + 20 + 20	First-attempt (%)	65.00	30.00	50.00
	≤3 attempts (%)	95.00	65.00	80.00
	Mean attempts	1.37	1.77	1.56

Table 3. Performance comparison of VLA policies across scale variation conditions.

Policy	Conditions	-20%	0 (Baseline)	+25%
FT-100	First-attempt (%)	35.00	55.00	35.00
	≤3 attempts (%)	65.00	80.00	80.00
	Mean attempts	1.31	1.50	1.75
FT-100 + 20	First-attempt (%)	70.00	70.00	70.00
	≤3 attempts (%)	90.00	80.00	80.00
	Mean attempts	1.28	1.19	1.19
FT-100 + 20+20	First-attempt (%)	30.00	55.00	60.00
	≤3 attempts (%)	60.00	85.00	90.00
	Mean attempts	1.58	1.53	1.39
FT-100 + 20+20+20	First-attempt (%)	20.00	65.00	60.00
	≤3 attempts (%)	65.00	85.00	90.00
	Mean attempts	1.29	1.29	1.33
FT-100 + 20 + 20 + 20 + 20	First-attempt (%)	55.00	65.00	40.00
	≤3 attempts (%)	70.00	95.00	95.00
	Mean attempts	1.37	1.37	1.68

Table 4. Performance comparison of VLA policies across geometry variation conditions.

Policy	Conditions	Cube (Baseline)	Cali Cat
FT-100	First-attempt (%)	55.00	40.00
	≤3 attempts (%)	80.00	65.00
	Mean attempts	1.50	1.54
FT-100 + 20	First-attempt (%)	70.00	45.00
	≤3 attempts (%)	80.00	65.00
	Mean attempts	1.19	1.31
FT-100 + 20+20	First-attempt (%)	55.00	40.00
	≤3 attempts (%)	85.00	95.00
	Mean attempts	1.53	1.95
FT-100 + 20+20+20	First-attempt (%)	65.00	55.00
	≤3 attempts (%)	85.00	85.00
	Mean attempts	1.29	1.53
FT-100 + 20 + 20 + 20 + 20	First-attempt (%)	65.00	50.00
	≤3 attempts (%)	95.00	85.00
	Mean attempts	1.37	1.59

As shown in Table 2, the policy achieves only a 10.00% success rate within three attempts for FT-100 and 45.00% for FT-100+20 on the yellow-colored cube. In comparison, Table 4 shows that the Cali Cat geometry (a cube-like cat-shaped object) achieves a notably higher 65.00% success rate within three attempts for both FT-100 and FT-100+20, suggesting that the policy generalizes more readily to novel geometry than to to the evaluated colour conditions. In Table 3, the FT-100 policy achieves a 65.00% success rate, and FT-100+20 achieves a 90.00% success rate on the small cube (-20%) within three attempts, while both policies achieve an 80.00% success rate on the large cube (+25%).

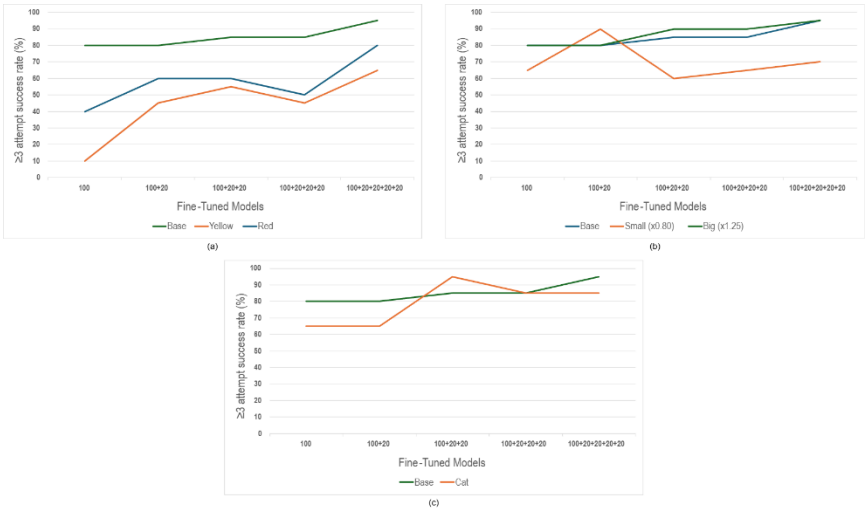

Figure 3. Performance of Vison-Language-Action policies across (a) Color variation, (b) Scale variation and (c) Geometric Variation

