Stable VLA: Towards Robust Vision-Language-Action Models without Extra Data

Yiyang Fu, Chubin Zhang, Shukai Gong, Yufan Deng, Kaiwei Sun, Qiyang Min, Qibin Hou, Yansong Tang, Jianan Wang, Daquan Zhou†

† Indicates coresponding author

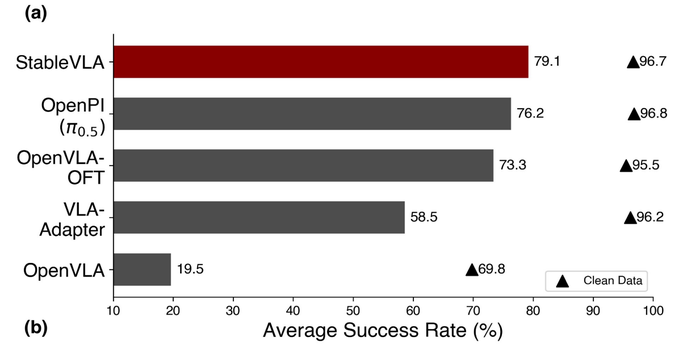
DAGroup PEKING UNIVERSITY

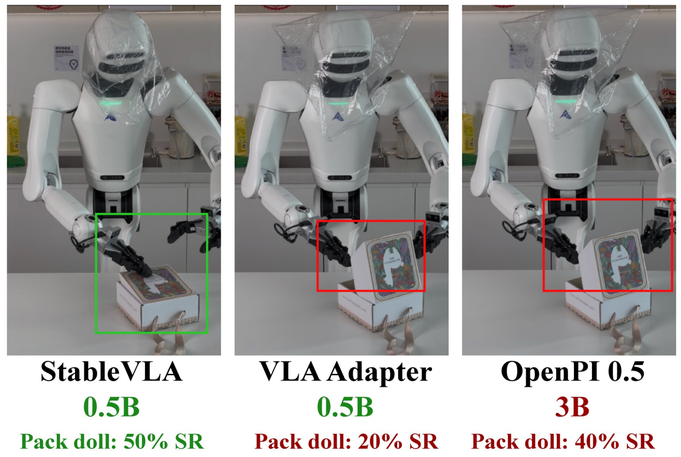
ICML International Conference On Machine Learning

Project Page | Yiyang Fu | Chubin Zhang | Shukai Gong

Summary

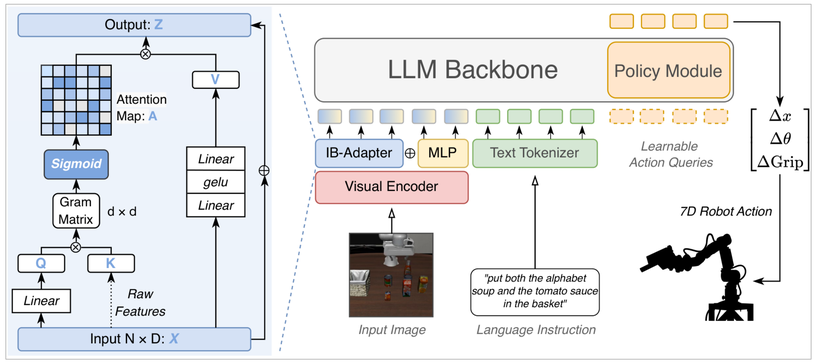
- Existing VLA models are fragile under unseen visual corruptions.
- The modality projector is a key source of robustness degradation.
- IB-Adapter filters task-irrelevant visual nuisances through channel-wise covariance modeling.
- StableVLA achieves competitive robustness with 7B-scale VLAs using a 14× smaller backbone.





Architecture of StableVLA

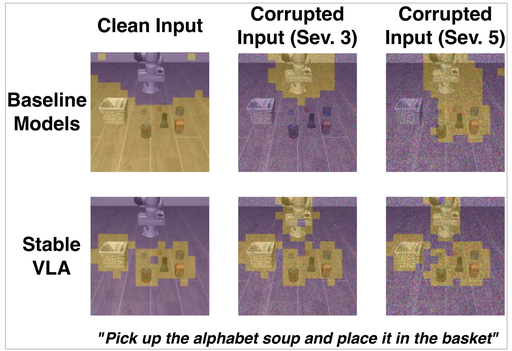
StableVLA replaces the standard MLP projector with a Fused IB-Adapter to improve the intrinsic robustness of VLA models. The design contains two complementary pathways: an MLP path that preserves fine-grained spatial information for precise manipulation, and an IB-Adapter path that models channel-wise covariance to suppress task-irrelevant visual nuisances before features are passed into the LLM backbone.



