Improving Robotic Manipulation Robustness via NICE Scene Surgery

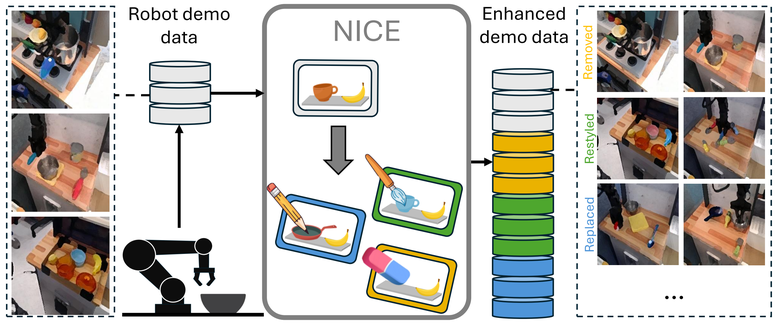
Sajjad Pakdamansavoji Mozhgan Pourkeshavarz Adam Sigal Zhiyuan Li Rui Heng Yang Amir Rasouli

Huawei Technologies Canada

Edit the demos you already have: +11% VLA success, −7% collisions and +21 points of affordance accuracy, with no new robot data.

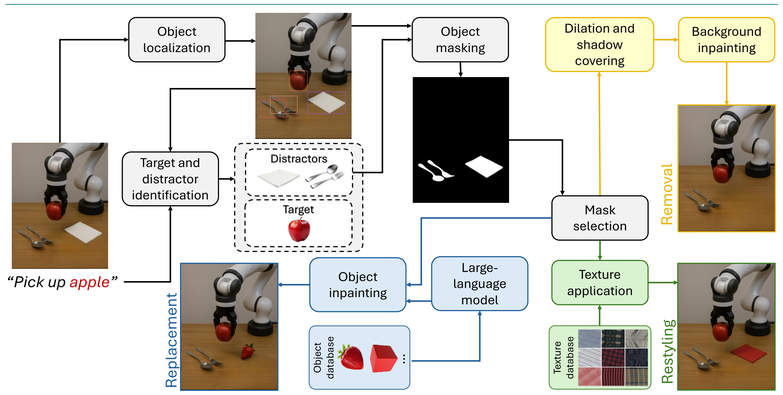
Motivation

- Behaviour-cloned policies **break under visual distractors** they never saw in training.
- Collecting more real demos for every clutter level is **slow and expensive**; simulators and custom generators add a sim-to-real gap.
- **Idea:** close the out-of-distribution gap by **editing existing demonstrations**, keeping actions valid.



NICE replaces, restyles or removes distracting objects to create new experiences.

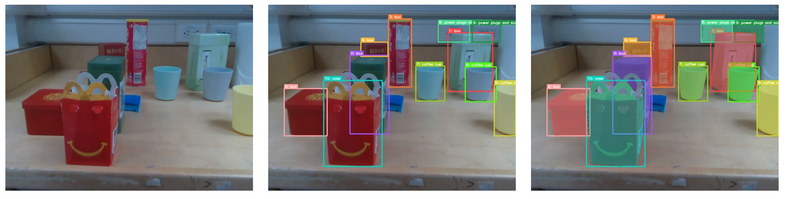
NICE: Naturalistic Inpainting for Context Enhancement



1. **Parse**: Florence-2 detects objects, SAM-2 segments them, and an LLM assigns roles (target vs. distractor).
2. **Edit distractors only**, so spatial relations and **action labels stay consistent**:
   - **Remove**: dilate masks and inpaint the background with LaMa.
   - **Restyle**: re-texture with DTD patterns and colour shifts.
   - **Replace**: Stable Diffusion inpainting of a scene-congruent object.

- No extra robot data, **no simulator**, **no custom model training**.

Parsing a scene



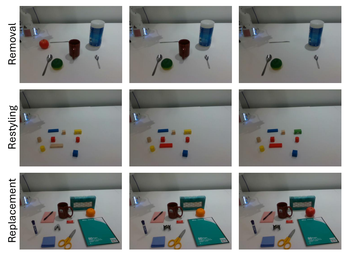
Raw image, Florence-2 detections, SAM-2 masks.

Edits on Bridge data



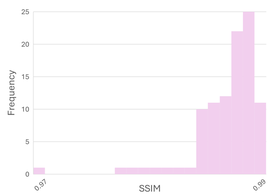
Each pair: original (left) and NICE edit (right).

Are the edits realistic?



Real scenes edited by hand to serve as ground truth.

- **100 real images**: remove, recolour or swap one object at a time, then replicate each change with NICE.
- Removal: **high SSIM** with the real background, shadows included.
- Restyle and replace: **low FID** against the real captures.



SSIM, removal



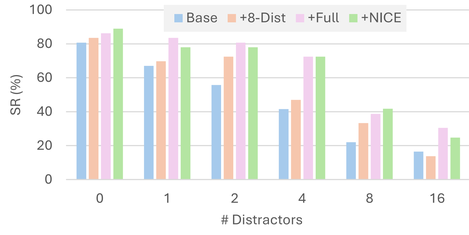
FID per operation

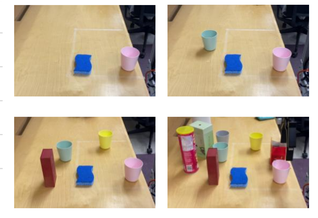
Conclusion

- **Scene surgery** on existing demos is a cheap, scalable fix for clutter.
- Works for both **VLMs** (affordance) and **VLAs** (manipulation).
- **Next:** add and rearrange objects, and test on a wider set of skills.

Real-robot manipulation (π_0)

- 4 skills (pick, move, put, stack) × 6 clutter levels (0–16 distractors) × 9 layouts: **864 real trials**.
- Fine-tune on Base plus one of: **8-Dist** (real, 8 distractors), **Full** (real, every clutter level) or **NICE** (edits of 8-Dist only).





Left: success vs. number of distractors. Right: example test scenes.

Data	SR ↑	CR ↓	TCR ↓
Base	47	38	34
+8-Dist	53	22	16
+Full (real)	**65**	27	14
+NICE	64	**15**	**10**

SR: success, CR: collision, TCR: target confusion (%).

- NICE beats its own input (8-Dist) by **+11%** SR and **matches Full**, which needs real demos at every clutter level.
- **Safest policy**: lowest collision rate and target confusion of all.
- Gains extend to **16 distractors**, beyond the 8 in the source data.

Spatial affordance (RoboPoint VLM)



Low, medium and high clutter.

Fine-tuning data	Low	Medium	High
Original	32.6	30.5	20.1
+NICE	**48.1**	**45.8**	**41.4**

Average prediction accuracy (%): **+21.4** points in high clutter.