However, performance on the small cube degrades after further fine-tuning, dropping to 70.00% by the final stage, whereas performance on the large cube improves to 95.00%. Overall, performance under scale and geometry variation does not show the same magnitude of improvement observed under color variation. As shown in Fig. 3, our findings show that although performance degrades under out-of-distribution shifts, overall generalization improves with increased fine-tuning episodes, with the extent and stability of this improvement varying by variation type.

4. Discussion

OOD generalization is a critical metric for the practical deployment of robotic agents in complex environments (Zhu et al., 2024). While existing studies have primarily focused on natural distribution shifts, such as variations in lighting and background environments (Kim et al., 2024; Wen et al., 2025), our study investigates the out-of-distribution generalization capability of SmolVLA, under variations in color, scale, and geometry while accounting for positional variability across trials.

Based on our observations, color variation was associated with the largest performance degradation, suggesting that the policy may be more sensitive to appearance-level or chromatic shifts than to structural variations such as scale or geometry as seen when the fine-tuned models confronting red and yellow cube. Besides, insufficient fine-tuning leads to reduced first-attempt success rates. It was also observed that the policy exhibits retry-like behavior within a single task execution, indicating partial recovery capability during failure cases. However, for smaller cube conditions, the performance of success rate within 3 attempts degraded from 90% (FT-100+20) to 70% (FT+100+20+20+20+20) after further fine-tuning, suggesting that models beyond FT-100+20 struggle to grasp and locate smaller cubes. Across all evaluation conditions, a consistent trend is observed whereby progressive fine-tuning leads to improved task success and enhanced robustness under distribution shifts. The largest performance gains are observed for the FT-100+20 policy, where incremental updates primarily improve stability and reduce failure frequency. Failure cases observed included position recognition errors, incomplete execution, and grasping point errors (Takagi et al., 2026).

Previous works on VLA policy training largely relied on scaling up large, fragmented, and heterogeneous robotic datasets without a principled understanding of how dataset composition affects performance, particularly under distribution shift (Kim et al., 2024; Shukor et al., 2025). In contrast to these brute-force approaches that overlook key uncertainties in dataset selection (Zitkovich et al., 2023), we examine how fine-tuning under a single baseline condition affects policy generalization and robustness. While our dataset is limited, object positions were generated using a custom Python-based location randomizer to introduce spatial variability across trials and reduce positional memorization bias. One limitation of this study is the small number of training demonstrations, which results in limited variation across different conditions. Future work should incorporate greater diversity in the training dataset, including a wider range of colours and visual conditions, to improve the generalizability of the policy across different scenarios, even when the underlying task remains the same. In addition, future studies should investigate and compare the fine-tuning performance of different established low-parameter VLA models under limited-data settings.

5. Conclusion

This work highlights that SmolVLA benefits from incremental fine-tuning, with improved manipulation success across the colour, scale, and geometry variations evaluated in this study. Within the evaluated OOD conditions, colour variation, particularly the evaluation involving the yellow and red cubes, produced the largest performance degradation, suggesting that the observed colour-related failure patterns may indicate sensitivity to changes in pixel-level visual representation of the target object. These findings primarily characterise the OOD generalization behaviour of SmolVLA under the experimental conditions investigated, while whether similar patterns occur across other lightweight VLA models remains to be validated.

