Distracted Robot: How Visual Clutter Undermines Robotic Manipulation

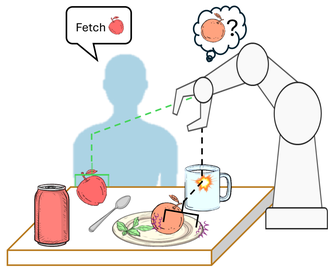
Amir Rasouli Montgomery Alban Sajjad Pakdamansavoji Zhiyuan Li Zhanguang Zhang Aaron Wu Xuan Zhao

Huawei Technologies Canada

Clutter cuts VLA success by up to 34%; a psychophysical clutter score predicts the drop, and policies with similar averages fail on very different scenes.

Motivation

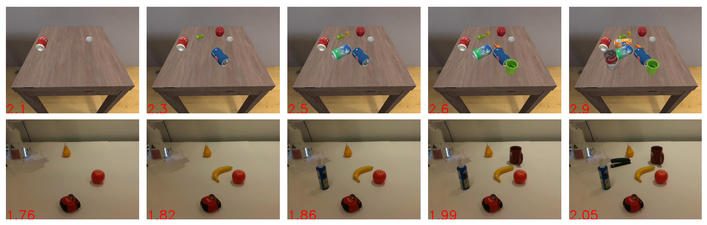
- Real homes are **cluttered**: distractors cause target confusion, collisions and failed grasps.
- Prior work varies clutter factors **one at a time**, so their combined effect is unknown.
- We need a **single, principled measure** of clutter to build fair evaluation scenes.



Expected (green) vs. executed (black) path when fetching an apple.

Protocol: a psychophysical clutter measure

- Start from **feature congestion** (Rosenholtz et al.): covariance of colour, contrast and orientation across scales.
- **Dual-view (DvFC)**: combine the robot's view with a **top-down view** to capture both visual clutter and operational complexity.



Synthetic (top) and real (bottom) scenes with their DvFC scores.

Scenario generation

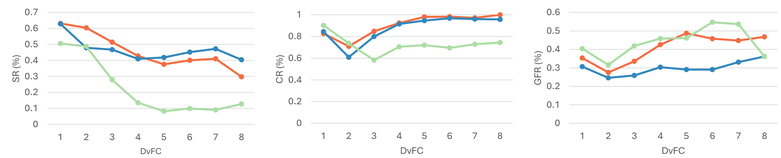
- 6 SIMPLER skills with **1–12 distractors** drawn from 61 YCB objects.
- Targets stay graspable and at most 50% occluded; scenes are binned by DvFC and **sampled uniformly**: **6,000 scenarios**.
- 5 VLAs: Octo, OpenVLA, SpatialVLA, π_0 and CogACT.

Result 1: clutter hurts every policy

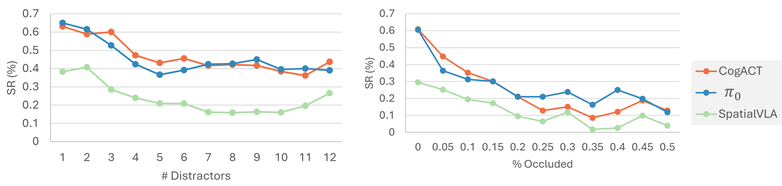
Policy	SR_{base} ↑	SR ↑	h-SR ↑	CR ↓	GFR ↓
SpatialVLA	0.59	0.24	0.08	**0.68**	0.44
π_0	0.71	0.47	0.09	0.84	**0.29**
CogACT	**0.74**	**0.48**	**0.10**	0.87	0.40

- Success falls by up to **34%** with distractors, and further when the target is partly occluded.
- High collision rates (CR): policies have **no real obstacle avoidance**.

