BEYOND CONFIDENCE: LOG-ALIGNMENT PROFILES FOR MONITORING INTERNAL COMPUTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution shift can alter a network's internal computation while leaving its predictions confident. Our log-alignment profiles record layerwise amplification and activation scale for each input, using two scalar measurements per probe during an ordinary forward pass. Calibrated against reference data, they detect departures from training data and locate atypical layer responses. The log-alignment profiles detect shifts that output scores detect poorly. On CIFAR-10 and SVHN datasets, we show that even when output confidence fails to separate samples from the datasets, our log-alignment profile achieves 0.880 detection AUROC. Separately, we train a predictor of incorrect ResNet-50 predictions using output entropy and the log-alignment profile score. Fitted on labeled clean and shifted images, it raises held-out error-prediction AUROC from 0.632 with entropy alone to 0.811. In language models from 124M to 7B parameters, the log-alignment profile detects Python code relative to a natural-text reference at 0.95–0.99 AUROC. On LIBERO behavior-cloning policies, log-alignment profile alarms detect corruption onset 16–35 steps before action entropy; on ManiSkill, log-alignment profiles add episode-failure information beyond ensemble disagreement. In grokking transformers, late-layer alignment interventions during training reduce mean post-fork grokking delay by 33% in a three-seed comparison and accelerate grokking in all eight runs of a separate paired experiment.

1 INTRODUCTION

Distribution shift can change a network's internal activations while leaving its predictions confident. Output scores such as maximum softmax probability (Hendrycks & Gimpel, 2017), energy (Liu et al., 2020) and ensemble disagreement (Lakshminarayanan et al., 2017) summarize the prediction. We instead study an internal monitor that identifies which layers respond atypically and whether that response precedes a change in output uncertainty.

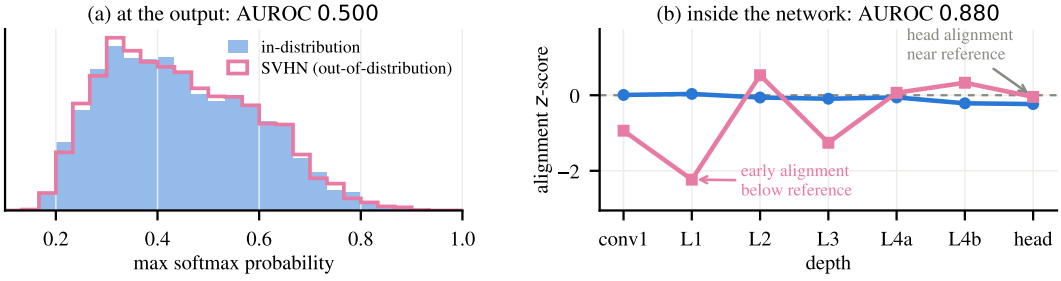


Figure 1: **Similar output confidence can hide different internal responses.** We match the maximum-softmax confidence histograms of CIFAR-10 reference and SVHN digit images to ask whether internal measurements can distinguish samples from the two datasets, despite similar output confidence. (a) Maximum softmax probability (MSP) separates the sets at chance (AUROC 0.500). (b) The log-alignment profile separates them at AUROC 0.880. Curves show mean alignment in reference standard deviations; zero is the reference mean. SVHN departs most strongly in the first block but returns near the reference at the classifier head. See Appendix G.7 for matching and baseline comparisons.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The log-alignment ratio (LAR) was introduced by Everett et al. (2024) and studied as a training-time diagnostic by Shehper & Vaswani (2026). We turn it into an input-level monitor by recording alignment and activation scale across depth and calibrating the resulting log-alignment profile against reference data. This results in a scalar score U_{LAP} that flags an atypical input; the layerwise log-alignment profile shows where its processing departs from the reference (Section 4.3). This provides both a monitoring signal and an interpretable description of the internal response.

Alignment measures how strongly each layer amplifies its input along learned directions, while activation scale records the input's magnitude. We apply these measurements to image classifiers, language models and closed-loop policies, with task-specific aggregation and calibration.

Our contributions are: **(1) A compact log-alignment profile of internal computation.** We construct a per-input log-alignment profile of layerwise alignment and activation scale, calibrated on in-distribution data. Each probe uses norms of activations already available in the forward pass. The calibrated log-alignment profile provides both a departure score and a depth-resolved description of the response (Section 4). **(2) Evidence that the log-alignment profile complements output uncertainty.** Across image classifiers, language models and visuomotor policies, we identify shifts that output scores detect poorly. Replacing learned input directions (patterns amplified differently by a layer's weights) with random ones in the monitor reduces detection AUROC by up to 0.41 without changing predictions. We compare with other detectors of unusual hidden activations and test amplification and activation scale separately to identify their contributions (Table 1 and Appendix G.3). Adding a log-alignment profile to entropy raises supervised ResNet-50 error-prediction AUROC from 0.632 to 0.811 on clean and shifted images (Section 5.1). **(3) Alignment interventions change generalization timing.** In grokking transformers, encouraging alignment at late layers reduces mean grokking delay by 33% (Table 4).

2 Related Work

The closest constructions are depth-resolved activation statistics calibrated on in-distribution (ID) data: layerwise feature Mahalanobis (Lee et al., 2018) and Gram-matrix deviations (Sastry & Oore, 2020). Our log-alignment profile represents each layer with amplification and activation scale, yielding two coordinates per probe. We compare against both methods under identical calibration (Table 1) and use a random-direction control to test whether learned weight geometry contributes to detection. Other feature-space methods estimate typicality or reshape activations (Sun et al., 2022; Wang et al., 2022; Sun et al., 2021; Djurisic et al., 2023; Ren et al., 2021).

Predictive uncertainty is studied through Bayesian and ensemble methods (Neal, 1996; Gal & Ghahramani, 2016; Blundell et al., 2015; Lakshminarayanan et al., 2017; Papamarkou et al., 2022), aleatoric/epistemic decompositions (Kendall & Gal, 2017), evidential heads (Sensoy et al., 2018), calibration (Guo et al., 2017) and conformal coverage (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023; Barber et al., 2023). We test whether internal log-alignment profiles add information to output-distribution scores (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020).

