Interpretable Signals for Vision-Language-Action Model Failure

Tony Yang [1] **Aidan Mokalla** [2]

Abstract

Vision-language-action (VLA) models take in visual observations and language instructions to produce physical robot actions, but when they fail at this task, this poses safety concerns for deployment. We study whether interpretable internal signals from a VLA $\pi_{0.5}$ can predict if the model will fail on LIBERO tasks. We cache action-expert activations and attention weights during evaluation, and then derive two families of episode-level features: attention-derived summaries over visual tokens and activation-probe summaries for action direction and gripper state. Our analysis suggests that action-direction probe confidence is the most robust failure-relevant signal, while attention-derived features are also predictive and gripper-state probes are weaker. We further compare against success-centroid distance and random-label probe baselines, and find that meaningful action-direction labels are important for the probe signal. These results work to show that simple mechanistic features inside VLAs can reveal failure-relevant representations.

1. Introduction

Vision-language-action models take in both visual perception and language information in order to generate robotic actions to complete tasks (Kim et al., 2024; Black et al., 2025; Cadene et al., 2026). As these models are physically deployed, we believe that failure prediction is highly important. What makes these models unique is that action token outputs of VLAs are directly transferred into the real-world when deployed, introducing direct safety concerns. A poorly calculated action has direct effects on its environment, and can cause unchecked harm, and so this threat vector must be preemptively evaluated. Thus, the ability to foresee VLA failure *prior* to physical task execution is crucial.

[1]University of Pennsylvania [2]Independent Researcher. Correspondence to: Tony Yang <tyang08@engineering.upenn.edu>, Aidan Mokalla <AidanMokalla@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

This work uses mechanistic interpretability methods to analyze whether internal representations and patterns may offer insights into the model's internal state before or during task failure. Recent VLA interpretability work suggests that action concepts become increasingly represented in mid-to-late layers (Grant et al., 2026; Swann et al., 2026), while related vision language model (VLM) interpretability work shows that probes on hidden states can reveal multimodal grounding failures (Ashok et al., 2025). This motivates a simple question: Do interpretable signals from a VLA's internal architecture that temporally preceede the final output correlate with, and possibly predict, rollout success?

Our work investigates this question in $\pi_{0.5}$ on LIBERO-Long. We focus on the Gemma action expert in $\pi_{0.5}$ and construct two feature families: *attention features*, which summarize visual-token attention entropy and mass, and *activation-probe features*, which summarize the confidence of success-trained probes for action direction and gripper state. We propose a failure prediction pipeline that extracts interpretability signals, which do not account for superposition, from VLA internal representations during inference.

Our contributions are framed as three empirical claims instead of a general theory of VLA failure. First, action-direction labels are linearly decodable from $\pi_{0.5}$ action-expert activations, giving a linkage from hidden states to the robot action interface. Second, confidence and entropy from success-trained action probes are associated with rollout failures. Third, attention features over visual tokens provide a complementary failure signal, although the observed direction is more nuanced than our original hypothesis.

This paper presents:

- evidence that action-direction information is linearly decodable from mid-to-late action-expert layers;
- preliminary failure-prediction results showing that probe confidence is strongly associated with failures;
- a chunk-level risk-trajectory analysis showing that these features can be evaluated throughout a rollout;
- an unexpected attention result: task failures exhibit lower visual-attention entropy than successes.

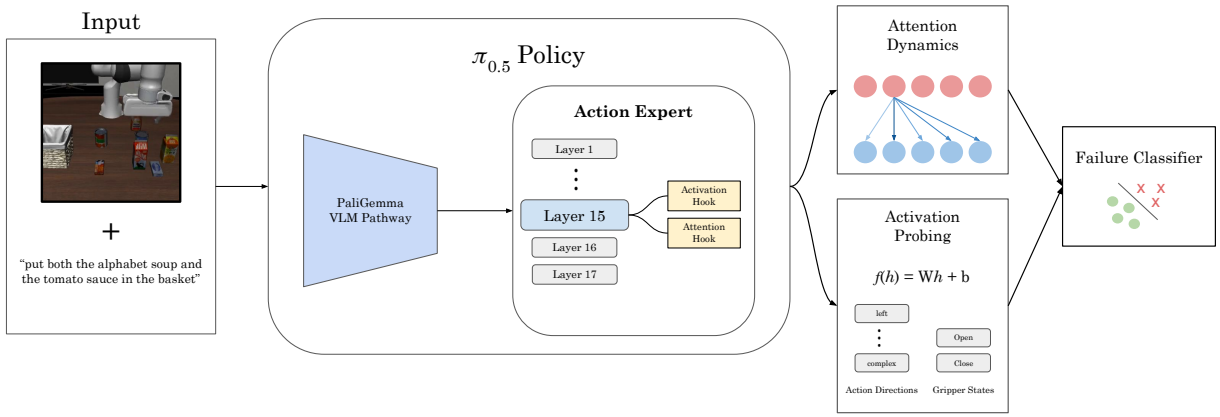

Figure 1. Pipeline overview. During evaluation, we cache internal attention and activation features from $\pi_{0.5}$, convert them into interpretable scalar signals, and train a lightweight episode-level classifier to predict success or failure.

