Don't Get Distracted: Improving Robotic Perception Robustness via In-Context Visual Scene Editing

M. Pourkeshavarz*, A. Sigal*, S. Pakdamansavoji, Z. Li, RH. Yang, Amir Rasouli
Huawei Technologies Canada

CoRL
Data@CoRL25

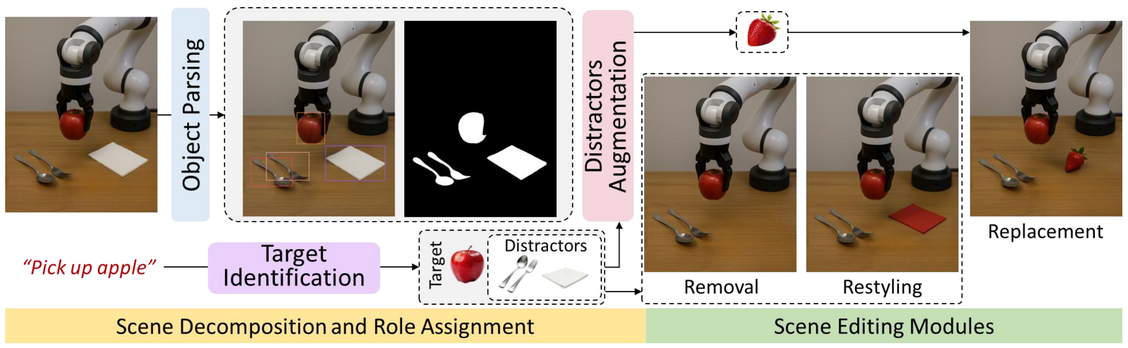
Motivation

- **Generalization across visually diverse environments** is fundamental for deploying robotic manipulation policies
- Learned **policies trained via imitation learning suffer from clutter** or context variation previously not observed during training
- **Generating large-scale robot demonstrations** capturing complex environment is **prohibitive**
- Existing data generation pipelines depend on **computationally expensive** simulators and suffer from sim2real gap

Contributions

- A novel **in-context visual scene editing (NICE)** strategy for large-scale data generation with minimal human involvement
- Experimental evaluation to **highlight the realism of data** generated using our pipeline
- Experiments on two downstream tasks, visual **affordance prediction** and **object manipulation** to validate the effectiveness of NICE data

