SAFE: Multitask Failure Detection for Vision-Language-Action Models

Qiao Gu, Yuanliang (Avery) Ju, Shengxiang (Owen) Sun,
Igor Gilitschenski, Haruki Nishimura, Masha Itkina, Florian Shkurti

NeurIPS 2025

vla-safe.github.io

UNIVERSITY OF TORONTO
Robotics Institute
VECTOR INSTITUTE
TOYOTA RESEARCH INSTITUTE

VLAs still have limited success rates and diverse failure modes.

"Take toast out of toaster"

"Replace the paper towel"

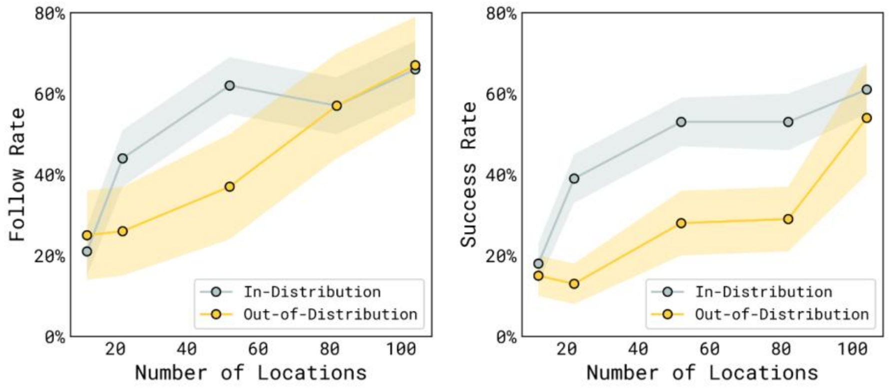
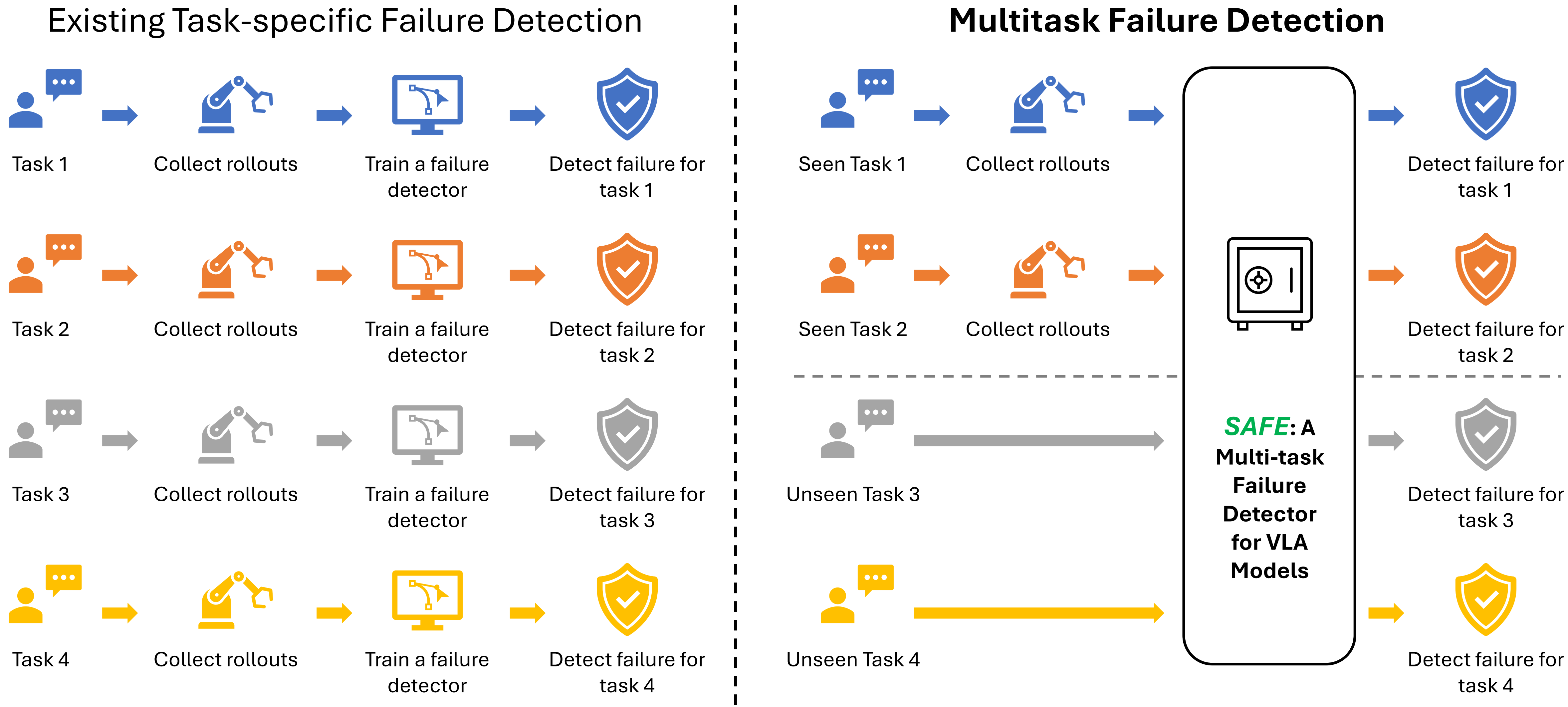


