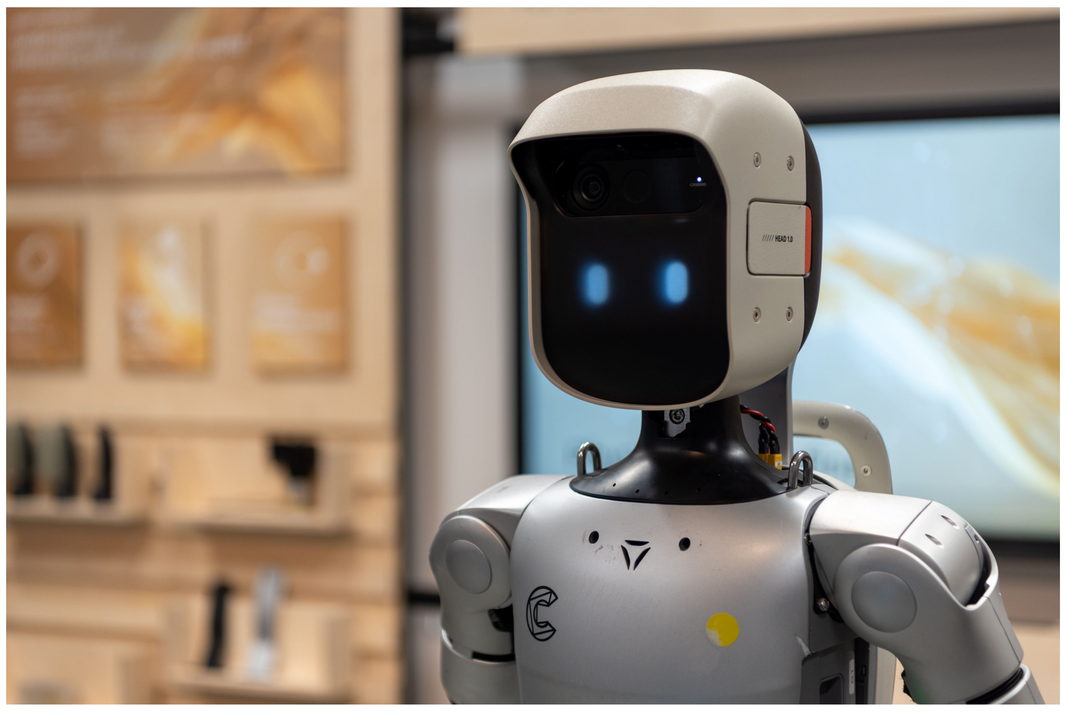
Towards Data-Efficient VLA Post-Training:

a case study of an industrial task

Anna Schooneveld | Geet Jose | Cuong Kasperzyk | Surabhi Rammurthy | I. J. Smallman | Riccardo Secoli

Introduction

- **Humanoid** robots are nearing industrial deployment in **warehousing**, where human-scale infrastructure drives rapid adoption. **Labour shortages** and costs drive demand, with order picking >55% of expenses. **Box picking is a key benchmark.**
- Vision-Language-Action (**VLA**) models are **promising**, since industrial environments require robustness to unstructured scenarios like damaged or misaligned boxes, where rigid automation fails but **human-like flexibility succeeds**.
- However, current post-training strategies for VLAs either use hundreds or thousands of demonstrations **limiting fast implementation and scalability** or **fall short of the required >99% reliability for industry.**
- **This work post-trains NVIDIA GR00T N1.5 on an unseen humanoid with only 60 demonstrations**, **achieving 97% success.** It outperforms the diffusion-based iDP3, which struggles to learn key behaviours, highlighting the strength of VLA pre-training for reliable real-world deployment.



Methodology

Data Collection Setup

- Experiments used a modified Unitree G1 humanoid with custom compliant 3D-printed hands, which improved grasp reliability despite no impedance control. An ORBBEC Gemini camera provided RGB and point clouds.
- Demonstrations were collected via motion-capture teleoperation, with a custom software stack to sync all data.
- The 60-episode dataset is carefully structured, covering positions, orientations, and hand configurations, with explicit demonstrations for absence handling and edge-case manipulation behaviours reliably.

Task Definition and Dataset

The box-picking task in this study imposes the following behavioural requirements:

- **Grasping**: the robot must successfully lift a box placed at any position within grasping reach, grasping the sides of the box with both hands.
- **Rotation invariance:** the robot must adapt its hand orientation to the box angle to grasp boxes across the full range of rotations encountered in deployment.
- **Absence handling**: the robot must not attempt to pick up in the absence of a box.

Experiments & Results

	Epoch tests			Lighting	Data ablation			iDP3
	Baseline/Short	Middle	Long	Yellow Spotlights	One Rep	Start A	Start C	
Num epochs	**23.0**	32.8	45.4	23.0 (baseline)	25.2	39.22	37.74	300
Box middle	10	10	9	10	2	10	10	9
Box close 0°	10	9.5	10	10	0	0	10	10
Box close 45°	9	8.5	10	9.5	0	0	5	2
Box close 90°	10	10	10	10	9	0	10	9
Box close 135°	10	10	3.5	10	10	0	5	3
No box	10	10	8	7	10	10	0	0
Basic total	**59/60 (98.3%)**	**58/60 (96.7%)**	**50.5/60 (84.2%)**	**56.5/60 (94.2%)**	**31/60 (51.7%)**	**20/60 (33.3%)**	**40/60 (66.7%)**	**33/60 (55.0%)**
Box close 20°	10	10	-	10	-	-	-	4.5
Box close 60°	8	6	-	6.5	-	-	-	7.5
Box close 120°	10	10	-	9	-	-	-	10
Box close 160°	10	10	-	10	-	-	-	5
Generalisation total	**38/40 (95.0%)**	**36/40 (90.0%)**		**35.5/40 (88.8%)**	-	-	-	**27/40 (67.5%)**
Combined total	**97/100**	**94/100**	-	**92/100**	-	-	-	**60/100**

Key Results

- **Sufficient data per behaviour is critical**: with only six demonstrations, the model failed to learn complex actions (e.g. far-box pulling), showing that under-represented behaviours are not reliably acquired in low-data settings.
- **Behaviours must be clearly separable**: overlapping states (e.g. pick vs hand-over) caused behavioural interference, leading to incorrect actions. Distinct state-action regions are essential for reliable learning.
- **Strong generalisation is achieved**: the model reaches 95% success on unseen angles and shows robustness to lighting changes, with only a ~5% performance drop, indicating effective interpolation beyond training data.
- **Data efficiency depends on careful design**: reducing dataset size causes severe overfitting (33–66% performance), while increasing complexity also harms learning. Simplicity, coverage, and distinctness outweigh dataset size.

Conclusions

Pre-trained VLAs enable reliable humanoid deployment in the low-data regime, but success depends critically on carefully structured datasets that balance behaviour coverage, separation, and task simplicity.