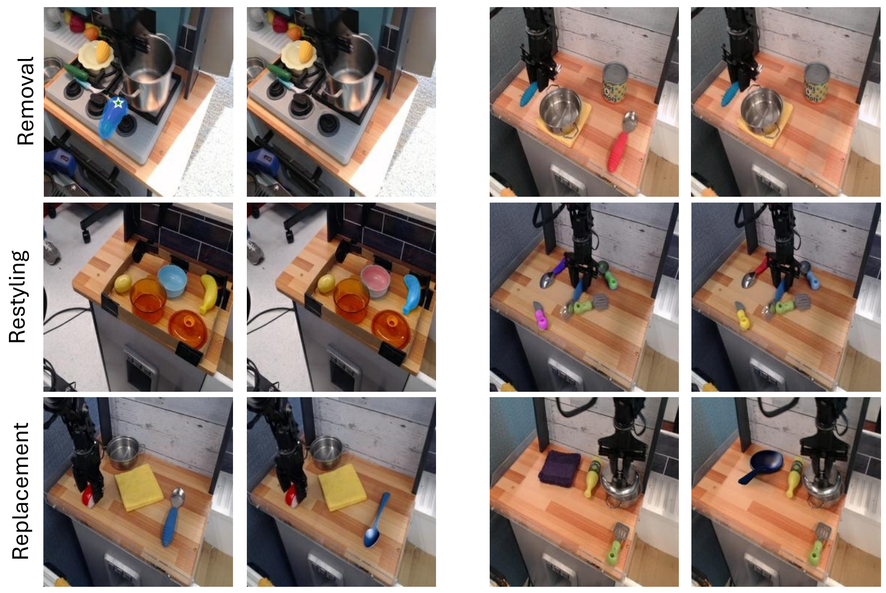
NICE Pipeline



An overview of the NICE data generation pipeline

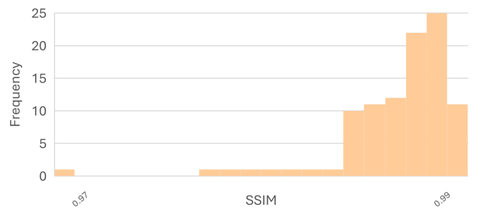
- **Object removal:** Detect objects, segment and mask them, dilate mask regions to cover shadows, and use LaMa inpainting to propagate background
- **Object restyling:** Using object masks apply textures from Describable Textures Dataset (DTD) and perform color and lighting adjustment
- **Object replacement:** Replace object with a novel but contextually relevant: 1) generate description of the existing object with Deepseek, 2) generate image of the object from this description using Stable Diffusion

Qualitative Examples on Bridge Data

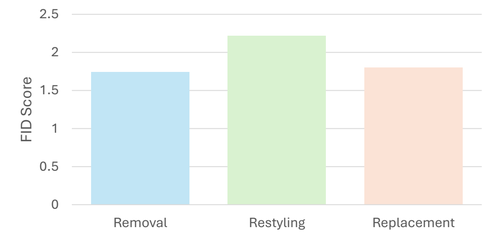


Realism of NICE

Real world examples of the three operations



Distribution of SSIM values for removal operation



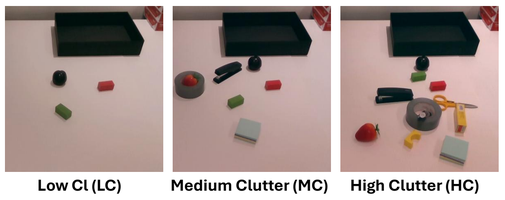
FID score of the three enhancement strategies

- Generate **real-world examples** by swapping/removing objects and recapturing the scenes
- Simulating the operations using **NICE pipeline yields high similarity**

Affordance

Average prediction accuracy (APA)(%) across different clutter levels using RoboPoint

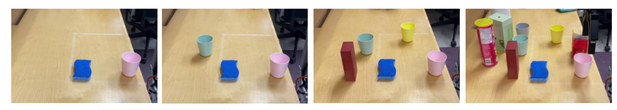
Dataset	APA_{LC}	APA_{MC}	APA_{HC}
Original	32.64	30.47	20.08
+NICE	**48.12** (+15.48)	**45.76** (+15.29)	**41.44** (+21.36)

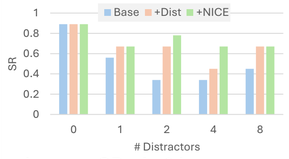


Low Cl (LC) | Medium Clutter (MC) | High Clutter (HC)

- Scenes with **levels of clutters**, placing 1-2 (LC), 5-8 (MC), and 11-15 (HC) objects
- Finetuning affordance prediction model on NICE is best for **very cluttered scenes** achieving as high as **21% improvement** in accuracy

Manipulation





Average SR of π0 in scenes with different level of clutter

Average performance of π_0 finetuned on different datasets

Data	*SR*↑	*CR*↓	*oCR*↓
Base	0.51	0.38	0.15
+Dist	0.65	0.09	0.07
+NICE	**0.74**	**0.06**	**0.02**

- We evaluate π_0 finetuned on 3 datasets,
  - **Base**: scenes with only targets
  - **Dist**: scenes with only 8 distractors and targets
  - **NICE**: data generated from Dist using our data generation pipeline
- **NICE boots performance**, especially on more cluttered scenes, by as much as 22%
- On average **NICE results in 23%** in SR compared to **Base** and **9%** to **Dist**
- **NICE lowers total collision rate** (CR), and obstacle CR (oCR) by **3% and 5%** respectively, making the policy trained on the data operate safer