Fig. 9: **Evaluating language following with different numbers of training locations.** We evaluate language following rate and success rate for picking up user-indicated items and placing them into drawers or sinks, averaged over seen object categories ("in-distribution") or unseen categories ("out-of-distribution"). Performance increases steadily as we increase the number of training locations.

SOTA VLAs achieve <80% average task progress.

We need a failure detector for safe and reliable deployment of VLA models.

Generalist VLAs need multitask failure detectors

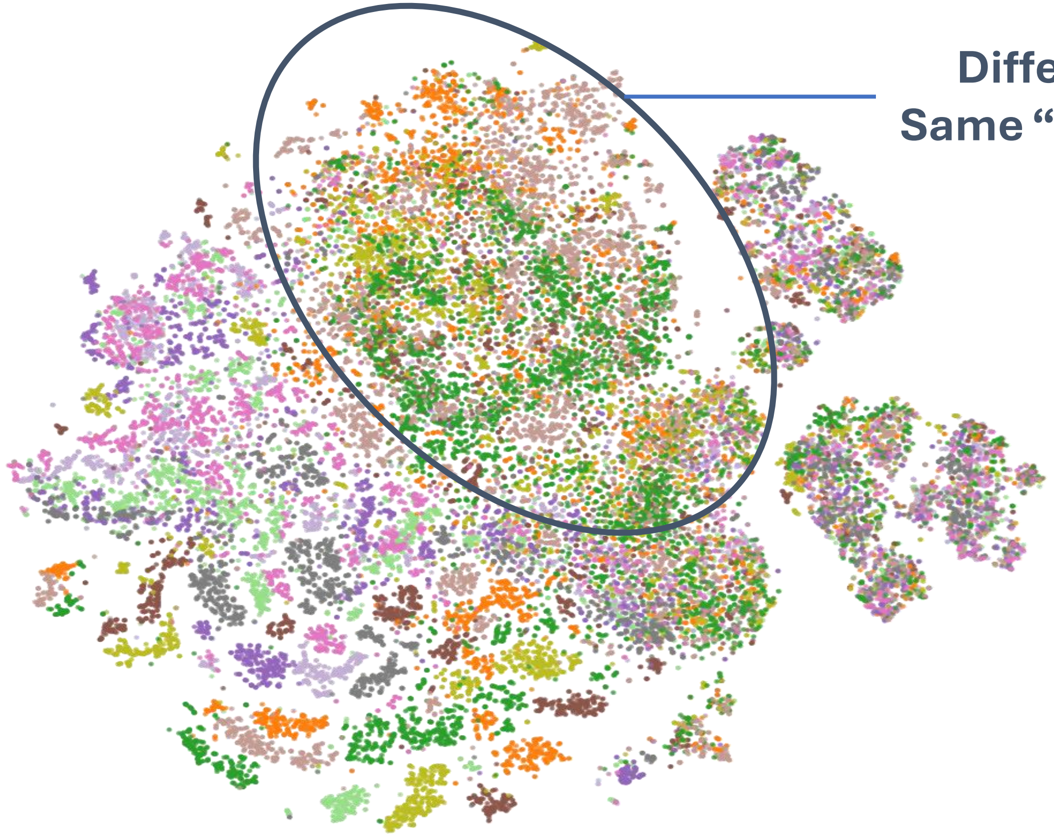


- × **No cross-task generalization**
- × **Labor-intensive**
- × **Only for task-specific policies**