Language and multimodal uncertainty methods use likelihood (Chou et al., 2025; Bai et al., 2024), input embeddings (Tran et al., 2022; Upadhyay et al., 2023; Venkataramanan et al., 2025), internal states (Chen et al., 2024), and generation consistency (Farquhar et al., 2024; Nikitin et al., 2024; Manakul et al., 2023; Lin et al., 2024; Lau et al., 2025; Zhang et al., 2024b; Quach et al., 2024; Su et al., 2024; Zhou et al., 2024). Our log-alignment profile measures layerwise amplification and scale in one forward pass (Section 5.2). Our comparison of layer responses against reference data relates to the familiarity hypothesis (Dietterich & Guyer, 2022) to the covariate-shift axis of generalized OOD detection (Yang et al., 2024; Zhang et al., 2024a; Yang et al., 2022). For grokking, prior work studies delayed generalization (Power et al., 2022), weight-norm timescales (Truong et al., 2026), circuit formation (Nanda et al., 2023; Sivasankar, 2026) and LAR as a training diagnostic (Shehper & Vaswani, 2026). We test how alignment interventions change generalization under norm control.

3 Background: Log-Alignment Ratio

We use the LAR identity (Everett et al., 2024; Shehper & Vaswani, 2026) to interpret and compute the layerwise measurements.

A linear layer amplifies some input directions more strongly than others. Alignment measures how much activation energy falls along those strongly amplified directions. The spectrum p in (1) describes the layer's amplification preferences; occupancy q in (2) describes the input's energy distribution.

Definition 3.1 (Weight spectrum and activation occupancy)**.** Let $W \in \mathbb{R}^{m\times d}$ be a learned matrix of weights with SVD $W = U\Sigma V^\top$, singular values $\sigma_1 \geq \cdots \geq 0$ and right-singular vectors v_i (the columns of V), and call

$$p_i = \frac{\sigma_i^2}{\|W\|_F^2}, \qquad \sum_i p_i = 1, \tag{1}$$

its *weight spectrum*: since $\|Wv_i\|_2^2 = \sigma_i^2$ and $\|W\|_F^2 = \sum_j \sigma_j^2$, p_i is the fraction of the matrix's total energy carried by right-singular direction v_i (zero when W is rank deficient).

The occupancy of a nonzero activation $h \in \mathbb{R}^d$ is

$$q_i(h) = \frac{(v_i^\top h)^2}{\|h\|_2^2}, \qquad \sum_i q_i(h) = 1, \tag{2}$$

the energy fraction of h in direction v_i. For a batch of B activations stacked as columns, $H = [h_1, \ldots, h_B] \in \mathbb{R}^{d\times B}$, $q_i(H) = \|v_i^\top H\|_2^2/\|H\|_F^2$ is the $\|h_b\|_2^2$-weighted average of the $q_i(h_b)$.

The batch-level log-alignment ratio can be written as the log overlap of the two spectra (Shehper & Vaswani, 2026),

$$\mathrm{LAR}(W, H) = 1 + \tfrac{1}{2}\log_d \textstyle\sum_i p_i\, q_i(H), \tag{3}$$

which is large when activation concentrates where the matrix amplifies. Shehper & Vaswani (2026) studied its relation to generalization and overfitting as a training-time diagnostic. Evaluating the same identity at a single activation gives the measurement discussed below.

Definition 3.2 (Alignment)**.** The *alignment* of a nonzero activation $h \in \mathbb{R}^d$ with W is the overlap of the two spectra, $S(W, h) = p^\top q(h)$, or on the log scale

$$L(W, h) = 1 + \tfrac{1}{2}\log_d S(W, h). \tag{4}$$

We write $L_\ell(x)$ for the alignment at depth ℓ on input x; with the batch occupancy $q(H)$ in place of $q(h)$, this recovers Equation (3).

Theorem 3.3 (Occupancy identity)**.** *For every nonzero h,*

$$S(W, h) = \frac{\|Wh\|_2^2}{\|W\|_F^2\,\|h\|_2^2}, \quad L(W, h) = 1 + \log_d \frac{\|Wh\|_2}{\|W\|_F\|h\|_2}. \tag{5}$$

With the weights W fixed, the layer amplifies each direction by a fixed amount. Alignment can differ between inputs because their activations distribute different fractions of energy across these directions, as recorded by $q(h)$. Changing only an activation's magnitude leaves its alignment unchanged.

By Theorem 3.3, alignment can be computed directly from the input, output, and weight norms, without an SVD or explicitly computing $q(h)$. The method below retains activation magnitude as a separate measurement, combines it with alignment across depth, and calibrates their joint variation against reference data.

4 The Log-Alignment Profile: Measurement and Intervention

We turn the layerwise alignment measurement into a monitor for individual inputs. We first choose where to record activations, then measure amplification and activation magnitude at each probe, and finally compare the resulting log-alignment profile with reference data to obtain a departure score. Monitoring leaves the network's weights and predictions unchanged.

4.1 Where and how to probe

We attach probes to interior linear transformations spanning network depth and record their input and output norms during the forward pass. Here W is the probed weight matrix and h its incoming activation, so Wh is the output before adding bias. For a convolution, W contains the flattened

filters and h is a receptive-field patch; for a transformer, W is a feed-forward projection and h is the token representation entering it. We pool squared norms over the patches or tokens of each input (Appendices F and G.12).

We denote the number of probes by P; probe counts and placements are given in Appendix G. Interior alignment correlates weakly with maximum softmax probability (MSP), the probability assigned to the predicted class, whereas head alignment is more strongly coupled to it (Appendix G.1). This motivates testing whether interior probes add information beyond output confidence. In our tests, we show that during training, the interior alignment gap between hard and easy examples is negatively correlated with the generalization gap; Appendix G.1 reports the coefficients and experimental setup.

By Theorem 3.3, probes compute alignment from norms of forward-pass activations without an SVD. Convolutional probes also avoid constructing a patch matrix (Appendix F). Profiling adds less than 10% inference overhead in the measured configurations (Appendix J).

4.2 Log-alignment profile coordinates: gain and radius

Each probed layer yields two numbers: the log output norm and the log input norm,

$$a_\ell(x) = \log \|W_\ell h_\ell(x)\|_2, \qquad b_\ell(x) = \log \|h_\ell(x)\|_2, \tag{6}$$

We express these measurements as *gain* and *radius*:

$$g_\ell(x) = a_\ell(x) - b_\ell(x), \qquad r_\ell(x) = b_\ell(x). \tag{7}$$

Here g_ℓ measures log amplification and r_ℓ measures the log scale of the incoming activation. To relate gain to LAR, apply Equation (5) at layer ℓ, and use $\log_{d_\ell} u = \log u / \log d_\ell$:

$$L_\ell(x) = 1 + \frac{\log \|W_\ell h_\ell(x)\|_2 - \log \|h_\ell(x)\|_2 - \log \|W_\ell\|_F}{\log d_\ell} = 1 + \frac{g_\ell(x) - \log \|W_\ell\|_F}{\log d_\ell}. \tag{8}$$

For fixed weights, $\log \|W_\ell\|_F$ and $\log d_\ell$ are constants, so this is an invertible affine transformation. Thus g_ℓ and L_ℓ contain the same information about x at a fixed layer. Gain summarizes directional alignment, while radius records activation scale, which LAR removes through normalization (Proposition A.1). Activations can have the same amplification ratio but different magnitudes: gain captures the ratio, and radius preserves the scale. We retain both: $\Phi(x) = [\phi_1(x)^\top, \ldots, \phi_P(x)^\top]^\top$, where $\phi_\ell = (g_\ell, r_\ell)^\top$. Appendix G.3 compares the joint log-alignment profile with gain-only and radius-only variants.

