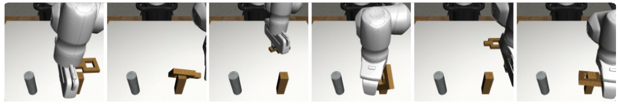
Can We Detect Failures Without Failure Data? Uncertainty-Aware Runtime Failure Detection for Imitation Learning Policies

Chen Xu, Tony Khuong Nguyen, Emma Dixon, Christopher Rodriguez, Patrick Miller, Robert Lee, Paarth Shah, Rares Andrei Ambrus, Haruki Nishimura, Masha Itkina

Motivation

Generative imitation learning policies are prone to failure:



Challenges :

- High-dimensional action and observation data.
- Demonstration data contain only successful trajectories.
- Diverse failure types occur during deployment.

Solution: A modular two-stage runtime failure detector

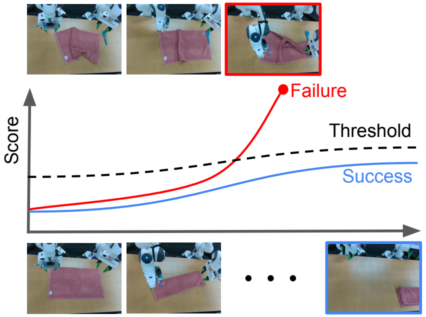
✔ Extracts **scalar scores** from high-dimensional data and uses conformal prediction to threshold when to alert failure.

✔ Requires **no failure** training data.

✔ Capable of detecting **different kinds of failures**.

Proposed Framework

- **Stage 1:** Extract scalar detection scores given data in each rollout.
- **Stage 2:** Determine detection threshold using conformal prediction band.



- **Sequentially** detect failures if scores exceed thresholds.
- Alarm is raised under **physical changes** in the environment.
- **Flexible** to:
  - Incorporating new scores and thresholding schema.
  - Building on *any* imitation learning policy.

Stage 1: Scalar Score Design

Desiderata:

1. **One-class**: No failure data is required.
2. **Light-weighted:** Fast inference for real-time robot control.
3. **Discriminative:** Gap in scores between successes/failures.

Based on **SOTA OOD detectors:**

(a) learned data density

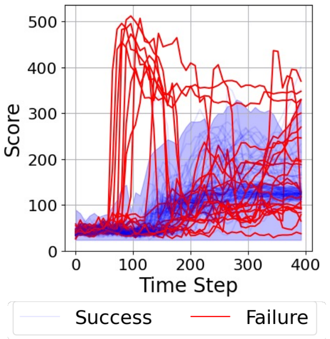
(b) second-order distribution

(c) one-class discriminator

(d) post-hoc metrics

Stage 2: Sequential Threshold

- Construct thresholds as a one-sided conformal prediction band.
- Threshold **adapt temporally** to score variations.
- **Theoretically controls** false positive rate.



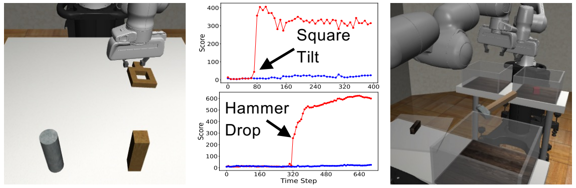
References

[1] *Christopher Agia et al.* **Unpacking Failure Modes of Generative Policies: Runtime Monitoring of Consistency and Progress. CoRL '24**

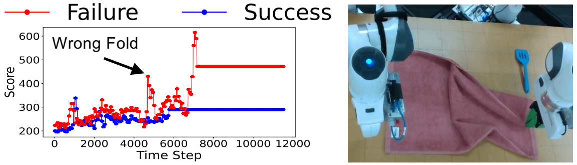
Experimental Results

Physically Meaningful Metric

Sudden rise in scores indicates failure has occurred.



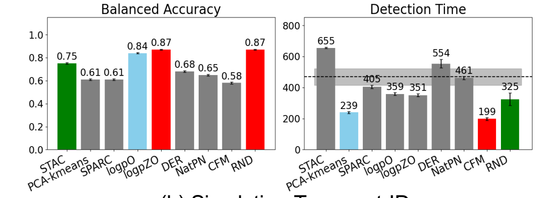
(a) Simulation-Robomimic



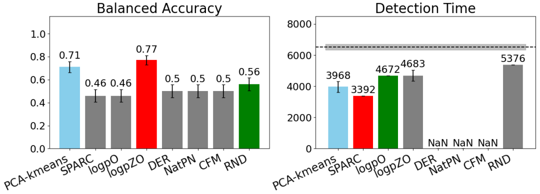
(b) On-Robot-OOD

Quantitative Comparison

- Top three: red > blue > green
- Our proposed *logpZO* performs best in Accuracy.
- No batch sampling: significantly faster than STAC [1].



(b) Simulation-Transport-ID



(b) On-Robot-OOD