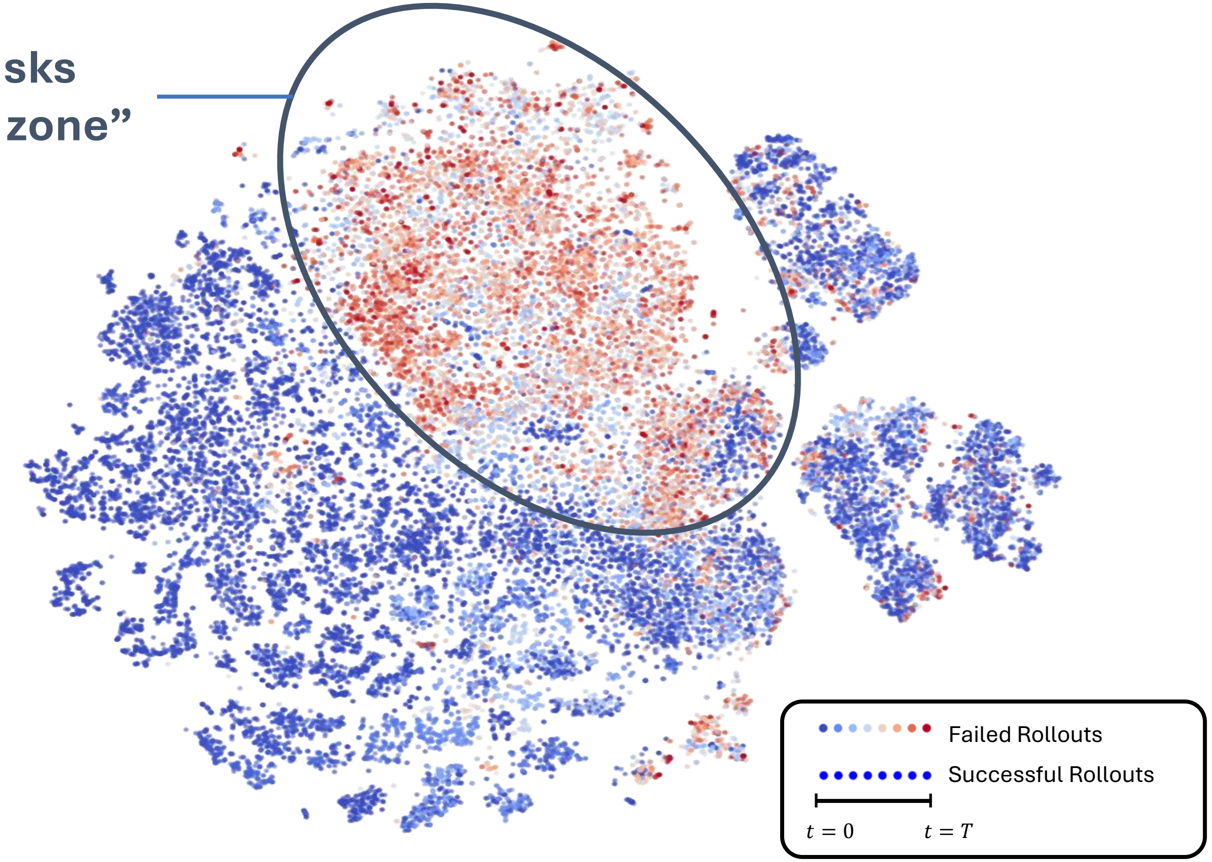
- ✓ **Work for unseen task zero-shot**
- ✓ **Avoid data collection and re-training**
- ✓ **For generalist policies like VLAs**

Key insight: VLA captures high-level knowledge about task failure in its feature space

- VLAs may have high-level semantic knowledge in its feature space.