4.3 The log-alignment profile score U_{LAP}

We use in-distribution (ID) data to obtain a calibrated alignment score. Let m_j^{ID} and s_j^{ID} be the mean and standard deviation of log-alignment profile coordinate Φ_j (one probe's gain or radius) on the ID calibration set. Each of the $2P$ scalar coordinates of Φ is standardized, $z_j(x) = (\Phi_j(x) - m_j^{\mathrm{ID}})/s_j^{\mathrm{ID}}$ for $j = 1, \ldots, 2P$, giving the calibrated log-alignment profile $Z(x) = (z_1(x), \ldots, z_{2P}(x))^\top$. The log-alignment profile score U_{LAP} (the subscript abbreviates log-alignment profile) is its Mahalanobis distance,

$$U_{\mathrm{LAP}}(x) = (Z(x) - \mu_{\hat{y}})^\top \Sigma_{\mathrm{pooled}}^{-1} (Z(x) - \mu_{\hat{y}}) \tag{9}$$

with predicted-class means $\mu_{\hat{y}}$ and one covariance Σ_{pooled} pooled over within-class residuals. The covariance records which layer responses normally vary together, so the score detects unusual combinations as well as unusually large individual deviations. Language models and control policies use one reference mean and covariance. Appendix E explains the choice of Mahalanobis distance and its calibration properties. Larger U_{LAP} indicates larger deviation from the calibration distribution. We evaluate shift detection using AUROC and assess its relationship to prediction errors separately (Section 5.1). Binary alarms use thresholds calibrated on reference scores, with the thresholding protocol specified for each experiment.

4.4 Intervening on occupancy

We test how alignment and occupancy affect learning by adding regularizers to the training objective (Section 5). Each regularizer is evaluated at a selected layer ℓ; Section 5.4 specifies the target

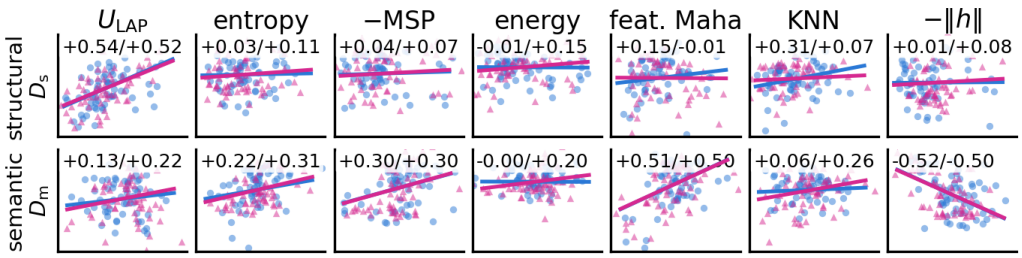


Figure 2: U_{LAP} **is the only score of seven that tracks structural deviation in both architectures.** The rows show associations with structural (top) and semantic (bottom) distance, each adjusted for the other. Blue circles denote ResNet-50; magenta triangles denote ViT-B/16. Annotations give partial Pearson correlations over approximately $80{,}000$ samples.

layers for each experiment and compares early- and late-layer interventions. Within that layer, i indexes right-singular directions in decreasing order of singular value; p_i and q_i are the weight and activation energy fractions in direction i. We omit the layer index on these quantities in the rest of this section for ease of notation. Let K be a selected set of direction indices and $k = |K|$ its size. The alignment-target penalty $\beta(L_\ell - L^\star)^2$ encourages a chosen alignment level. The occupancy penalty $R_q = -\sum_{i \in K} q_i$ encourages activation energy in a selected right-singular subspace, with the basis detached from differentiation. The spectrum penalty $R_p = -\sum_{i \le k} p_i$ encourages weight energy in the leading singular directions.

Comparing these interventions tests the effects of changing activation occupancy and the weight spectrum. Comparisons between equal-dimensional subspaces test how the choice of directions affects learning (Section 5.4 and Appendix G.11). The theoretical grounding for the alignment and occupancy penalties, including their activation-space gradients and proofs, is given in Appendix D.

5 Experiments

We show that log-alignment profiles provide an effective way to detect unfamiliar processing and assess prediction reliability across domains with different input modalities, architectures, and output spaces. The vision experiments test whether log-alignment profiles detect shifts beyond output confidence, help predict classification errors, and localize corruptions within images (Section 5.1). We then evaluate text and visual shift detection in language models (Section 5.2) and failure detection in visuomotor policies (Section 5.3). Finally, we test whether alignment interventions change generalization (Section 5.4).

5.1 Vision Experiments

We test whether the log-alignment profile reveals changes that classification confidence misses and whether they help identify prediction errors in image-based classification. We evaluate ResNet and ViT classifiers from CIFAR to ImageNet scale under corruptions and semantic shifts (Appendices G and G.3). The baselines include MSP, entropy, energy, and maximum logit (Appendix G.3), as well as feature norm, Mahalanobis distance (Lee et al., 2018), KNN (Sun et al., 2022), ViM (Wang et al., 2022), ReAct (Sun et al., 2021), and ASH-S (Djurisic et al., 2023). Gram statistics (Sastry & Oore, 2020) and matched-depth feature-norm and pooled-feature Mahalanobis profiles test whether measuring across layers alone explains the gains.