2. Related Work

VLAs such as OpenVLA and $\pi_{0.5}$ map images and language instructions to robot actions (Kim et al., 2024; Black et al., 2025). The LeRobot framework provides a pipeline for running and evaluating such models (Cadene et al., 2026) using LIBERO, which is a suite of manipulation benchmarks, including the long-horizon tasks used in this work (Liu et al., 2023).

Previous works have analyzed internal VLA representations. For example, and (Grant et al., 2026) provides a map of the "action circuits" within VLAs (Häon et al., 2025) found FFN activations reveal sparse semantic directions causally linked to action selection. Although not aiming to predict failures specifically, (Swann et al., 2026) applied sparse autoencoders (SAEs) to VLA hidden layers, discovering that action-relevant features concentrate in mid-to-late layers. These works motivate analyzing action-expert layers instead of treating the policy as a black box when trying to determine if task failure can be predicted.

SAFE studies scalable failure estimation for VLAs using final latent features and identifies failure regions in the embedding space to do so (Gu et al., 2025). SAFE focuses on simplicity when only considering the feature vectors of the last layer. However, the authors urge future work to extend beyond last-layer features. Our work differs by constructing interpretable scalar signals from attention and action-related probes to analyze their relationship with failures.

A separate line of work pursues safety through formally verified exploration during training rather than post-hoc diagnosis. For example, REVEL (Anderson et al., 2020) enforces *provably* safe exploration in an RL setting in continuous control by alternating between a symbolic policy class that is efficiently verifiable and a neurosymbolic class with approximate gradients, avoiding the need to repeatedly verify a full neural network within the learning loop. This targets safety during the exploration process itself, and requires a specific policy representation and training procedure to admit verification. Our work instead treats the VLA as a fixed, already-trained model: we simply ask whether lightweight probes on its existing internal activations can predict failure at deployment time, without altering the policy class or training pipeline.

3. Method

We collect attention weights and activation features within $\pi_{0.5}$ when trained and evaluated on the LIBERO benchmark. Next, these values are used to construct attention entropy and linear probe confidence on activation features mapped to both VLA action directions and gripper states. Statistical analysis and experiments will examine the correlation between these signals and task failure. The model $\pi_{0.5}$ uses a dual-pathway architecture with a PaliGemma vision-language backbone and an 18-layer Gemma action expert. We hook action-expert layers denoted `expert_L0` through `expert_L17`. Unless otherwise stated, our main experiments use `expert_L15`.

3.1. Attention Features

For action-expert layer ℓ, we collect self-attention weights during denoising. Let $A^{(\ell)} \in \mathbb{R}^{h \times n \times n}$ denote attention over heads, query tokens, and key tokens. We focus on action/suffix query tokens and visual prefix key tokens. For the action token (position n), we compute its attention to visual tokens (positions $1 \ldots n_v$):

$$p = \text{softmax}\left(\frac{1}{h}\sum_{i=1}^{h} \frac{\mathbf{Q}_i^{(n)} \mathbf{K}_i^{(1 \ldots n_v)\top}}{\sqrt{d_k}}\right)$$

where $\mathbf{Q}_i, \mathbf{K}_i$ are query and key matrices for head i. For each head and action token, we normalize attention mass over visual tokens and compute normalized entropy:

$$H_{\text{attn}} = \frac{-\sum_{j=1}^{n_v} p_j \log p_j}{\log n_v}. \tag{1}$$

We also record visual attention mass and temporal summaries across denoising steps. Our initial hypothesis was that failures would have higher entropy, suggesting that the model is uncertain about which visual features are relevant for the task. However, our results suggest the opposite.

3.2. Activation-Probe Features

For our second feature family, we use linear probing across action-expert activations. We do not use probes to control the robot but rather to analyze whether intermediate activations resemble representations observed during successful task runs.

We derive direction labels from the executed translational action components $[dx, dy, dz]$, assigning each step to one of $k = 8$ bins: left, right, forward, back, up, down, neutral, or complex. For a hidden activation $h_t^{(\ell,k)}$ at layer ℓ and action-token position k, we train a linear probe

$$f(h) = Wh + b \tag{2}$$

to predict the direction label. For the failure signal, probes are trained only on successful episodes. During scoring, we summarize the softmax confidence and entropy over an episode:

$$c_{\text{dir}} = \max_i \text{softmax}(f(h))_i. \tag{3}$$

Low confidence or high probe entropy indicates that the representation does not match the action-direction structure learned from successful executions.