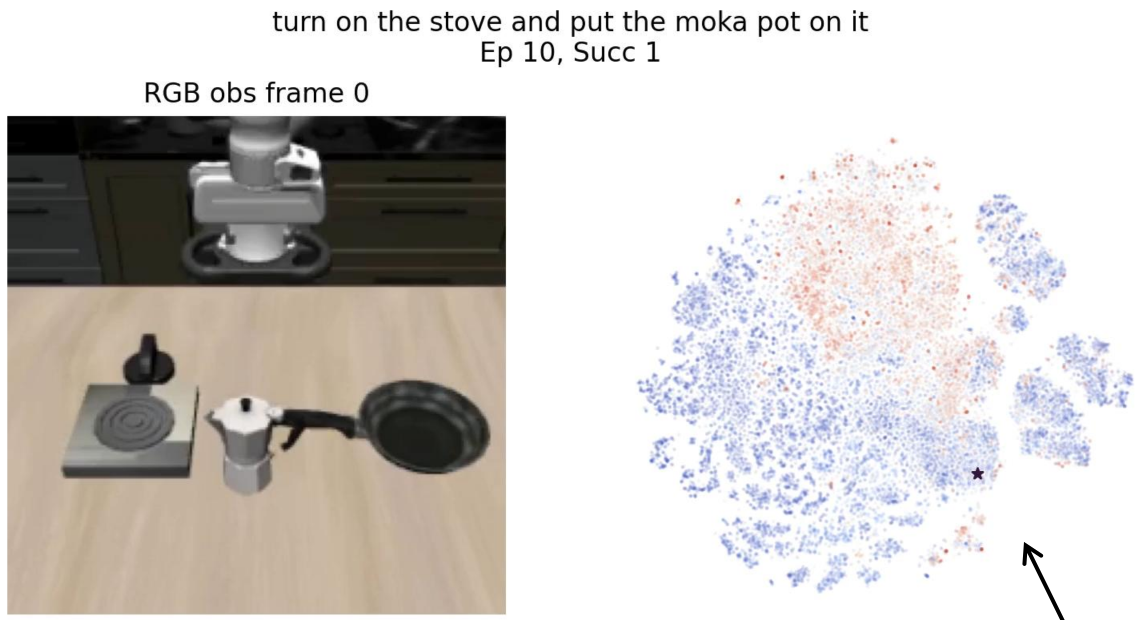


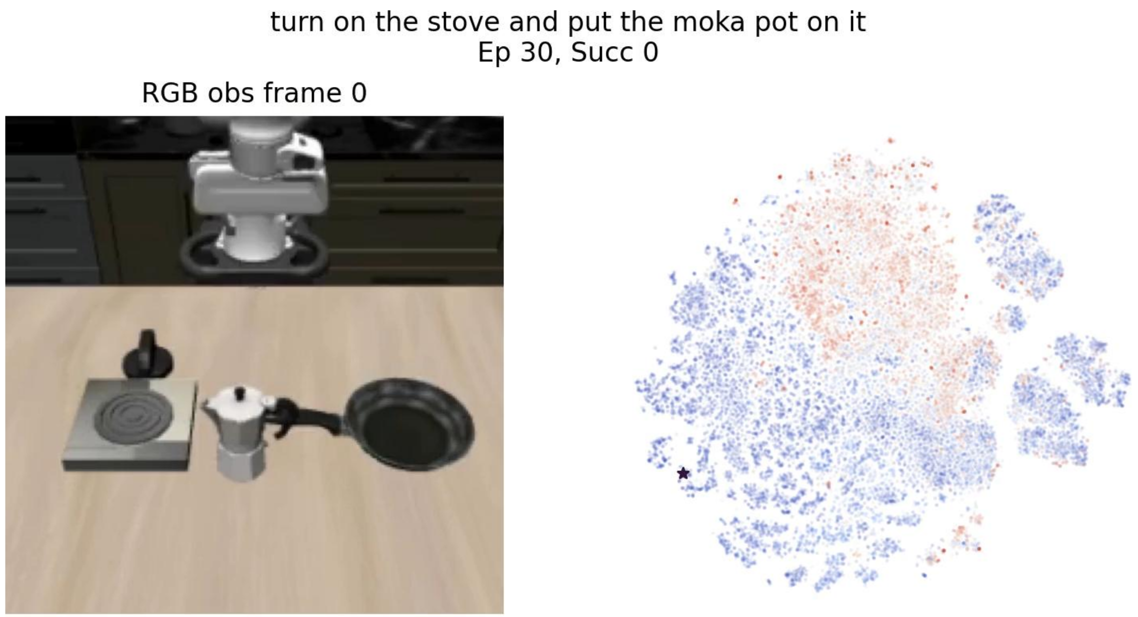
t-SNE of policy latent features, colored by task ID



t-SNE of policy latent features, colored by task failures

How Features Evolve in the Feature Space?



