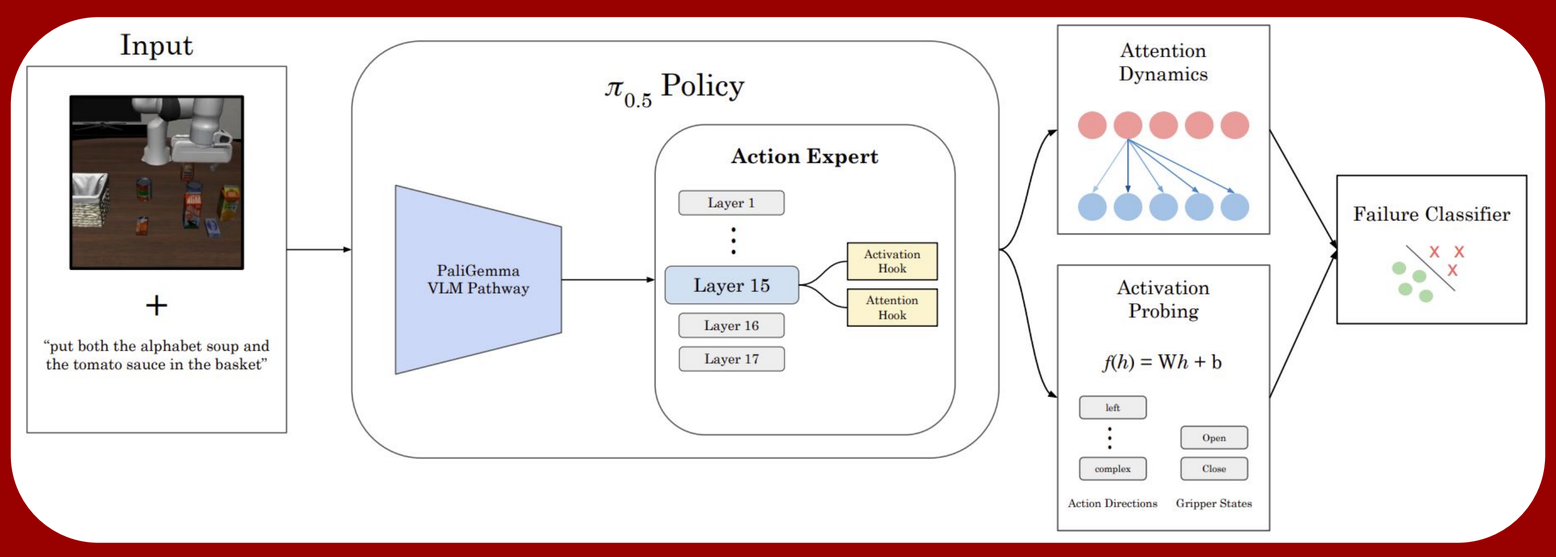


Action-direction Probes Inside $\pi_{0.5}$ Predict Robot Task Failures With AUROC 0.917

Interpretable Signals for Vision-Language-Action Model Failure

Tony Yang, Aidan Mokalla

1 TL;DR

MAIN 3 TAKEAWAYS

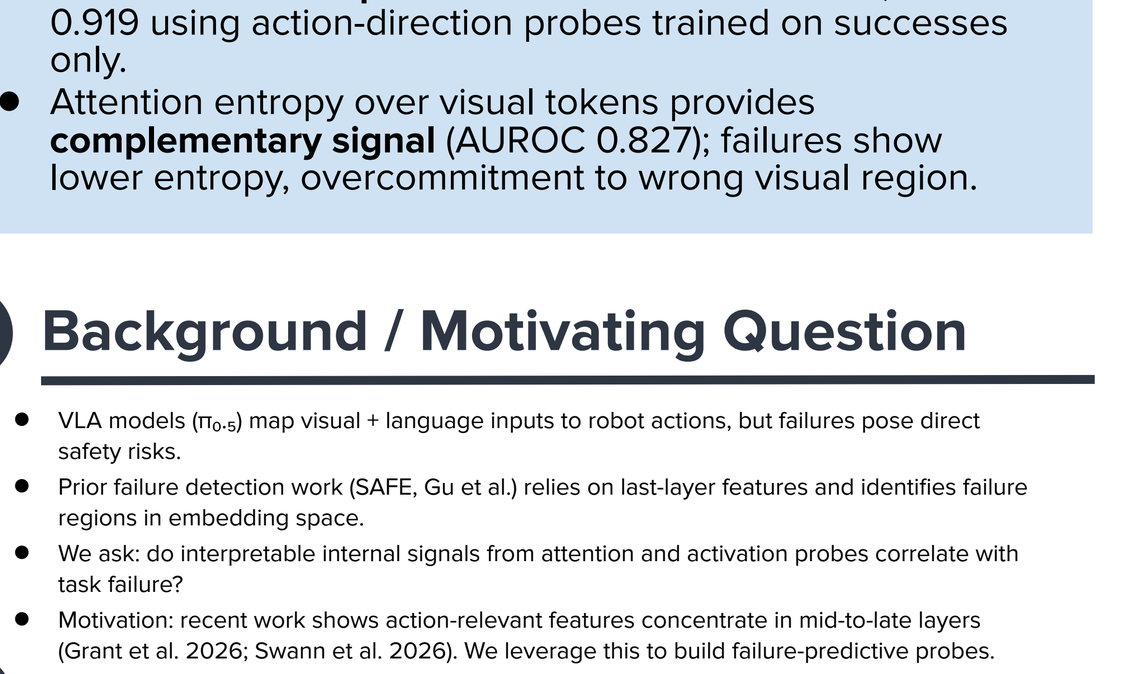
- Action-direction is **linearly decodable** from $\pi_{0.5}$ action-expert mid-to-late layers (L15, acc. 0.707).
- Probe confidence **predicts failures**: AUROC 0.917, bal. acc. 0.919 using action-direction probes trained on successes only.
- Attention entropy over visual tokens provides **complementary signal** (AUROC 0.827); failures show lower entropy, overcommitment to wrong visual region.

