FailureGen: Scalable Robot Failure Data Generation with World Models

Akul Santhosh
akuls@nvidia.com

Chaitra Sathe
csathe@nvidia.com

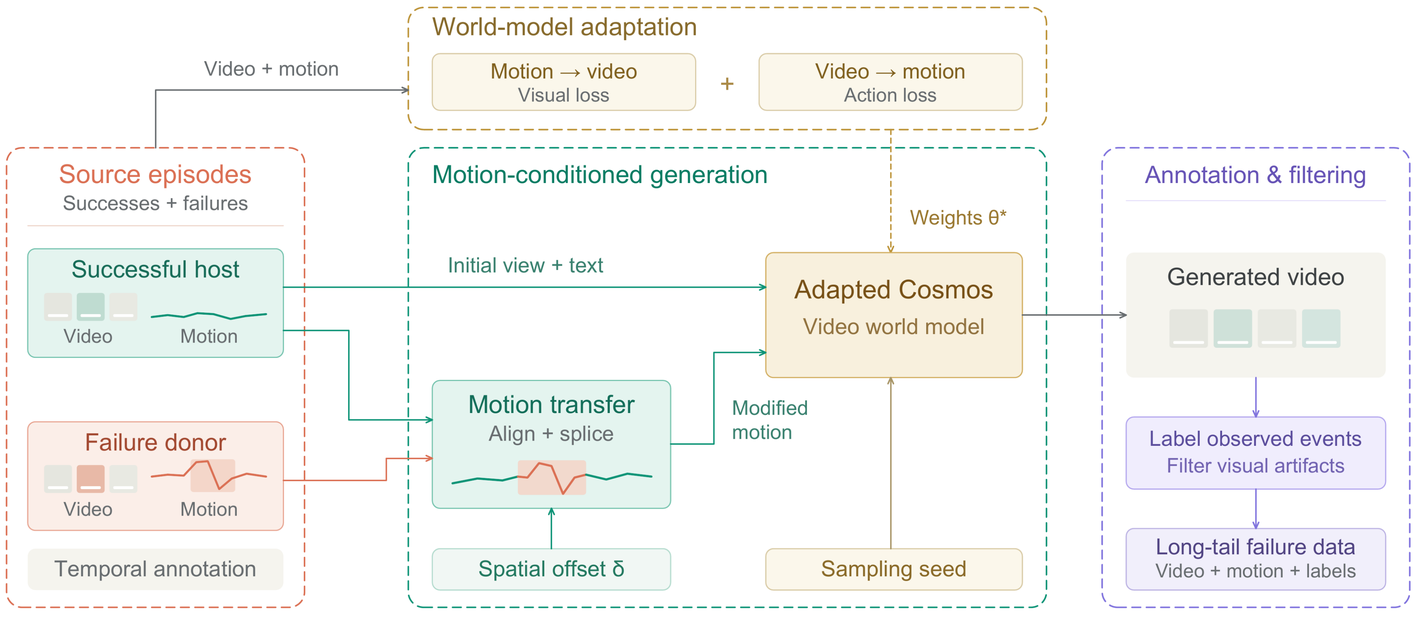
Aastha Jhunjhunwala
aasthaj@nvidia.com

Project page

“Oops, I erred” Workshop at CoRL 2026

TL;DR Robot failures are hard to collect at scale. FailureGen transfers recorded failure motion into compatible host scenes. An adapted world model produces videos. Event labeling and filtering yield temporally annotated video and motion pairs. Preliminary evaluations in two settings show improved failure detection on real recordings.