The idea for the gripper probe is similar to the action-direction probe. However, we separately consider the gripper state due to its binary nature, thus $k = 2$ instead of 8 and we train a binary probe (open/closed).

3.3. Episode-Level Failure Classifier

For each episode, we construct a scalar feature vector

$$s = [H_{\text{attn}}, c_{\text{dir}}, H_{\text{dir}}, c_{\text{grip}}, H_{\text{grip}}], \tag{4}$$

where H_{dir} and H_{grip} are probe-output entropies. We train logistic regression classifiers with balanced class weighting and evaluate with AUROC, average precision, and balanced accuracy. Since failures are rare relative to successes, classifier results should be viewed as preliminary, and future work should validate on larger and more diverse failure sets.

3.4. Chunk-Level Risk Trajectories

Failure prediction on the episode level is useful for analyzing failure after it has already occurred, but ideally estimates should be updated and tracked while the episode is still running. This is important since VLAs do not simply produce a single label after seeing a single image, but instead repeatedly refine their continuous inputs and previous state to produce continuous action chunks. Thus, we construct a risk trajectory from the same attention and activation-probe features. Since $\pi_{0.5}$ predicts actions in chunks, we use each policy call as our timestep unit instead of a simulator step. For each policy call u, we compute the same feature family on cached attention and activation tensors and feed it to the trained logistic classifier to obtain a risk score $r_u \in [0, 1]$. We report final chunk-level AUROC and the fraction of failed episodes whose risk crosses a high-risk threshold before the end of the episode. The risk trajectory is shown in Figure 3.

4. Experimental Setup

We evaluate $\pi_{0.5}$ fine-tuned on LIBERO using LeRobot. Our main setting is LIBERO-Long (`libero_10`), with 10 tasks and 50 episodes per task. Each rollout uses a maximum of 520 steps and $\pi_{0.5}$ action chunking with 10 action steps.

We cache per-token action-expert activations with shape $[\text{steps}, \text{batch}, \text{action tokens}, \text{hidden}]$. The main probe layer/token is `expert_L15 token_5`, selected from a layer/token sweep. Attention collection stores entropy and visual attention mass only, not raw visual attention maps, to reduce storage use.

We report three classes of results. First, probe decodability measures whether action direction is linearly encoded in the representation. Second, failure-prediction experiments train success-only probes and use attention/probe summaries as episode-level features. Third, the risk-trajectory experiment applies the classifier throughout each rollout at the policy-call level. Our final classifier uses a unified same-rollout collection so attention and activation features share the same success/failure labels.

5. Results

In a separate 100-episode activation sweep, `expert_L15 token_5` achieves validation accuracy 0.707 and macro-F1 0.675 for action-direction prediction. The sweep in Figure 2 shows that L2 is substantially weaker than L13–L17, supporting the previous findings mentioned earlier of mid-to-late action-expert layers for activation-probe features. A full action-expert layer sweep in Figure 10 further shows that gripper state is highly decodable across layers, suggesting that its weaker failure-prediction result is not caused by

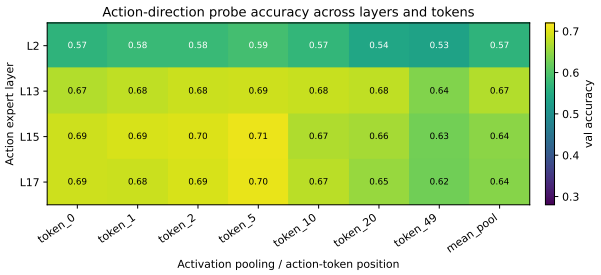

Figure 2. Action-direction decodability across action-expert layers and token positions. Mid-to-late layers are more predictive than early layer L2, and `expert_L15 token_5` is the strongest setting used in our failure analyses.

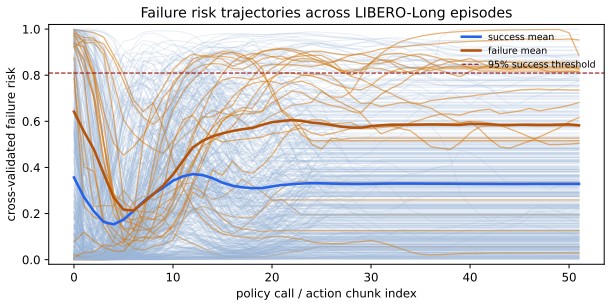

Figure 3. Chunk-level failure risk over LIBERO-Long rollouts. Because $\pi_{0.5}$ uses action chunking, each point corresponds to a policy call rather than a simulator timestep.

absent gripper information.