Result 2: DvFC predicts the drop

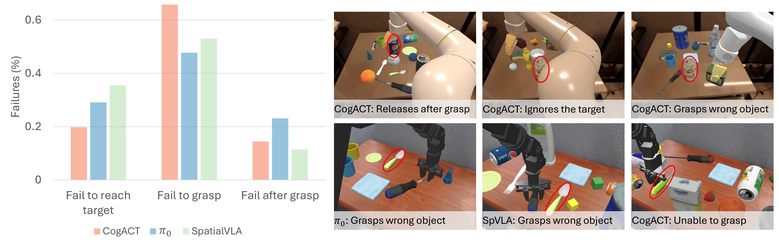


Success rate falls steadily as DvFC clutter increases.

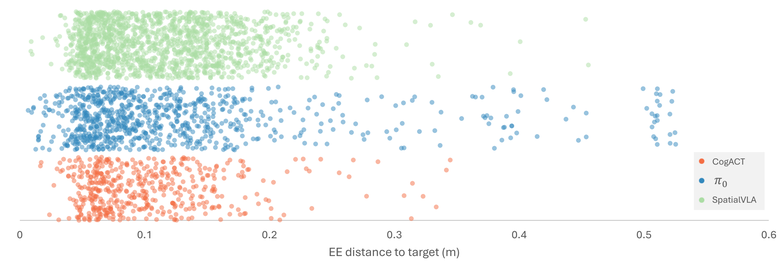


Number of distractors Occlusion of the target

Why policies fail

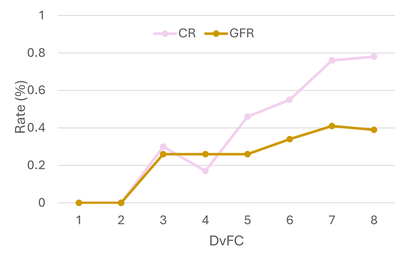


Left: failure sources per policy. Right: examples (targets circled in red).



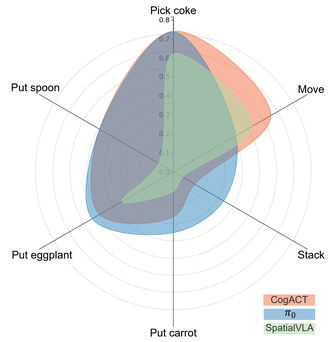
"Fail to reach target": closest end-effector distance before each failure.

Real robot: errors grow with clutter

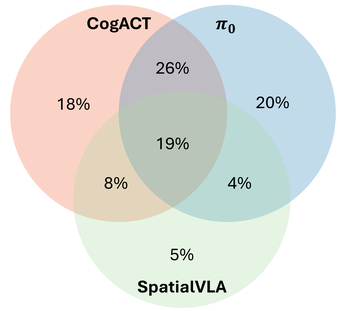


- Collision rate and grasp failure rate both **rise with DvFC** on the UR5e.
- Clutter hurts **acting**, not only perceiving.

Result 3: same average, different weaknesses



Per-task success

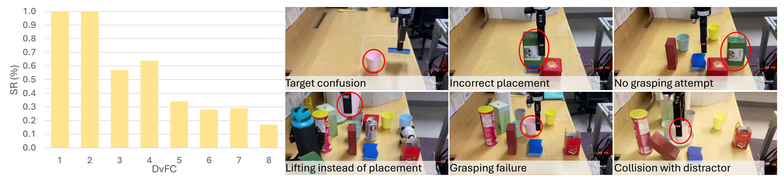


Overlap of solved scenes

- π_0 and CogACT both reach ≈48% SR, yet **agree on few successes**: averages hide vulnerabilities.

Real-world validation (UR5e, π_0)

- 4 skills × 6 distractor levels (0–16) × 9 layouts = **216 real scenes**.
- The same **downward trend** with DvFC holds on a real robot.



Success per DvFC bin (1–8) and real failure examples.

Does more data fix clutter?

	Base	+Distractors
SR ↑	0.47	**0.65**
h-SR ↑	0.33	**0.48**
CR ↓	0.38	**0.27**
GFR ↓	0.29	**0.23**

- Fine-tuning on cluttered demos: **+18%** success.
- Collisions and grasp failures improve **much less**.

Conclusion

- Evaluate VLAs **across clutter levels**, not only on average.
- Data helps, but robust manipulation needs **explicit scene reasoning**.