2 Background / Motivating Question

- VLA models ($\pi_{0.5}$) map visual + language inputs to robot actions, but failures pose direct safety risks.
- Prior failure detection work (SAFE, Gu et al.) relies on last-layer features and identifies failure regions in embedding space.
- We ask: do interpretable internal signals from attention and activation probes correlate with task failure?
- Motivation: recent work shows action-relevant features concentrate in mid-to-late layers (Grant et al. 2026; Swann et al. 2026). We leverage this to build failure-predictive probes.

3 Methods

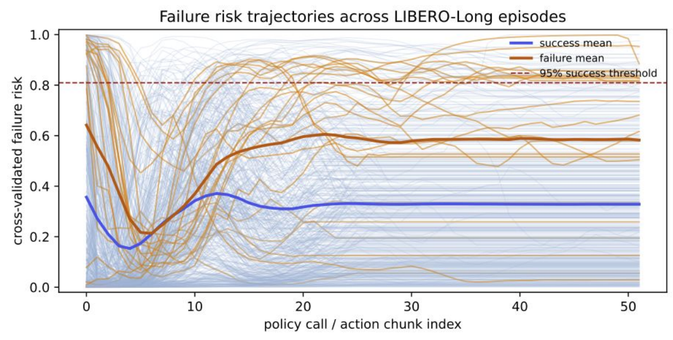
- Cached activations from $\pi_{0.5}$ Gemma action expert (L0–L17) on 500 LIBERO-Long rollouts (478 success, 22 failure). Main probe: expert L15 token 5, selected by layer/token sweep.
- Episode-level logistic regression classifier over s = [H_attn, c_dir, H_dir, c_grip, H_grip] with balanced class weighting. Also built chunk-level risk trajectories updated per policy call.



4 Results

- Action-direction probe (L15 token 5): AUROC 0.917, AP 0.725, bal. acc. 0.919 — dominant failure signal. Mean confidence lower in failures (0.646 vs. 0.755); low-conf. fraction higher (0.332 vs. 0.166).
- Risk trajectories (AUROC 0.744): 17/22 failed episodes cross the 95th-pct threshold. Early failures appear at policy call 0; others emerge mid-rollout — two distinct failure modes.

Features / baseline	AUROC	AP	Bal. Acc.
Attention-derived	0.827	0.510	0.754
Action-direction probe	**0.917**	0.725	**0.919**
Gripper-state probe	0.713	0.244	0.675
All features	0.882	**0.726**	0.853
Success-centroid baseline	0.847	0.376	0.789
Random-label probe baseline	0.477	0.047	0.480



- Attention AUROC 0.827 — complementary failure signal. Failures show lower visual attention entropy, suggesting overcommitment to wrong visual region, not diffuse uncertainty.
- Fig. 3 shows diverging risk scores: failed episodes (orange) climb above the threshold while successes stay flat. Some failures trigger immediately at policy call 0; others emerge mid-rollout.

5 Limitations and Discussion

- Only 22 failures out of 500 LIBERO-Long episodes; results should be validated on larger, more diverse failure sets across tasks.
- Attention results are nuanced: low entropy in failures may indicate overcommitment to wrong visual regions, not diffuse uncertainty.
- Gripper probe AUROC 0.713 (weakest). Gripper state is linearly decodable (macro-F1 ~0.98) across all layers but less discriminative for failure in long-horizon tasks.

PAPER

Mechanistic Interpretability Workshop, ICML 2026

tyang08@engineering.upenn.edu