In the 500-episode unified run, we observe that the strongest signal is the action-direction probe family, obtaining an AUROC of 0.917, an average precision of 0.725, and a balanced accuracy of 0.919. The results show that attention features, which obtained AUROC 0.827, and action-direction probe features are both predictive, while gripper-probe features are weak on this run. The all-feature classifier remains predictive but interestingly does not outperform action-direction features alone, suggesting that action-direction representation is a dominant failure-relevant signal in this particular setting. Additionally, mean action-direction confidence is lower in failures than successes (0.646 vs. 0.755), low-confidence fraction is higher in failures (0.332 vs. 0.166), and probe entropy is higher in failures (0.486 vs. 0.346). These results suggest that failure episodes are internally less consistent with the action-direction representations learned from successful executions.

For baselines in Table 1, success-centroid distance is above chance but weaker than the action-direction probe, while the random-label probe is near chance, as expected. Thus, meaningful action-direction labels seem to matter for the confidence signal.

Table 1. Episode-level failure prediction on 500 LIBERO-Long episodes. Action-direction probe features are the dominant failure-relevant signal in this setting.

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

Monitoring internal features throughout each episode, risk-trajectory results can be seen in Figure 3. We use policy-call-level scores and find that final risk reaches AUROC 0.744 with an average precision 0.155 on the unified run. At a threshold set to the 95th percentile of successful-episode risk, 17 of 22 failed episodes cross the threshold. Several failures cross immediately at the first policy call, while others cross later. This suggests a noticeable difference in the types of VLA failures, with some episodes starting in an unexpected or atypical state and others accumulating risk only after a sequence of actions.

6. Discussion

We find strong evidence that action-direction probe features are predictive of the observed failures. The results from activation-probing support the idea that successful runs show a more regular action-representation, and that low confidence hints that the model has not clearly formed action directions in its internal representations, signaling for task failure.

In our work, similarly to other failure detection work, a notable caveat is the scarcity of failed VLA episodes. Like numerous other VLA models, $\pi_{0.5}$ achieves high success rates on LIBERO. We initially obtained 7 failures in 100-episode LIBERO-Long runs, and addressed this by scaling our run to 22 failures out of 500 episodes.

On the other hand, the attention results are a bit more nuanced. Instead of supporting our original hypothesis that high-entropy indicates model uncertainty and thus likely failure, the results suggest that attention entropy may depend on the *failure mode*. It is possible that low entropy could indicate overcommitment to the wrong visual region, while high entropy could indicate diffuse uncertainty in other settings. It is also possible that evaluating attention entropy without regard for signal superposition may weaken the signal, and future work should consider learning sparse decompositions of such signals in VLA models.

We show that $\pi_{0.5}$ contains linearly decodable action-direction information, and that success-trained readouts of

this information are less confident on failed rollouts. We do not claim that the probes identify the causal circuit responsible for failure, nor that the resulting classifier is ready for deployment without adaptation to the specific task, model, environment, and use case.

7. Conclusion

Our work presents a preliminary pipeline to predict VLA failures using interpretability signals for $\pi_{0.5}$ on LIBERO. On a 500-episode evaluation, action-direction activation probes are the strongest failure-relevant signal, attention features provide complementary evidence, and gripper-state probes show weaker predictive evidence. Our results support the broader hypothesis that VLAs contain internal representations correlated with task failure. Future work should extend analysis to multiple VLA architectures and simulator environments, should investigate SAE-derived features,and should test whether interventions on action-direction or attention features can causally affect failure behavior.

Acknowledgements

Computational resources used in this work were supported by the Oregon Regional Computing Accelerator (Orca), funded by NSF CC*-2346732.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Anderson, G., Verma, A., Dillig, I., and Chaudhuri, S. Neurosymbolic Reinforcement Learning with Formally Verified Exploration. In *Advances in Neural Information Processing Systems*, 2020.

Ashok, D., Chaubey, A., Arai, H. J., May, J., and Thomason, J. Can VLMs Recall Factual Associations From Visual References? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15691–15708, 2025.

Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Galliker, M., et al. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In *Proceedings of the 9th Conference on Robot Learning*, PMLR 305:17–40, 2025.

Cadène, R., Aliberts, S., Capuano, F., Aractingi, M., Zouitine, A., Kooijmans, P., Choghari, J., Russi, M., Pascal, C., Palma, S., Shukor, M., Moss, J., Soare, A., Aubakirova, D., Lhoest, Q., Gallouédec, Q., and Wolf, T. LeRobot: An Open-Source Library for End-to-End Robot Learning. *arXiv preprint arXiv:2602.22818*, 2026.

Grant, B., Zhao, X., and Wang, P. Not All Features Are Created Equal: A Mechanistic Study of Vision-Language-Action Models. *arXiv preprint arXiv:2603.19233*, 2026.

Gu, Q., Ju, Y., Sun, S., Gilitschenski, I., Nishimura, H., Itkina, M., and Shkurti, F. SAFE: Multitask Failure Detection for Vision-Language-Action Models. In *Advances in Neural Information Processing Systems*, 2025.