Log-alignment profile detects image shifts that confidence misses. We distinguish changes in image statistics from unfamiliar content because corruption can alter how familiar objects appear without changing their class. Detecting these structural perturbations would complement methods that primarily respond to semantic novelty. For ImageNet-1K ResNet-50 and ViT-B/16, we separate the two effects using independent measurements. Structural distance D_{s} measures departures from ID color, edge, and spatial-frequency statistics; semantic distance D_{m} measures CLIP-embedding distance to nearby ID images (Appendix H). We measure each score's association with either distance using partial Pearson correlation r, controlling for the other distance. The log-alignment profile tracks

Score	ResNet-50 con. s3	ResNet-50 con. s5	ResNet-50 noise	ResNet-50 C-con.	ResNet-50 C-def.	ViT-B/16 con. s3	ViT-B/16 con. s5	ViT-B/16 noise	ViT-B/16 C-con.	ViT-B/16 C-def.
U_{LAP} (**log-alignment profile**)	0.822	**0.983**	**1.000**	**1.000**	**0.998**	**0.897**	**0.998**	0.971	**1.000**	**0.990**
randomized V	0.542	0.717	0.952	0.900	0.906	0.578	0.827	0.562	0.924	0.822
Gram, class-cond.	**0.846**	0.951	0.970	0.976	0.970	0.688	0.919	0.750	0.964	0.914
Gram, global	0.550	0.958	0.971	0.995	0.990	0.646	0.988	0.669	0.999	0.975
7-layer feat. Maha	0.482	0.722	0.998	0.859	0.957	0.560	0.972	**0.986**	0.997	0.974
entropy	0.503	0.498	0.711	0.784	0.875	0.567	0.921	0.789	0.999	0.845
energy	0.501	0.490	0.497	0.617	0.636	0.587	0.940	0.775	0.999	0.767
penult. feat. Maha	0.486	0.504	0.914	0.919	0.986	0.509	0.686	0.733	0.924	0.932
KNN	0.497	0.499	0.730	0.797	0.903	0.505	0.621	0.585	0.831	0.743
ViM	0.486	0.500	0.878	0.899	0.984	0.519	0.837	0.758	0.980	0.954
DICE (Sun & Li, 2022)	0.497	0.483	0.221	0.286	0.244	0.486	0.728	0.755	0.880	0.847
SCALE (Xu et al., 2024)	0.498	0.480	0.497	0.639	0.780	0.565	0.937	0.756	0.999	0.698
GEN (Liu et al., 2023b)	0.506	0.512	0.841	0.881	0.953	0.588	0.933	0.815	0.999	0.889

Table 1: **Log-alignment profile detects shifts missed by output and penultimate-feature scores.** Detection AUROCs use the same calibration split. Columns show synthetic contrast at severities 3 and 5, noise at severity 5, and ImageNet-C contrast and defocus at severity 5. Randomized V changes only the monitor's basis. Gram calibration (Sastry & Oore, 2020) is class-conditional or global. Bold marks the best AUROC per column; Table 5 gives additional ablations.

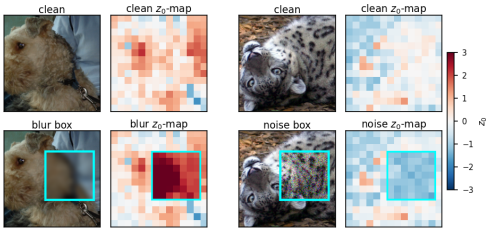
structural distance with $r = 0.54$ and $r = 0.52$, respectively (Figure 2). Output scores and feature Mahalanobis instead follow semantic distance. A controlled ImageNet-200 comparison confirms that corruption affects the log-alignment profile more than class novelty does (Appendix G.6).

A separate test matches CIFAR-10 and SVHN confidence histograms by retaining equal numbers of images in each ensemble-MSP bin (Figure 1). MSP detection is then at chance, but the log-alignment profile achieves 0.820–0.948 AUROC across three training label-noise levels (Appendix G.7). Table 1 tests whether this advantage extends to appearance corruptions, comparing output scores and internal-feature baselines. Contrast reduction changes image structure without changing the class; severity 5 reduces contrast more strongly than severity 3. On ResNet-50, the log-alignment profile achieves 0.82 and 0.98 AUROC at severities 3 and 5, while entropy, energy, feature Mahalanobis, KNN, and ViM remain near chance. Gram statistics win one contrast comparison but use a statistic 10^4 times larger and trail on most other tested sets (Table 5). Randomizing the monitor's right-singular basis while preserving singular values and network predictions reduces AUROC by up to 0.41.

Together, these tests show that the log-alignment profile detects altered image processing beyond class novelty and output confidence. The random-basis control links part of this sensitivity to the feature combinations that the trained layers amplify, rather than arbitrary input directions. Appendix G.2 shows where the changes occur across depth.

Predictor	ResNet-50	ViT-B/16
H	0.632	0.706
$\boldsymbol{H + U_{\mathrm{LAP}}}$	**0.811**	**0.849**
H + feat. Maha (min.)	0.793	0.791
H + feat. Maha (pred.)	0.799	0.792
H + DICE (Sun & Li (2022))	0.782	0.788
H + SCALE (Xu et al. (2024))	0.736	0.808
H + GEN (Liu et al. (2023b))	0.796	0.785

(a)



(b)

Table 2: **Alignment measurements improve error prediction and localize spatial anomalies.** (a) The table reports held-out error-prediction AUROC using entropy H alone or with an additional score. Feature Mahalanobis uses the minimum over classes ("min.") or the predicted class ("pred."). (b) First-block, position-calibrated alignment maps for ViT-B/16 show localized increases under blur (red) and decreases under noise (blue).

Log-alignment profile adds uncertainty information. We next test whether the log-alignment profile helps identify mistakes by the same ImageNet classifiers. We fit logistic error predictors

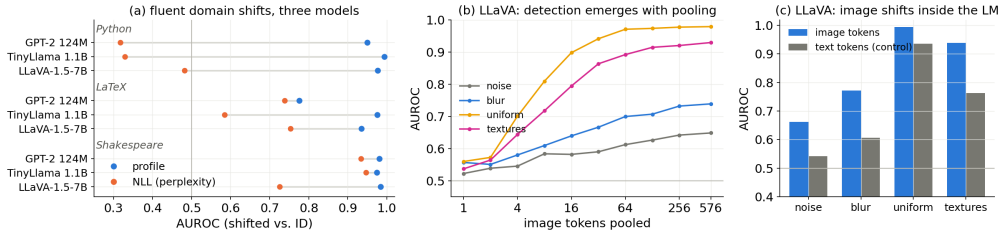


Figure 3: **Language and multimodal log-alignment profiles detect departures from the reference distribution.** (a) The log-alignment profile (blue) detects Python code that perplexity (orange) misses. (b) LLaVA detection improves when more image-token positions are pooled. (c) Image-token log-alignment profiles outperform the text-token control, which remains near chance on noise and blur. Appendix G.12 gives the protocols and numerical results.

using entropy alone or combined with an internal score on a balanced mixture of clean and shifted images. Log-alignment profile calibration is ID-only, but this supervised predictor uses labeled clean and shifted examples; fitting and evaluation examples are disjoint (Appendix G.14). Here, error-prediction AUROC measures separation between correct and incorrect classifications. Adding the log-alignment profile raises held-out AUROC from 0.632 to 0.811 on ResNet-50 and from 0.706 to 0.849 on ViT-B/16, outperforming combinations of entropy with feature Mahalanobis, DICE, SCALE, or GEN (Table 2a). The log-alignment profile thus helps distinguish correct from incorrect predictions beyond what output entropy reveals, without changing the classifier or its predictions.