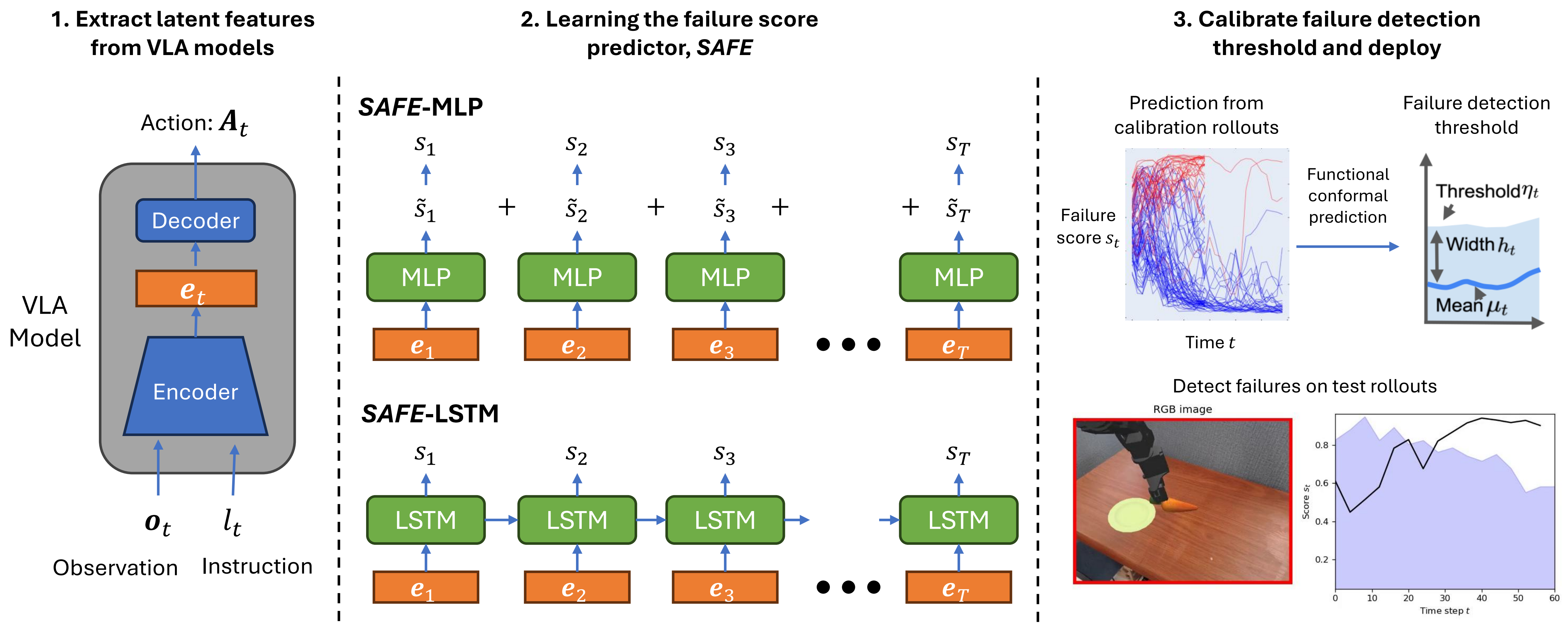


Successful rollouts: Embeddings always stay out of the red “failure zone”.

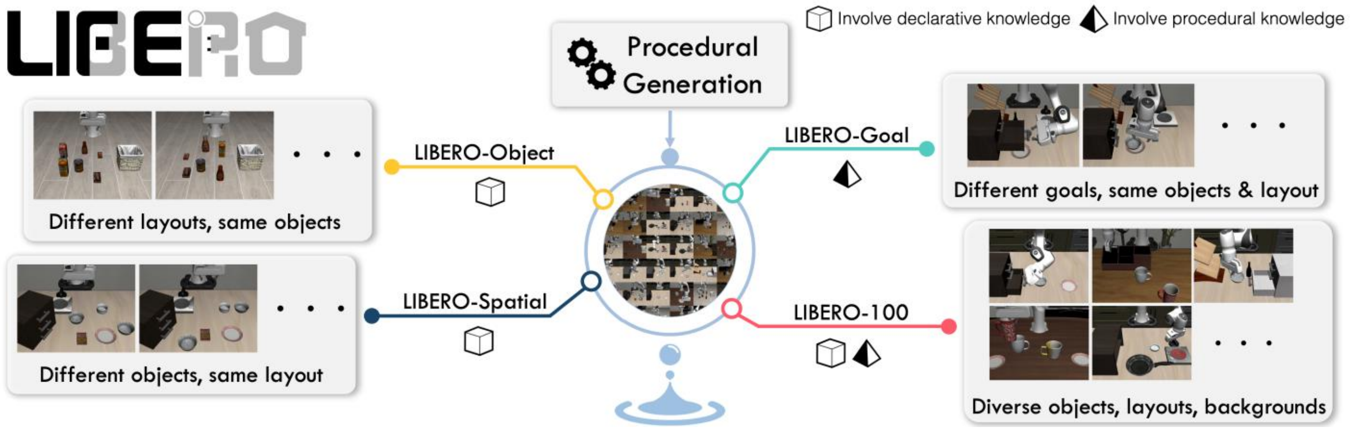
Failed rollouts: Robot drops the pot by accident. Embeddings go into the “failure zone”.

VLA Embeddings in this rollout are visualized as popping stars.