Häon, B., Stocking, K., Chuang, I., and Tomlin, C. Mechanistic interpretability for steering vision-language-action models. *arXiv preprint arXiv:2509.00328*, 2025.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., and Finn, C. OpenVLA: An Open-Source Vision-Language-Action Model. In *Proceedings of the 8th Conference on Robot Learning*, PMLR 270:2679–2713, 2025.

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Advances in Neural Information Processing Systems*, 2023.

Swann, A., McGranahan, L., Buurmeijer, H., Kennedy III, M., and Schwager, M. Sparse Autoencoders Reveal Interpretable and Steerable Features in VLA Models. *arXiv preprint arXiv:2603.19183*, 2026.

A. Experimental and Label Details

Main failure-prediction experiments use $\pi_{0.5}$ fine-tuned on LIBERO-Long (`libero_10`) with 500 same-rollout episodes: 478 successes and 22 failures. Each episode has a maximum horizon of 520 simulator steps. Since $\pi_{0.5}$ predicts action chunks, cached activations and attention-derived summaries are aligned to policy calls rather than individual simulator steps.

For activation probes, we cache per-token hidden states from the Gemma action expert and use `expert_L15 token_5` for the main failure analysis. Attention-derived features summarize attention weights from action/suffix query tokens to visual prefix key tokens. The main run stores visual-attention entropy and visual attention mass, not raw spatial attention maps.

LIBERO actions use the convention

$$[dx, dy, dz, droll, dpitch, dyaw, gripper].$$

Direction labels come from $[dx, dy, dz]$ in the robot/action frame, giving eight classes: left, right, forward, back, up, down, neutral, and complex. Gripper labels use the LIBERO/robosuite Panda convention: final action > 0 is closed and final action ≤ 0 is open. These labels describe the policy's executed actions, not external task correctness.

B. Episode-Level Classifier and Baselines

The episode-level classifier is logistic regression over scalar summaries of attention-derived features and activation-probe outputs. Attention-derived features include visual-token attention entropy, late-denoising entropy, late-minus-early entropy, visual attention mass, and late visual attention mass. Activation-probe features include mean confidence, low-confidence fraction, and normalized entropy for the action-direction and gripper-state probes. We use balanced class weighting because failures are rare.

The baselines in Table 1 are computed from the same cached `expert_L15 token_5` activations. The success-centroid baseline represents each episode by its mean activation vector and scores distance to a centroid fit on successful episodes. The random-label probe baseline preserves the probe architecture and confidence summarization pipeline, but shuffles action-direction labels during probe training.

C. Temporal Attention Analysis

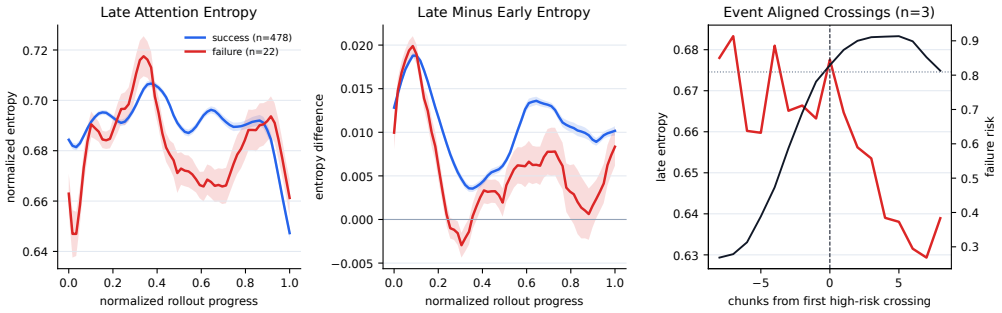

Figure 4. Temporal analysis of visual-attention entropy along risk trajectories. Failed rollouts show slightly lower late-denoising attention entropy and weaker late-minus-early entropy growth than successful rollouts. Event-aligned analysis around first high-risk crossings is shown for the few failures with non-initial crossings, suggesting that the low-entropy signal is more often early or persistent than a sharp collapse at a single timestep.

We further analyzed whether the low-entropy attention pattern emerges near high-risk portions of failed rollouts. Using chunk-level risk trajectories, we compared late-denoising attention entropy over normalized rollout progress. Failed episodes exhibited slightly lower mean attention entropy than successes, and a smaller late-minus-early entropy increase during denoising. Within failed episodes, chunk-level risk was negatively correlated with late attention entropy ($r = -0.231$) and with the late-minus-early entropy change ($r = -0.368$). However, most failed episodes that crossed the high-risk threshold did so at the first policy call, leaving only three non-initial crossings for event-aligned analysis. Thus, the current evidence supports an early or persistent low-entropy failure signature, rather than a universal abrupt entropy collapse at a later failure point.