Log-alignment profile localizes anomalies in input images. To test whether alignment can locate a corruption within an image, we measure it separately at each ViT-B/16 patch token and compare with token norm, calibrating both at each position on clean images. We compare patch-level AUROC for separating corrupted from uncorrupted patches. The first-block maps reveal the blurred and noisy regions (Table 2b). On held-out images with Gaussian noise inside random boxes, the alignment deficit achieves 0.924 patch-level AUROC, versus 0.466 for token norm. The signal therefore identifies where processing becomes atypical, beyond changes in activation magnitude alone. Blur raises alignment while noise lowers it, so the maps retain information about how the local response changes as well as where it changes. Appendix G.13 gives the protocol and maps across depth.

5.2 Language Experiments

We test whether the log-alignment profile can flag changes in text domain or visual input even when token predictability or prompt-text responses give little warning. Such a monitor could help identify inputs on which to check model accuracy: an input can differ from an application's reference data even when the model assigns its tokens high probability.

For GPT-2, TinyLlama-1.1B, and LLaVA-1.5-7B's language tower, we probe every decoder block and calibrate each model's log-alignment profile using enwik8 text only. We then test whether the score separates held-out enwik8 passages from each alternative domain: Python source, LaTeX source, and Shakespeare. We also test shuffled or reversed enwik8 tokens, uniform random tokens, and base64 strings. Here a shift means a departure from the calibration corpus, not necessarily from the model's pretraining data. The baseline is next-token perplexity, which measures input predictability. For LLaVA's visual test, we change only the images and keep the image-description prompt fixed. We calibrate on clean ImageNet images and compare image-token log-alignment profiles with a prompt-text control (Appendix G.12). Both tests compare detection AUROC between reference and shifted inputs.

Log-alignment profile detects text and visual shifts inside language models. Across the three models, Python code is at least as predictable as the reference text, so perplexity detects it at only 0.32–0.48 AUROC; the log-alignment profile reaches 0.95–0.99 (Figure 3a). For LLaVA's visual inputs, image-token log-alignment profiles detect noise, blur, random images, and textures at 0.66–0.99 AUROC. The prompt-text control performs worse on every shift and remains near chance on noise and blur (Figure 3c). Individual image positions are also near chance for those two corruptions, but averaging measurements over more positions within each layer improves detection (Figure 3b).

Perplexity measures whether the model can predict the next token; the log-alignment profile measures whether its internal responses resemble those produced by the reference inputs. Specifically, the models predict Python code well, but process it differently from enwik8 text. In LLaVA, monitoring the prompt-text positions alone misses noise and blur that are detectable at image-token positions. The gain from pooling suggests that these corruptions cause small changes across many positions, which become easier to detect when averaged together.

5.3 Visuomotor Experiments

We then test the log-alignment profile on robotics domains. We test whether log-alignment profiles warn of control failures beyond what action uncertainty reveals. We evaluate fixed PPO policies on ManiSkill3 (Tao et al., 2025), BC-RNN policies on LIBERO (Liu et al., 2023a), pretrained 7B OpenVLA-OFT checkpoints (Kim et al., 2025), and the RoboCasa365 generalist Diffusion Policy (Nasiriany et al., 2026; Chi et al., 2023). We calibrate on ID episodes and score each policy query. The baselines are action entropy where available, disagreement among independently trained policies, penultimate-feature Mahalanobis distance, and calibrated pixel statistics. Detection AUROC separates ID from shifted observations per policy query; failure-prediction AUROC separates successful from failed episodes. We also compare correlations with batch failure rates and alarm timing in environment steps (Appendix I).

Log-alignment profile detects distribution shifts and tracks policy failure. Table 3a compares degraded camera inputs during familiar tasks with held-out tasks and physical settings evaluated without added corruption. The log-alignment profile detects added Gaussian noise at 1.00 AUROC on both RoboCasa365 and OpenVLA-OFT, compared with 0.41 for pixel statistics on RoboCasa365 and 0.52 for feature Mahalanobis on OpenVLA-OFT.

The held-out conditions test whether this sensitivity extends beyond image degradation. On RoboCasa365, the log-alignment profile detects held-out composite tasks at 0.96 AUROC and familiar tasks in held-out kitchens or with held-out object instances at 0.85 and 0.82, respectively. Feature Mahalanobis is stronger on held-out tasks (1.00), but weaker on kitchens and object instances (0.55 and 0.41). The log-alignment profile thus adds sensitivity to changes in the physical setting, not just to synthetic noise. The LIBERO cross-suite tests likewise evaluate policies on tasks outside their training suite (Appendix I).

We separately test whether these signals predict failure. Across 192 batches from fixed PPO policies, log-alignment profile scores correlate positively with failure rate ($r = 0.58$), while disagreement decreases as failure rises ($r = -0.23$; Table 3c). Within the same shifted condition, adding the log-alignment profile to disagreement improves episode-failure AUROC by 0.144–0.244 across eight ManiSkill tasks (Table 3b). When corruption begins mid-episode, log-alignment profile alarms precede action-uncertainty alarms; LIBERO contrast alarms precede failure by 190 steps among failing episodes (Figure 18).

These results explain why internal monitoring complements action uncertainty: policies can agree while their observations produce unfamiliar internal responses. The log-alignment profile detects that mismatch early and adds information about which ManiSkill episodes fail. It provides a warning signal, not a guarantee of failure.

5.4 Grokking Experiments

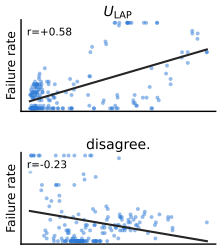
We test whether log-alignment profile is also a useful intervention target. Grokking separates memorization from later generalization (Power et al., 2022), allowing us to measure whether an intervention shortens that delay. Following Shehper & Vaswani (2026); Truong et al. (2026), we resume a two-layer modular-addition transformer from a memorized checkpoint and compare alignment targeting, spectrum concentration (R_p), and activation concentration (R_q) with continued training without these penalties. All conditions restore each weight matrix's Frobenius norm to its checkpoint value after every step, excluding norm growth as an explanation. A second test compares equal-dimensional subspaces on S_5 xyx, where the model predicts the composition $x \circ y \circ x$ of two permutations of five elements. The protocols are in Appendices G.10 and G.11. We compare the additional training steps needed to reach 99% validation accuracy and the fraction of runs that reach this threshold within the training horizon.

