How VLAs (Really) Work In Open-World Environments

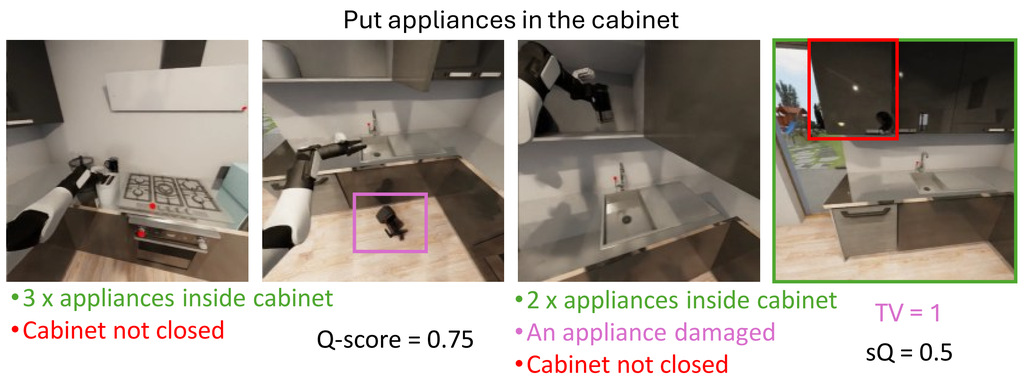
Amir Rasouli Yangzheng Wu Zhiyuan Li Rui Heng Yang Xuan Zhao
Charles Eret Sajjad Pakdamansavoji

Noah's Ark Lab, Huawei Technologies Canada

Final-state scores hide how a chore was done: top BEHAVIOR-1K VLAs are hard to reproduce, inconsistent and unsafe.

Motivation

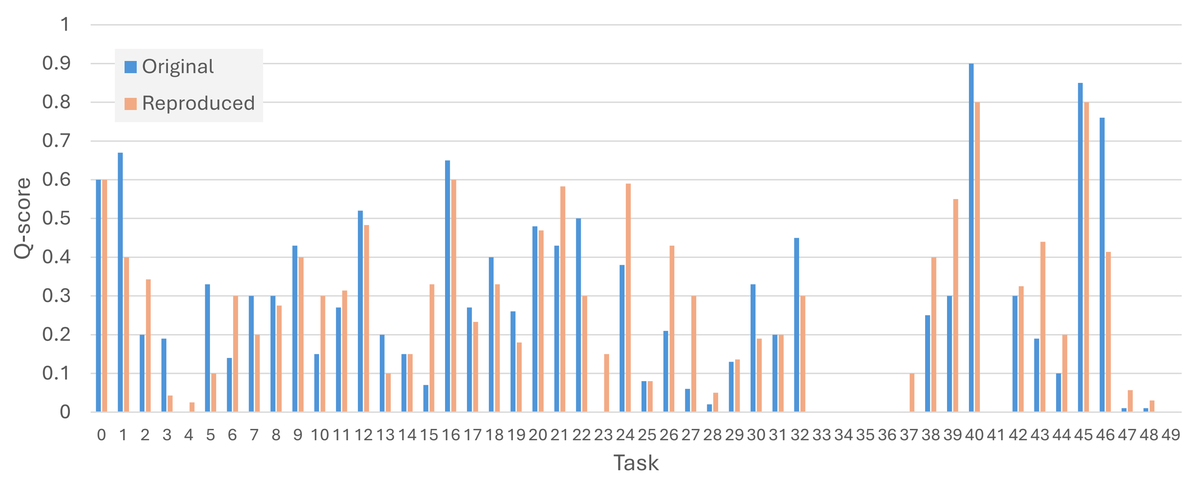
- VLAs now tackle **long-horizon household chores**, benchmarked on **BEHAVIOR-1K (B1K)**.
- The B1K **Q-score** only checks the **final state** of objects: it ignores *how* the robot got there.
- Dropping a toaster or knocking over a chopping board costs nothing, so reported progress can be **inflated**.



Top: 3 of 4 sub-goals met, so Q = 0.75, but a mixer was dropped. Bottom: the chopping board falls.