D. Risk Trajectory Details

The risk trajectory experiment applies the episode-level classifier to cumulative policy-call features. At policy call u, the classifier receives summaries computed up to that call and outputs a failure-risk score. This gives an online-compatible diagnostic at the action-chunk level. On the 500-episode unified run, final risk has AUROC 0.744 and AP 0.155. Using the 95th percentile of successful-episode risk as a threshold, 17 of 22 failed episodes cross the threshold at some point during the rollout.

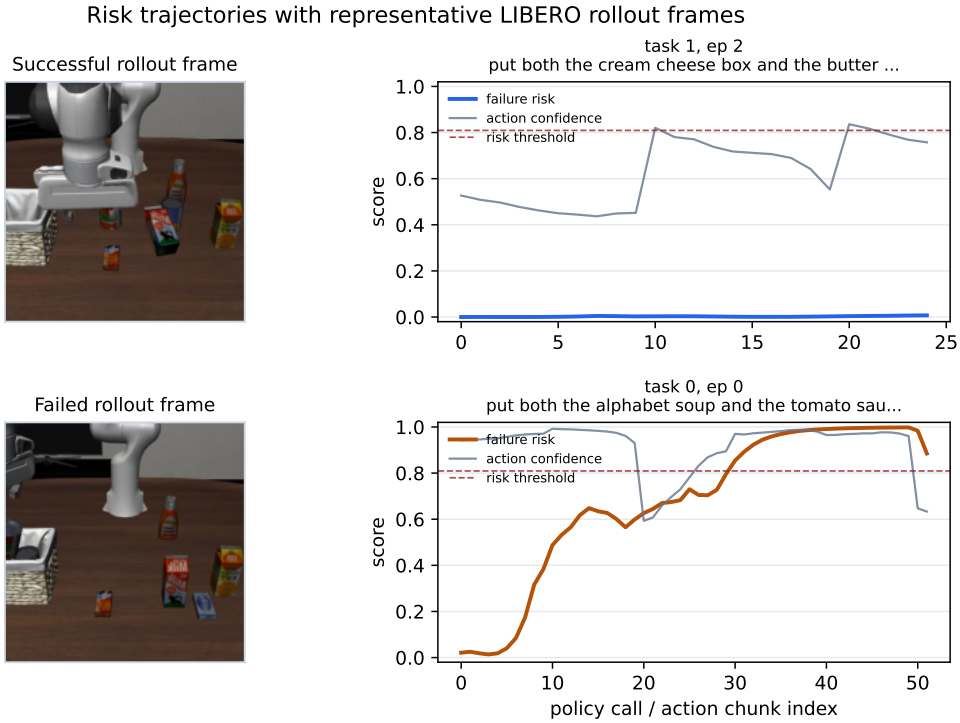

Figure 5. Representative risk trajectories with LIBERO rollout frames. The frames are illustrative examples from saved evaluation videos, while the risk curves are computed from the unified signal run. The failed example shows risk rising over policy calls, whereas the successful example remains low.

This analysis should be interpreted as a diagnostic rather than a causal intervention. In particular, the current saved attention summaries can show when attention-derived entropy or mass changes, but they cannot by themselves identify which object or image patch receives attention. Spatial attention-map validation requires a separate qualitative run that stores raw visual attention maps.

E. Action-Direction Probe Visualizations

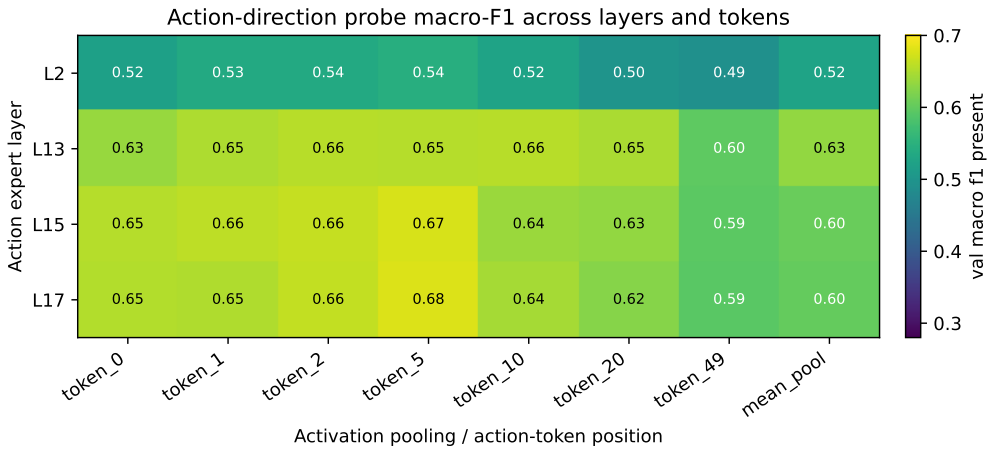

Figure 6. Layer/token macro-F1 for action-direction probes in the diagnostic activation sweep. Mid-to-late action-expert layers are more linearly predictive of executed translational direction than the early action layer, supporting the use of action-expert activations for failure analysis.