Policy	Condition	fail	log-alignment profile	feat	pix	ent	dis
PPO (ManiSkill)	held-out geometry	0.60	**0.61**	0.60	0.49	—	0.59
	Gaussian noise	0.42	**0.95**	0.46	0.80	—	0.50
BC-RNN (LIBERO)	contrast	0.89	**1.00**	0.44	0.49	0.54	0.70
	noise	1.00	**1.00**	0.48	0.46	0.44	0.85
	suite: Object	1.00	**1.00**	0.72	0.81	0.58	0.82
	suite: Goal	1.00	**1.00**	0.34	0.67	0.63	0.82
Diffusion Policy (RoboCasa365)	Gaussian noise	0.93	**1.00**	0.63	0.41	—	—
	held-out tasks	1.00	0.96	**1.00**	0.51	—	—
	held-out kitchens	0.61	**0.85**	0.55	0.69	—	—
	held-out instances	0.33	**0.82**	0.41	0.42	—	—
OpenVLA-OFT 7B (LIBERO)	Gaussian noise (Goal ckpt)	0.81	**1.00**	0.52	0.60	—	—
	Object ckpt on Spatial	1.00*	**0.99**	0.85	0.47	—	—

(a) Detection across testbeds

Task	dis	dis+log-alignment profile
PullCube	0.539	**0.724**
PickCube	0.450	**0.694**
PushT	0.478	**0.709**
AnymalC-Reach†	0.511	**0.676**
PickYCB	0.683	**0.827**
PlaceSphere	0.502	**0.688**
LiftPegUpright	0.513	**0.706**
PokeCube	0.535	**0.775**

(b) Failure prediction

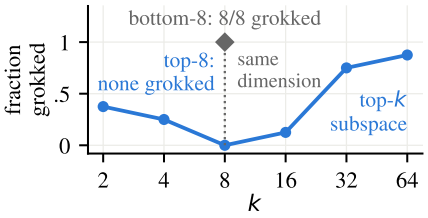


(c) Failure rate tracking

Table 3: **Shift detection and episode-failure prediction across control testbeds.** (a) Scores are per-step detection AUROCs; *fail* is the episode failure rate. Baselines are feature Mahalanobis (*feat*), pixel statistics (*pix*), action entropy (*ent*), and three-policy disagreement (*dis*). (b) Adding the log-alignment profile improves within-condition episode-failure AUROC on ManiSkill PPO tasks (likelihood-ratio $p \leq 0.012$ for each). The dagger denotes locomotion. (c) Points represent 192 batches of 32 episodes, relating failure rate to log-transformed mean score; black lines show pooled linear fits. Appendix I gives details.

Arm	steps	Δ
baseline	4367 ± 202	—
early $L^\star$	4483 ± 176	$+3\%$
late L*	$\mathbf{2917 \pm 425}$	$\mathbf{-33\%}$
all-FF $L^\star$	3567 ± 562	-18%
q-side (R_q)	3750 ± 350	-14%
p-side (R_p)	4317 ± 289	-1%
p+q	3833 ± 369	-12%

(a) Grokking interventions



(b) Direction identity

Table 4: **Alignment interventions change grokking dynamics depending on the layer and subspace.** (a) Steps to generalization after the checkpoint are reported as mean ± standard deviation over three matched seeds; Δ is the percentage change from baseline. (b) On S_5 xyx, none of eight top-8 runs groks within the training horizon, while all eight bottom-8 runs do (Appendix G.11).

Log-alignment profile interventions accelerate grokking dynamics. Targeting a late layer reduces the delay to generalization by 33%, whereas early-layer targeting changes it little (+3%; Table 4a). At the first feed-forward projection of the first block, concentrating the weight spectrum through R_p also has little effect (−1%), but encouraging activation energy in selected right-singular directions through R_q reduces the delay by 14%. On S_5 xyx, we apply R_q to the late feed-forward matrix's eight most-amplified or eight least-amplified input directions. None of eight top-8 runs groks within the training horizon, while all eight bottom-8 runs do (Table 4b).

The equal-dimensional subspaces produce opposite outcomes, so concentration alone cannot explain the effect. Thus, which directions receive activation energy matters, as does the layer where we intervene. Alignment is therefore more than a passive correlate of learning, but increasing it indiscriminately can hinder generalization.

6 Conclusion

Log-alignment profiles turn two measurements per layer into a compact monitor of internal computation. Gain records alignment with learned directions, while radius preserves activation scale. Across image classifiers, language models and robot policies, the log-alignment profiles detect shifts that output scores miss and reveal where responses become atypical across depth and token positions. Their practical value extends to improved error prediction and earlier alarms in closed-loop control. The grokking interventions connect measurement to learning: late-layer targeting reduces generalization delay, and equal-dimensional subspaces produce different outcomes depending on the selected directions. Together, these results establish layerwise amplification and scale as useful signals for monitoring networks and studying their learning dynamics. We discuss the limitations in Appendix K.

Ethics Statement

This paper contributes to the monitoring of deployed machine learning systems. Better attribution of distribution shift can improve the safety of systems that must recognize when their operating conditions have changed. The profile is diagnostic and adds no data collection requirement beyond a held-out in-distribution calibration set.

Reproducibility Statement

The quantity we measure is defined in closed form in Sections 3 and 4, and every theoretical statement is proved in Appendices A to F. Evaluation sets, calibration splits, seeds and baseline definitions are listed in Appendix G (vision and language) and Appendix I (control), with the grokking recipe in Appendix G.10; code will be open-sourced. Inference overhead and calibration footprint are reported in Appendix J.

AI Use Statement

Large language models were used to edit prose for clarity and concision, to assist in writing and refactoring experiment and figure scripts, and to check citation metadata. The research questions, methods, proofs, experiments and their interpretation were conceived, executed and verified by the authors, who take full responsibility for the content.

References

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing ImageNet out-of-distribution detection evaluation. In *International Conference on Machine Learning*, 2023.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

Shih-Han Chou, Shivam Chandhok, Jim Little, and Leonid Sigal. Mm-r3: On (in-) consistency of vision-language models (vlms). In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4762–4788, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Thomas G. Dietterich and Alex Guyer. The familiarity hypothesis: Explaining the behavior of deep open set methods. *Pattern Recognition*, 132:108931, 2022.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations*, 2023.