Why Does StableVLA Work?

This visualization further explains the effect of the Fused IB-Adapter. While the MLP branch preserves spatial details, the IB branch suppresses low-correlation noisy channels through covariance-based gating.

As a result, StableVLA maintains object-centric semantic grouping under high-severity corruptions, whereas the baseline projector produces scattered and unstable representations.



Results

Zero-shot Robustness in Simulation

StableVLA is evaluated on LIBERO and CALVIN under unseen visual corruptions in a zero-shot setting. Compared with VLA-Adapter, StableVLA substantially improves corrupted-task performance while preserving clean-task accuracy. On corrupted LIBERO, the average success rate increases from 58.5% to 79.1%, demonstrating that robustness can be improved through architecture rather than extra corrupted data.

Training Method	Method	Spatial				Object				Goal				Long				CALVIN			
		C	S3	S4	S5	C	S3	S4	S5	C	S3	S4	S5	C	S3	S4	S5	C	S3	S4	S5
OpenX Pretrain	OpenVLA (7B)	80.0	40.9	24.6	14.7	69.6	18.2	10.4	2.7	74.0	38.7	27.0	16.3	55.5	20.5	12.4	7.0	–	–	–	–
	OpenVLA-OFT (7B)	92.6	89.3	84.0	72.1	98.4	82.5	69.2	52.8	96.8	**94.5**	84.6	70.3	**94.4**	**77.6**	61.9	40.3	–	–	–	–
OpenX+Web Co-train	OpenPi–0.5 (3B)	**98.4**	88.3	79.0	62.4	**99.4**	**97.1**	**88.4**	**76.4**	97.2	87.2	82.5	64.2	92.0	76.1	**65.6**	**47.7**	–	–	–	–
VLM Direct FT	VLA-Adapter (0.5B)	96.0	93.7	83.3	58.5	96.8	71.0	44.1	29.3	97.4	79.5	64.7	47.3	94.4	63.5	41.0	26.2	4.14	2.56	1.89	1.44
	StableVLA (0.5B)	96.2	**94.4**	**92.1**	**82.0**	98.8	92.4	83.6	70.2	**98.0**	93.4	**85.0**	**71.9**	93.6	76.3	62.4	45.3	**4.17**	**2.77**	**2.11**	**1.51**

Real-world Robustness under Physical Corruptions

We further deploy StableVLA on real-world robot manipulation tasks with both synthetic and physical visual corruptions. The physical corruptions include oil-stained and plastic-shelter camera interference, which simulate common visual degradation during robot deployment. Across Pick and Place, Throw Basketball, Pour Water, and Pack Doll, StableVLA consistently shows the smallest performance drop, demonstrating that its robustness transfers beyond simulation.

Task	Method	Clean	Noise (Δ)	Blur (Δ)	Oil (Δ)	Shelter (Δ)	Avg. (Δ)
Pick and place	$\pi_{0.5}$	100.0	-63.3	-16.7	-10.0	-30.0	-30.1
	VLA-Adapter	80.0	-66.7	-40.0	-30.0	-60.0	-49.2
	StableVLA	80.0	**-30.0**	**-10.0**	**-10.0**	**-20.0**	**-17.5**
Throw basketball	$\pi_{0.5}$	80.0	-60.0	-33.3	-20.0	-30.0	-35.8
	VLA-Adapter	60.0	-53.0	-40.0	-20.0	-40.0	-38.3
	StableVLA	60.0	**-36.7**	**-16.7**	**-10.0**	**-10.0**	**-18.4**
Pour water	$\pi_{0.5}$	70.0	-60.0	-20.0	-20.0	-20.0	-30.0
	VLA-Adapter	40.0	-40.0	-30.0	-10.0	-20.0	-25.0
	StableVLA	40.0	**-23.3**	**-16.7**	**-0.0**	**-10.0**	**-12.5**
Pack doll	$\pi_{0.5}$	80.0	-63.3	-33.3	-30.0	-40.0	-41.7
	VLA-Adapter	50.0	-40.0	-26.7	-30.0	-30.0	-31.7
	StableVLA	60.0	**-16.7**	**-10.0**	**-20.0**	**-10.0**	**-14.2**