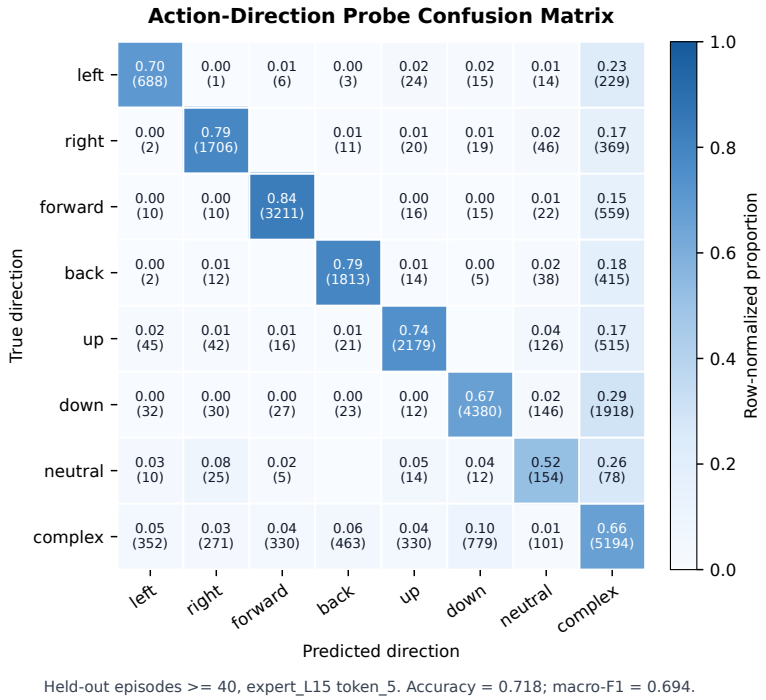

Figure 7. Row-normalized confusion matrix for the `expert_L15 token_5` action-direction probe. Confusions should be interpreted in the robot/action frame, not the camera frame; visually similar camera motions can correspond to different robot-frame directions.

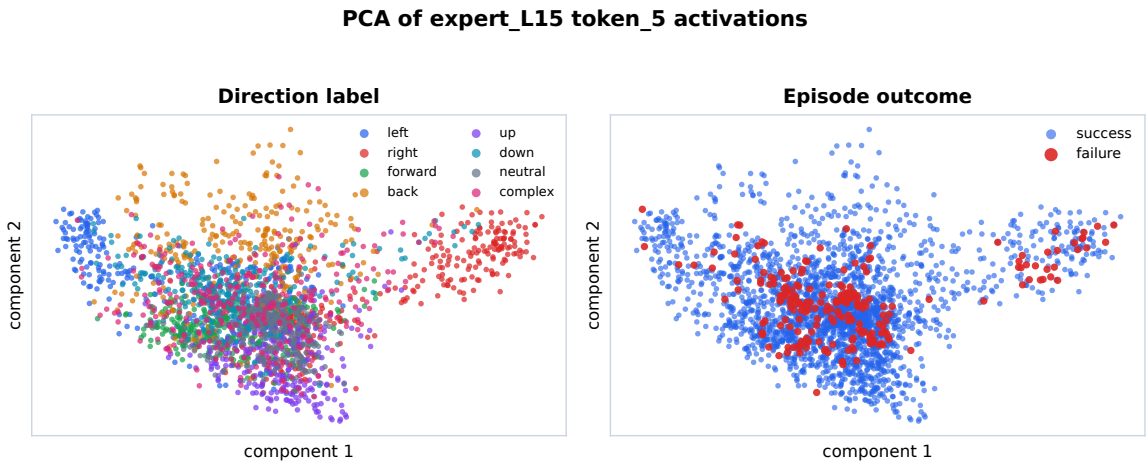

Figure 8. PCA projection of sampled `expert_L15 token_5` activations. The plot is intended as a qualitative view of action-direction structure; the linear-probe metrics provide the quantitative evidence.

