

# 3DS-VLA: A 3D Spatial-Aware Vision Language Action Model for Robust Multi-Task Manipulation

Anonymous Author(s)

Affiliation

Address

email

## A. Comparisons with Baseline

Although CoRL considers papers published within the last four months to be contemporaneous and thus not strictly required for comparison, we still make the effort to compare against SpatialVLA [1], which was released in January and accepted to RSS in late April 2025. This voluntary inclusion reflects our commitment to a thorough and up-to-date evaluation.

Since our method and SpatialVLA are developed using different simulation environments, directly migrating their approach to our setup would require substantial adaptation. Due to time constraints, we instead perform the comparison in the same real-robot setting used in our main paper. For Spatial VLA, we follow their LoRA fine-tuning configuration. As shown in Table 1, our method outperforms Spatial VLA by 6%. While both approaches incorporate 3D observations, our model further integrates spatiotemporal constraints on task-relevant objects, enhancing its ability to reason about both the robot and its environment for more effective task execution.

Table 1: We compare 3DS-VLA with SpatialVLA on 10 real-world tasks. \* denotes long-horizon tasks.

Models	Avg. Success $\uparrow$	Stack Cup	Pour Water	Pick Place*	Stack Block*	Water Plants	Bottle at Rack	Slide Box	Unplug Charger	Wipe Table	Open Drawer
SpatialVLA	0.48	<b>0.50</b>	0.30	<b>0.50</b>	0.30	0.50	0.40	<b>0.80</b>	<b>0.70</b>	0.40	0.40
Ours	<b>0.54</b>	0.40	<b>0.50</b>	0.40	<b>0.40</b>	<b>0.70</b>	<b>0.80</b>	0.60	0.40	<b>0.60</b>	<b>0.60</b>

## References

- [1] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.