Katie E Everett, Lechao Xiao, Mitchell Wortsman, Alexander A Alemi, Roman Novak, Peter J Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12666–12700. PMLR, 2024. URL https://proceedings.mlr.press/v235/everett24a.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Intl. Conf. on Machine Learning (ICML)*, pp. 1050–1059. PMLR, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Rui Huang and Yixuan Li. MOS: Towards scaling out-of-distribution detection for large semantic space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Robotics: Science and Systems (RSS)*, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Gregory Kang Ruey Lau, Hieu Dao, Nicole Kan Hui Lin, and Bryan Kian Hsiang Low. Uncertainty quantification for multimodal large language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, 2023a.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. GEN: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23946–23955, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Liu_GEN_Pushing_the_Limits_of_Softmax-Based_Out-of-Distribution_Detection_CVPR_2023_paper.html.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations*, 2026.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, 1996.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Theodore Papamarkou, Jacob Hinkle, M Todd Young, and David Womble. Challenges in markov chain monte carlo for bayesian neural networks. *Statistical Science*, 37(3):425–442, 2022.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-OOD detection. *arXiv preprint arXiv:2106.09022*, 2021.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In *International Conference on Machine Learning*, 2020.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Ali Shehper and Ashish Vaswani. A training-time diagnostic for generalization via the log-alignment ratio. *arXiv preprint arXiv:2605.28975*, 2026.

Achyuthan Sivasankar. Circuit synchronization precedes generalization: A causal precursor to grokking. *arXiv preprint arXiv:2606.12966*, 2026.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 979–995, 2024.

Yiyou Sun and Yixuan Li. DICE: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136840680.pdf.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 2022.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Demonstrating GPU parallelized robot simulation and rendering for generalizable embodied AI with ManiSkill3. In *Robotics: Science and Systems (RSS)*, 2025.

Dustin Tran, Jeremiah Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, et al. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*, 2022.

Xuan Khanh Truong, Hoang Viet Doan, Duc Trung Luu, and Thanh Duc Phan. The weight norm sets the grokking timescale: A causal delay law. *arXiv preprint arXiv:2606.13753*, 2026.

Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1899–1910, 2023.

Aishwarya Venkataramanan, Paul Bodesheim, and Joachim Denzler. Probabilistic embeddings for frozen vision-language models: Uncertainty quantification with Gaussian process latent variable models. In *Conference on Uncertainty in Artificial Intelligence*, pp. 4309–4328, 2025.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training-time and post-hoc out-of-distribution detection enhancement. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/46d943bc6a15a57c923829efc0db7c7a-Paper-Conference.pdf.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132:5635–5662, 2024.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2:1–32, 2024a.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024b.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *International Conference on Learning Representations*, 2024.

A Algebra of Sample-Wise LAR

Throughout, $W \in \mathbb{R}^{m\times d}$ is a nonzero matrix with singular value decomposition $W = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal and $\Sigma \in \mathbb{R}^{m\times d}$ has diagonal entries $\sigma_1 \geq \cdots \geq \sigma_{\min(m,d)} \geq 0$. We use the full decomposition, so that $\{v_1, \ldots, v_d\}$ is an orthonormal basis of $\mathbb{R}^d$ even when W is rank deficient; the weight spectrum Equation (1) then assigns $p_i = 0$ to any direction with $\sigma_i = 0$, and $\sum_{i=1}^d p_i = 1$ because $\|W\|_F^2 = \sum_i \sigma_i^2$.

Proof of Theorem 3.3. Fix $h \neq 0$. Expanding in the right-singular basis and using orthogonality of U,

$$\begin{aligned}\|Wh\|_2^2 &= h^\top W^\top W h = h^\top V\Sigma^\top U^\top U\Sigma V^\top h \\ &= h^\top V\Sigma^\top\Sigma V^\top h = \sum_{i=1}^d \sigma_i^2\,(v_i^\top h)^2,\end{aligned} \tag{10}$$

since $\Sigma^\top\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. Dividing Equation (10) by $\|W\|_F^2\|h\|_2^2 = \big(\sum_j \sigma_j^2\big)\|h\|_2^2$, which is nonzero because $W \neq 0$ and $h \neq 0$, and regrouping the factors,

$$\frac{\|Wh\|_2^2}{\|W\|_F^2\|h\|_2^2} = \sum_{i=1}^d \frac{\sigma_i^2}{\sum_j \sigma_j^2}\cdot\frac{(v_i^\top h)^2}{\|h\|_2^2} = \sum_{i=1}^d p_i\, q_i(h) = S(W,h). \tag{11}$$

This proves the first identity. Applying $\frac{1}{2}\log_d$ to both sides of Equation (11) and using $\frac{1}{2}\log_d(a^2/b^2) = \log_d(a/b)$ for $a, b > 0$ gives

$$\tfrac{1}{2}\log_d S(W,h) = \log_d \frac{\|Wh\|_2}{\|W\|_F\|h\|_2},$$

and adding 1 yields the second identity. Note $S > 0$ exactly when $Wh \neq 0$; if h lies in the nullspace of W then $S = 0$ and $L = -\infty$, consistent with Proposition A.2 below.

For the final claim, observe that $q(h)$ is a probability vector: it is nonnegative, and $\sum_i q_i(h) = \|V^\top h\|_2^2/\|h\|_2^2 = 1$ by Parseval's identity, since V is orthogonal. Hence $S(W,h) = p^\top q(h)$ is a linear functional of $q(h)$ with coefficient vector p determined entirely by W. For a fixed trained matrix, p does not depend on the input; therefore any two inputs h, h' with $q(h) = q(h')$ satisfy $L(W,h) = L(W,h')$, and all input-dependent variation in L enters through q. □

Proposition A.1 (Invariance)**.** *For every nonzero h and nonzero scalar c, $L(cW,h) = L(W,ch) = L(W,h)$. Moreover $L(PWQ^\top, Qh) = L(W,h)$ for all orthogonal $P \in \mathbb{R}^{m\times m}$, $Q \in \mathbb{R}^{d\times d}$.*

Proof. All three claims follow from the norm-ratio form Equation (5), whose argument is $\rho(W,h) = \|Wh\|_2/(\|W\|_F\|h\|_2)$; since $\log_d$ is injective it suffices to show ρ is unchanged.

Scaling of W. $\|(cW)h\|_2 = |c|\,\|Wh\|_2$ and $\|cW\|_F = |c|\,\|W\|_F$, so the factor $|c|$ cancels: $\rho(cW,h) = \frac{|c|\|Wh\|_2}{|c|\|W\|_F\|h\|_2} = \rho(W,h)$.

Scaling of h. $\|W(ch)\|_2 = |c|\,\|Wh\|_2$ and $\|ch\|_2 = |c|\,\|h\|_2$, so again the factor cancels.