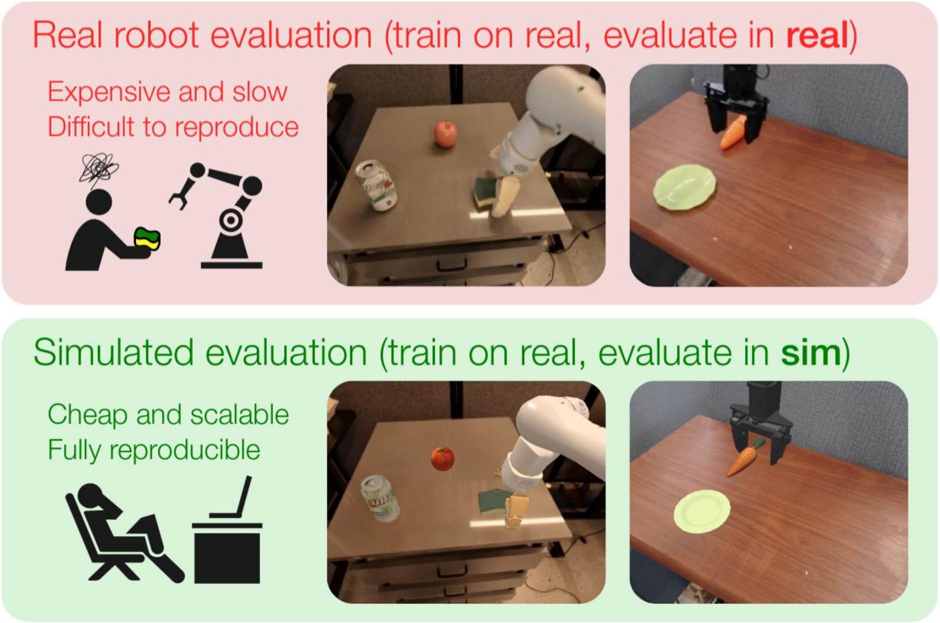
Multi-task Failure Detector based on VLA Internal Features