Acknowledgements

This study was supported by Edge Intelligence Sdn. Bhd.. The authors also acknowledge the support of the International Transdisciplinary Research Conference (INTRAC) 2026 organized by Universiti Malaya.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article. Author PSQY is an Editorial Board member for INTRAC 2026 and was not involved in the editorial evaluation or the peer-review process of this manuscript. Author HVK is the director of Edge Intelligence Sdn. Bhd., the entity that provided equipment support for this research.

References

Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., & Galliker, M. Y. (2025). $\pi_ {0.5} $: a Vision-Language-Action Model with Open-World Generalization. 9th Annual Conference on Robot Learning,

Cadene, R., Aliberts, S., Capuano, F., Aractingi, M., Zouitine, A., Kooijmans, P., Choghari, J., Russi, M., Pascal, C., & Palma, S. (2026). Lerobot: An open-source library for end-to-end robot learning. *arXiv preprint arXiv:2602.22818*.

Goyal, A., Hadfield, H., Yang, X., Blukis, V., & Ramos, F. (2025). Vla-0: Building state-of-the-art vlas with zero modification. *arXiv preprint arXiv:2510.13054*.

Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., & Sanketi, P. (2024). Openvla: An open-source vision-language-action model.*Proceedings of Machine Learning Research* Conference on Robot Learning,

Li, X., Li, P., Qian, L., Liu, M., Wang, D., Liu, J., Kang, B., Ma, X., Wang, X., & Guo, D. (2026). What matters in building vision–language–action models for generalist robots. *Nature Machine Intelligence*, 1–15.

Li, Z., Liu, J., Dong, Z., Teng, T., Rouxel, Q., Caldwell, D., & Chen, F. (2026). Towards Deploying VLA Without Fine-Tuning: Plug-and-Play Inference-Time VLA Policy Steering via Embodied Evolutionary Diffusion. *IEEE Robotics and Automation Letters*, *11*(5), 6234–6241.

Ma, Y., Song, Z., Zhuang, Y., Hao, J., & King, I. (2026). A Survey on Vision–Language–Action Models for Embodied AI. *IEEE Transactions on Neural Networks and Learning Systems*.

Sejnova, G., Vavrecka, M., & Stepanova, K. (2024). Bridging language, vision and action: Multimodal VAEs in robotic manipulation tasks. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),

Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., & Marafioti, A. (2025). Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*.

Sridhar, K., Dutta, S., Jayaraman, D., & Lee, I. (2025). Ricl: Adding in-context adaptability to pre-trained vision-language-action models. *arXiv preprint arXiv:2508.02062*.

Takagi, Y., Kambara, M., Yashima, D., Seno, K., Tokura, K., & Sugiura, K. (2026). AnoleVLA: Lightweight Vision-Language-Action Model with Deep State Space Models for Mobile Manipulation. *arXiv preprint arXiv:2603.15046*.

Wang, X., Dengxiong, X., Bai, S., Zheng, P., & Zhang, Y. (2026). VLAbot: A human Vision–Language–Action models interaction framework for robotic assembly. *Robotics and Computer-Integrated Manufacturing*, *100*, 103268.

Wen, J., Zhu, Y., Li, J., Zhu, M., Tang, Z., Wu, K., Xu, Z., Liu, N., Cheng, R., & Shen, C. (2025). Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*.

Yi, T., Yang, Q., & Chen, E. (2026). ActionX: pre-training action experts with reinforcement learning for vision-language action models. *Frontiers in Neurorobotics*, *20*, 1806605.

Yu, Z., Wang, B., Zeng, P., Zhang, H., Zhang, J., Wang, Z., Gao, L., Song, J., Sebe, N., & Shen, H. T. (2025). A survey on efficient vision-language-action models. *arXiv preprint arXiv:2510.24795*.

Zhu, Y., Ji, Y., Zhao, Z., Wu, G., & Wang, L. (2024). Awt: Transferring vision-language models via augmentation, weighting, and transportation. *Advances in Neural Information Processing Systems*, *37*, 25561–25591.

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., & Wahid, A. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. Conference on Robot Learning.