Finding 1: results are hard to reproduce

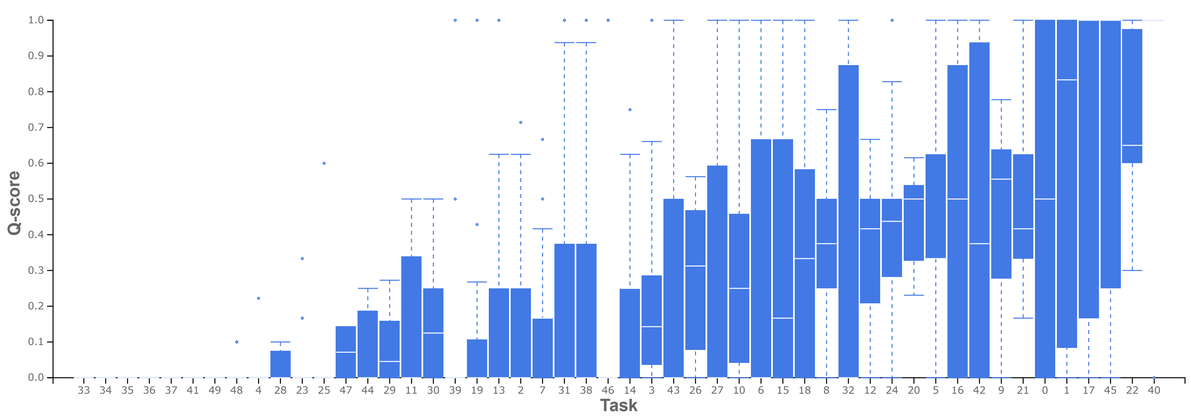
- We reran the **official checkpoints** of the top entry (RLC) with the official setup.
- Gaps of **over 27%** on several tasks: VLAs are sensitive to small rendering changes (lighting, shadows).



Posted vs. reproduced per-task Q-scores for RLC.

Finding 2: success is inconsistent

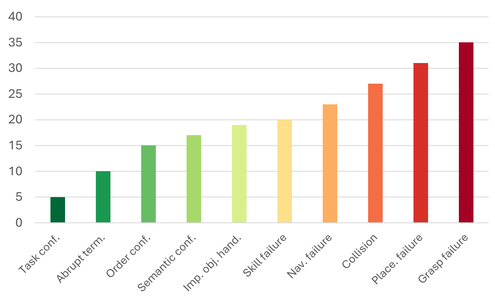
- About **half the tasks** are all-or-nothing across trials, with many outliers: success depends heavily on chance and scene layout.



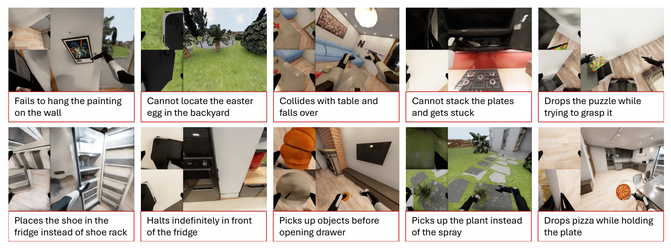
Q-score spread across trials per task (RLC), sorted by mean.

Finding 3: why tasks fail

- **8 robotics experts** reviewed **500 videos** (50 tasks × 10 trials) and labelled failures in 10 categories.
- **Grasp failure** is most common: flat objects need repositioning, and depth and localization errors show weak 3D understanding.
- **Collisions** are frequent even in uncluttered scenes; there is also task, order and semantic confusion.



Failure types (once per task).



Example RLC failures in B1K.

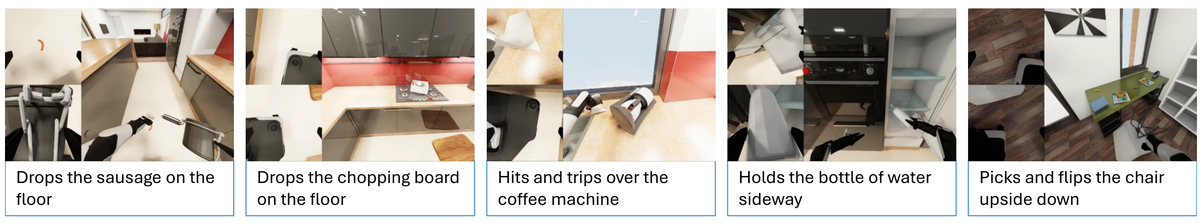
Safety-aware metrics

For N target sub-goals g_i and M support-object sub-goals s_j:

$$Q = \frac{1}{N}\sum_{i=1}^{N} g_i \qquad sQ = \frac{1}{N}\sum_{i=1}^{N} g_i\, p_i\, h_i$$

$$seQ = \frac{\sum_i g_i\, p_i\, h_i + \sum_j s_j}{N + M}$$

- p_i: placed **upright** (within 30°). h_i: critical object (appliance, fragile, full container, food) **not dropped**.
- s_j: support objects (table, board) not knocked over or moved more than 10 cm. Contact alone is fine.



Safety violations that now reduce the score.

Finding 4: safety lowers the score

Model	Q	sQ	seQ-Oracle	seQ	Violations
RLC	0.256	0.239	0.395	0.356	0.88
Comet	0.192	0.173	0.338	0.304	0.90

Average over 20 moderate B1K tasks; violations = target + non-target per task.

- Success is low (at most **26%**), yet safety still cuts the score by **up to 35%** on individual tasks.
- Both top models struggle on the **same tasks**.

Takeaways

- Report **reproducibility, consistency and safety**, not only final-state success.
- VLAs need **long-horizon memory**, replanning and better 3D grounding.