Simulation Experiment Setup



LIBERO-10

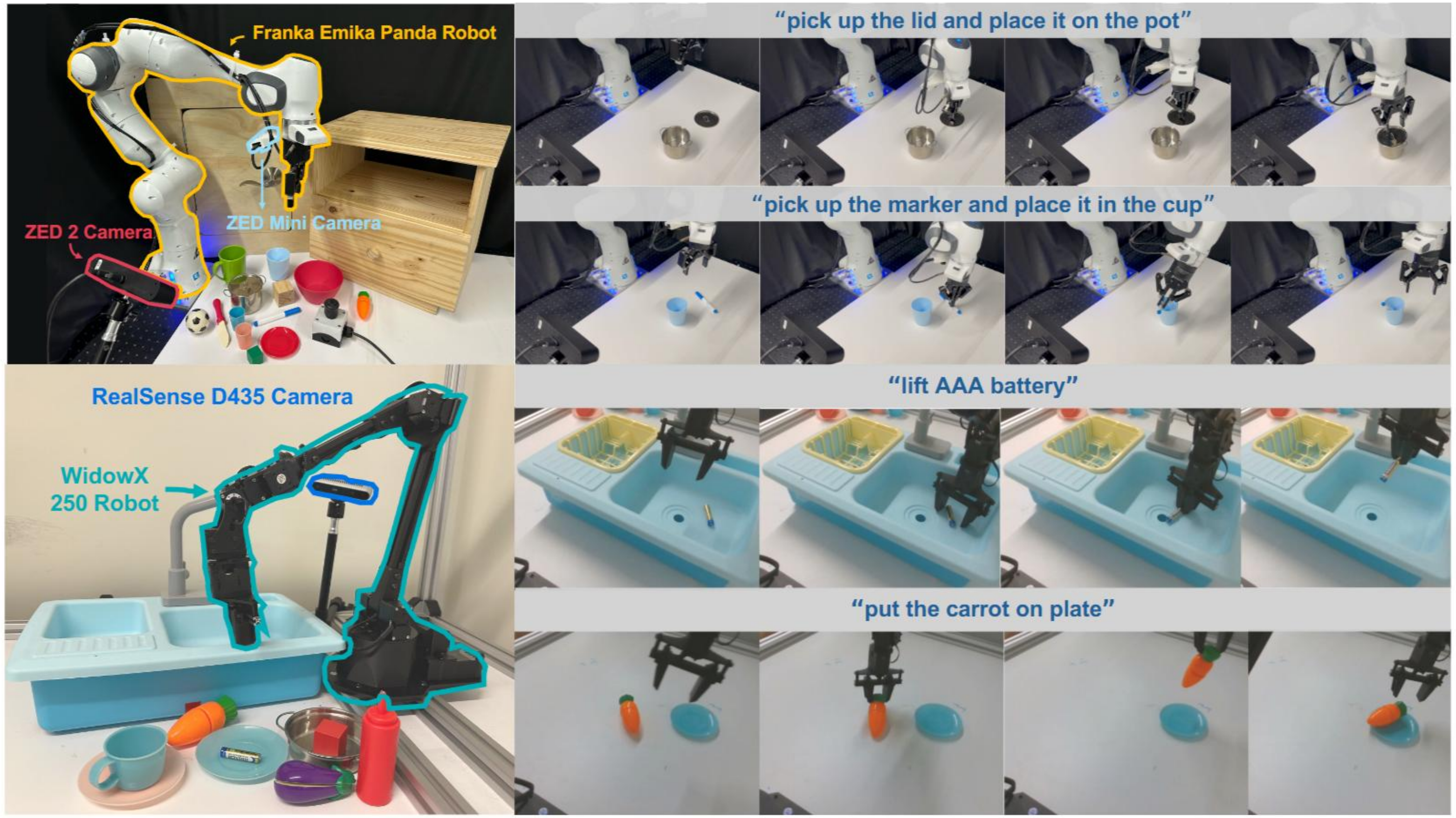


SimplerEnv

Real-world Experiment Setup

π_0-FAST on Franka

OpenVLA on WidowX



SAFE outperforms diverse baselines in both simulation and real world

	VLA Model / Benchmark / Eval Task Split	OpenVLA LIBERO Seen	OpenVLA LIBERO Unseen	π_0-FAST LIBERO Seen	π_0-FAST LIBERO Unseen	π_0 LIBERO Seen	π_0 LIBERO Unseen	π_0^* SimplerEnv Seen	π_0^* SimplerEnv Unseen	Average Seen	Average Unseen
Token Unc.	Max prob.	50.25	53.83	61.32	69.44	-	-	-	-	55.79	61.64
	Avg prob.	44.05	51.58	52.46	58.04	-	-	-	-	48.26	54.81
	Max entropy	52.94	53.09	46.69	62.96	-	-	-	-	49.81	58.03
	Avg entropy	45.27	50.03	50.93	58.63	-	-	-	-	48.10	54.33
Embed. Distr.	Mahalanobis dist.	62.03	58.85	**93.56**	83.79	**77.12**	**74.31**	88.42	52.84	80.28	67.45
	Euclidean dist. k-NN	66.00	55.23	92.04	84.12	75.64	70.73	**89.73**	68.41	80.85	69.62
	Cosine dist. k-NN	67.09	69.45	92.09	**84.64**	75.76	70.31	**90.19**	71.32	81.28	73.93
	PCA-KMeans [9]	57.18	55.10	68.46	57.12	64.92	60.35	66.88	61.19	64.36	58.44
	RND [39]	52.57	46.88	88.67	81.57	71.92	69.44	85.07	65.89	74.56	65.95
	LogpZO [8]	61.57	52.91	91.52	83.07	76.80	73.23	88.79	74.66	79.67	70.97
Sample Consist.	Action total var.	62.76	65.43	76.95	74.50	77.20	75.18	68.41	67.94	71.33	70.76
	Trans. total var.	55.33	58.99	78.21	80.03	49.38	54.71	63.27	55.90	61.55	62.41
	Rot. total var.	47.85	55.30	80.87	77.29	52.94	61.06	58.07	62.10	59.93	63.94
	Gripper total var.	61.84	64.48	76.82	74.42	77.19	75.19	69.16	69.29	71.25	70.84
	Cluster entropy	50.16	51.44	80.22	80.53	76.19	72.12	68.25	73.66	68.71	69.44
Action Consist.	STAC [18]	-	-	83.07	85.31	46.55	47.91	60.74	62.21	63.45	65.14
	STAC-Single	-	-	85.46	81.16	68.46	69.39	68.71	70.40	74.21	73.65
SAFE (Ours)	*SAFE*-LSTM	**70.24**	**72.47**	**92.98**	**84.48**	**76.98**	71.09	88.85	**80.11**	**82.26**	**77.04**
	SAFE-MLP	**72.68**	**73.47**	90.06	80.44	73.50	**73.27**	89.50	**84.82**	**81.43**	**78.00**