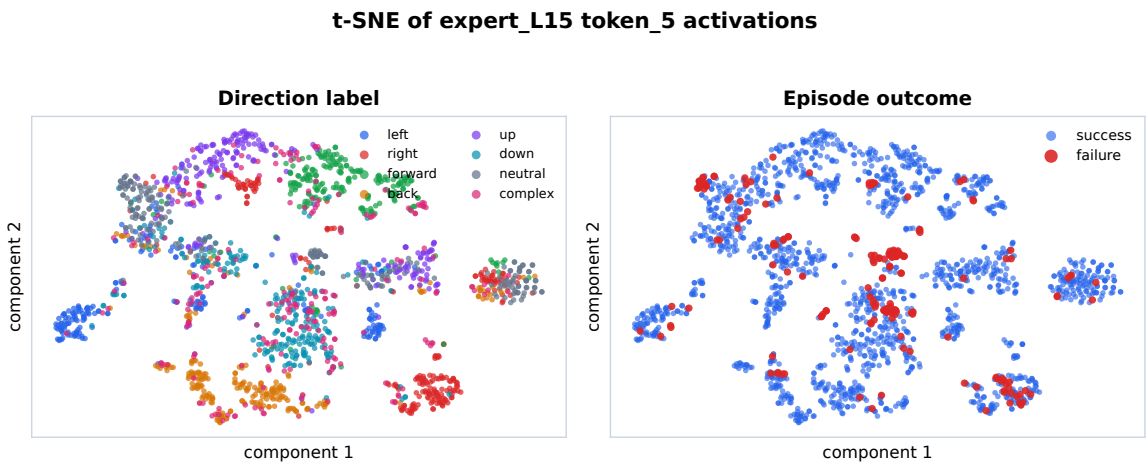

Figure 9. t-SNE projection of sampled `expert_L15 token_5` activations colored by action direction and episode outcome. This visualization is non-metric and should not be read as a classifier result. It is included to show that action labels induce visible structure in the activation space.

F. Full Action-Expert Layer Sweep

To test whether the gripper-state probe is weak because of the chosen layer, we run a sweep over all 18 $\pi_{0.5}$ action-expert layers. For each layer, we collect a 100-episode LIBERO-Long activation run, train action-direction and gripper-state probes, save compact summaries, and delete raw activation tensors before moving to the next layer. The completed sweep uses token-0 activations for each layer, so it should be interpreted as a layer-wise diagnostic rather than a full layer-by-token search.

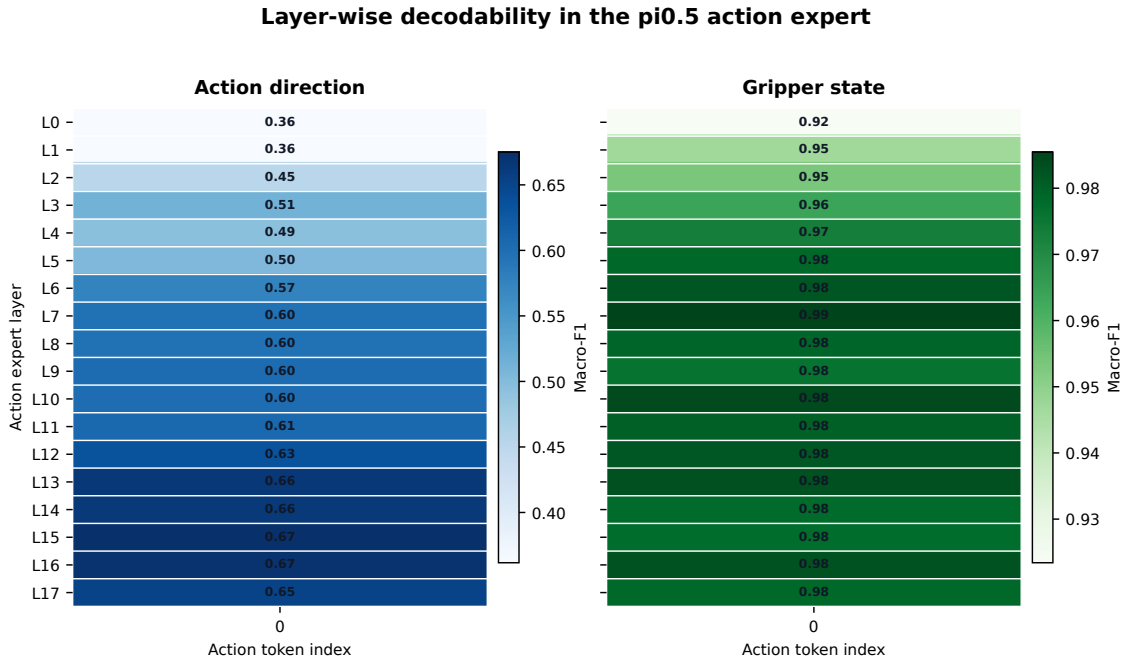

Figure 10. Layer-wise linear decodability of action direction and gripper state from token-0 action-expert activations. Action-direction macro-F1 increases through the action expert and peaks in late layers, while gripper state is already highly decodable in early layers and remains strong throughout.

The sweep clarifies the gripper result in the main failure classifier. Action-direction decodability grows from early to late action-expert layers, with the best macro-F1 near `expert_L15`. Gripper state, however, is linearly decodable across the full action expert, reaching high macro-F1 even in early layers. Thus, the weaker gripper-state failure signal in Table 1 is unlikely to mean that $\pi_{0.5}$ lacks gripper-state information. A more plausible interpretation is that gripper state is encoded, but is less discriminative of rollout failure than translational action-direction structure in these LIBERO-Long episodes.