1 Recorded motion becomes annotated failure data



Architecture from the paper

Motion intervention

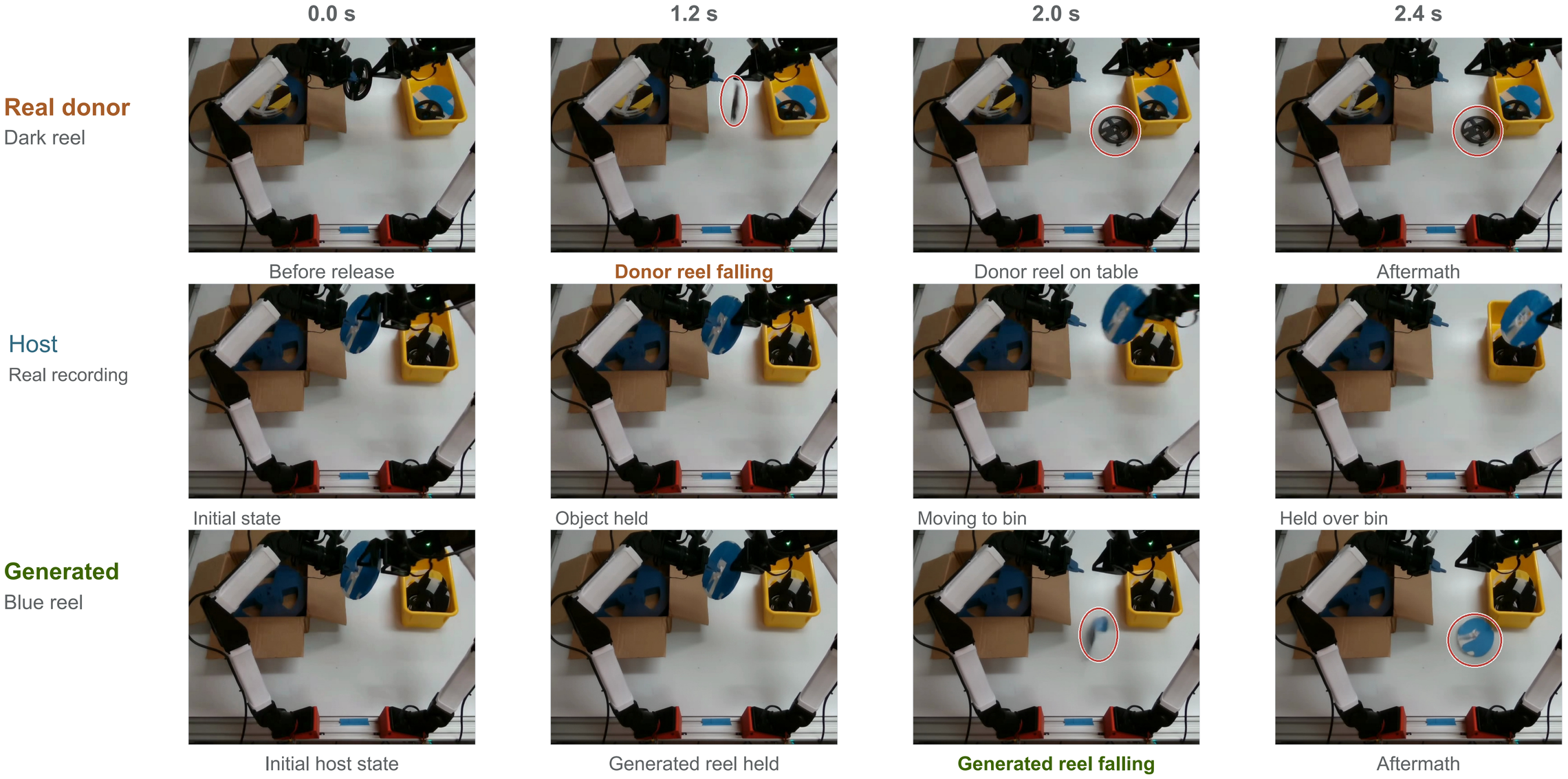
Nominal host motion

Donor segment in the host window

host **donor** host

Selected channels change inside the intervention.
Other channels retain host conditioning.
Schematic; optional offsets vary the intervention.

2 The failure unfolds with different timing in the host scene



3 Utility on real recordings

Industrial manipulation

Cosmos3-Edge detector, separate real recording session
66 real training episodes. Augmentation adds 75 clips.

Event recall **77.8%** (Real only (28/36)) ➡ **91.7%** (With FailureGen (33/36))

Average precision **0.41 → 0.51**

Clean episodes with alarms **5 / 41 → 0 / 41**

Preliminary comparison: sampling and training budgets differ.

ViFailback: 3 tasks, 2 model sizes

Episode accuracy (%) on real evaluation videos
Real only → FailureGen augmentation

Task	Qwen3-VL-2B	Qwen3-VL-8B
Object into box	**94.1 → 100.0**	**98.5 → 100.0**
Vial / test-tube rack	**58.5 → 58.5**	**49.6 → 54.1**
Drawer / can	**84.4 → 97.0**	**94.8 → 97.0**

Accuracy and failure recall improve in 5 of 6 combinations.
The remaining combination is unchanged.