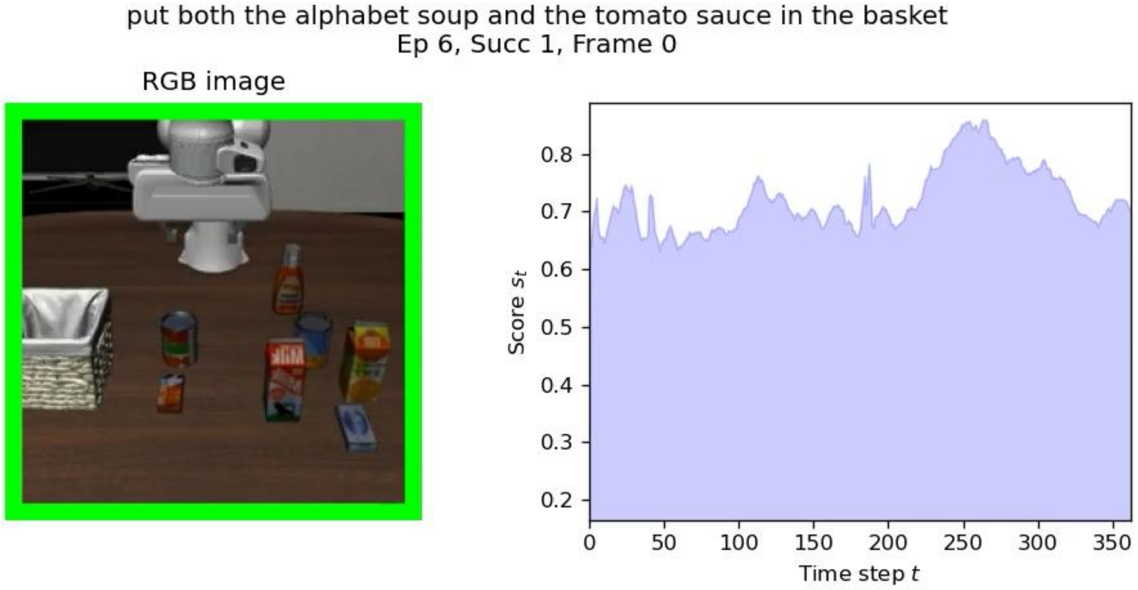
Failure detection ROC-AUC on simulation benchmarks

Method	π_0-FAST Franka Seen	π_0-FAST Franka Unseen	OpenVLA WidowX Seen	OpenVLA WidowX Unseen
Max prob.	53.74	48.59	50.77	54.25
Avg prob.	51.60	47.30	48.94	44.36
Max entropy	59.23	53.50	51.88	49.19
Avg entropy	50.67	46.08	47.72	53.84
Mahala. dist.	75.54	53.93	82.37	70.00
Euclid. k-NN	**80.35**	**60.27**	72.01	53.64
Cosine k-NN	80.23	59.51	74.76	65.88
PCA-KMeans	49.98	51.03	75.62	47.22
RND	62.00	45.83	66.68	47.67
LogpZO	64.43	52.24	62.94	51.32
STAC-Single	45.24	38.01	–	–
SAFE-LSTM	77.27	58.70	**84.29**	**71.80**
SAFE-MLP	**86.76**	**64.16**	**89.11**	**88.42**

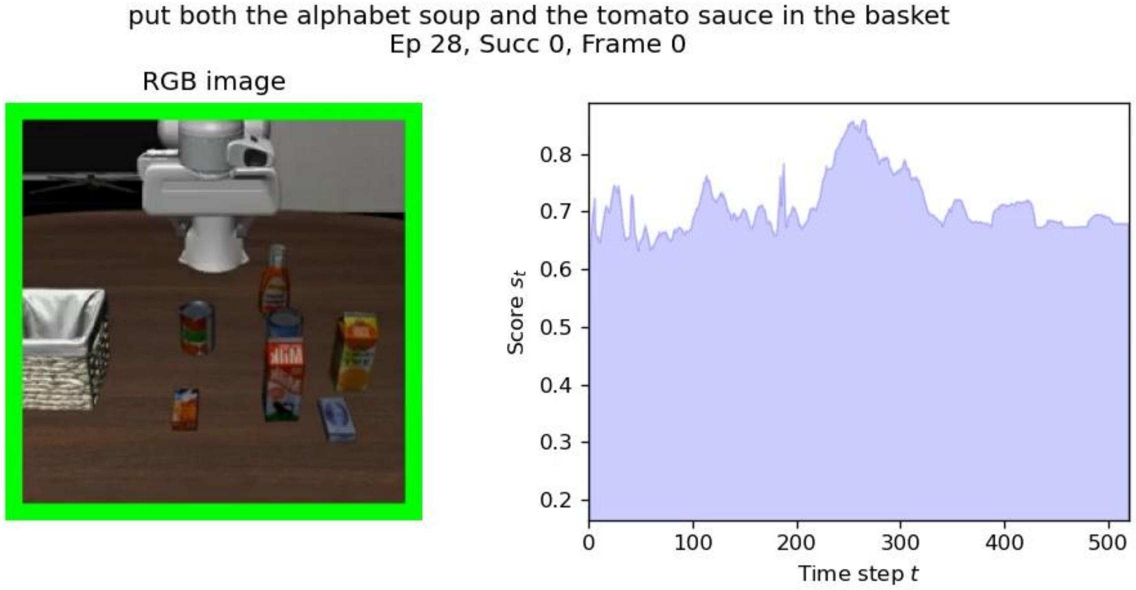
Failure detection ROC-AUC on real-world benchmarks

SAFE Detecting Failure in Simulation

- *SAFE*-LSTM on OpenVLA + LIBERO



Successful rollout



Failed rollout: When the robot gets stuck while picking up alphabet soup, it raises a failure signal

Thank you!

Paper & code: vla-safe.github.io