Orthogonal reparameterization. Using $Q^\top Q = I$,

$$(PWQ^\top)(Qh) = PW(Q^\top Q)h = PWh.$$

Orthogonal maps preserve the Euclidean norm, so $\|PWh\|_2 = \|Wh\|_2$ and $\|Qh\|_2 = \|h\|_2$. They also preserve the Frobenius norm, since $\|PWQ^\top\|_F^2 = \mathrm{tr}(QW^\top P^\top PWQ^\top) = \mathrm{tr}(QW^\top WQ^\top) = \mathrm{tr}(W^\top W) = \|W\|_F^2$ by cyclicity of the trace. Hence all three factors of ρ are unchanged. □

The third statement is a genuine change of coordinates: it rotates the weights and the activations *together*. The randomized-direction control of Appendix C does not do this. It replaces V by an independent random basis while leaving the activations where they are, which is precisely why it destroys the learned correspondence rather than preserving it.

Proposition A.2 (Spectral bounds and stable-rank ceiling)**.** *For every nonzero h,*

$$\min_i p_i \;\le\; S(W,h) \;\le\; p_{\max} := \max_i p_i = \frac{\sigma_1^2}{\|W\|_F^2}, \tag{12}$$

and consequently $L(W,h) \le 1 - \frac{1}{2}\log_d \operatorname{srank}(W)$, where $\operatorname{srank}(W) = \|W\|_F^2/\|W\|_2^2$ is the stable rank. The upper bound is attained exactly when h lies in the top right-singular subspace. If W has full column rank then $\min_i p_i > 0$ and L is bounded below; otherwise S can equal 0 and L is unbounded below.

Proof. By Theorem 3.3, $S(W,h) = p^\top q(h)$ where $q(h)$ is a probability vector. A convex combination of the numbers $\{p_i\}$ lies between their minimum and maximum, which gives Equation (12); the upper bound is attained iff q places all its mass on indices achieving $p_{\max}$, i.e. iff h lies in the span of the top singular directions. Since $\|W\|_2 = \sigma_1$, we have $p_{\max} = \sigma_1^2/\|W\|_F^2 = 1/\operatorname{srank}(W)$, so applying the increasing map $x \mapsto 1 + \frac{1}{2}\log_d x$ to $S \le p_{\max}$ gives

$$L(W,h) \le 1 + \tfrac{1}{2}\log_d \frac{1}{\operatorname{srank}(W)} = 1 - \tfrac{1}{2}\log_d \operatorname{srank}(W).$$

If W has full column rank then $\sigma_d > 0$, so $\min_i p_i > 0$ and $L \ge 1 + \frac{1}{2}\log_d \min_i p_i > -\infty$. If instead $\sigma_i = 0$ for some i, taking $h = v_i$ gives $q = e_i$, hence $S = p_i = 0$ and $L = -\infty$. □

Proposition A.3 (Overlap and correlation forms agree)**.** *Let $\sigma_p^2 = \frac{1}{d}\sum_i (p_i - \frac{1}{d})^2$ and σ_q^2 likewise, and let $\rho = \operatorname{corr}(p,q)$ be the Pearson correlation of the two spectra across their d coordinates. Then*

$$\sum_i p_i q_i = \frac{1}{d} + d\,\rho\,\sigma_p\sigma_q, \tag{13}$$

and consequently $\mathrm{LAR} = \frac{1}{2} + \frac{1}{2}\log_d\big(1 + d^2\rho\,\sigma_p\sigma_q\big)$. In particular $\mathrm{LAR} \gtrless \frac{1}{2}$ according as $\rho \gtrless 0$.

Proof. Both p and q are probability vectors on d coordinates, so their coordinate averages are $\bar p = \bar q = 1/d$. The covariance across coordinates is therefore

$$\begin{aligned}\operatorname{cov}(p,q) &= \frac{1}{d}\sum_i \Big(p_i - \frac{1}{d}\Big)\Big(q_i - \frac{1}{d}\Big) \\ &= \frac{1}{d}\left(\sum_i p_i q_i - \frac{1}{d}\sum_i p_i - \frac{1}{d}\sum_i q_i + \frac{d}{d^2}\right) = \frac{1}{d}\left(\sum_i p_i q_i - \frac{1}{d}\right),\end{aligned}$$

since $\sum_i p_i = \sum_i q_i = 1$. Rearranging gives $\sum_i p_i q_i = \frac{1}{d} + d\operatorname{cov}(p,q)$, and substituting $\operatorname{cov}(p,q) = \rho\,\sigma_p\sigma_q$ yields Equation (13). Inserting this into $\mathrm{LAR} = 1 + \frac{1}{2}\log_d \sum_i p_i q_i$ and factoring $1/d$ out of the logarithm,

$$\mathrm{LAR} = 1 + \tfrac{1}{2}\log_d\Big(\tfrac{1}{d}\big(1 + d^2\rho\sigma_p\sigma_q\big)\Big) = \tfrac{1}{2} + \tfrac{1}{2}\log_d\big(1 + d^2\rho\sigma_p\sigma_q\big),$$

using $\log_d(1/d) = -1$. The final claim follows because $\sigma_p, \sigma_q \ge 0$ and $\log_d$ is increasing, with equality $\mathrm{LAR} = \frac{1}{2}$ iff $\rho\sigma_p\sigma_q = 0$, i.e. iff the spectra are uncorrelated or either one is constant. □

The two readings are complementary. The correlation form explains the $\frac{1}{2}$ baseline and the role of spectral spread, which is how the diagnostic's training dynamics are interpreted; the overlap form is linear in q, which is what yields the closed form of Theorem 3.3 and the per-sample restriction used throughout this paper.

B Head Alignment and Confidence

This appendix proves the head claims of Section 4.1. Evaluating Equation (5) at the classifier—W the classifier matrix, h the penultimate feature, $\tilde z = Wh$ the bias-corrected logits—gives

$$L_{\mathrm{head}}(x) = 1 + \log_d \|\tilde z\|_2 - \log_d \|h\|_2 - \log_d \|W\|_F, \tag{14}$$

which for fixed W turns on one number, the stretch $\|\tilde{z}\|_2/\|h\|_2$. Below we show that head alignment and maximum softmax probability share a radial factor but neither is a function of the other: a near-tie and a landslide can have logit vectors of the same length, and any confidence level is reachable at any length.

Setting. Let $z \in \mathbb{R}^C$ be the logits of an input under a classifier with $C \geq 2$ classes, $\mathrm{softmax}(z)_c = e^{z_c}/\sum_{c'} e^{z_{c'}}$, and $\mathrm{MSP}(z) = \max_c \mathrm{softmax}(z)_c$. Write the polar decomposition $z = \|z\|_2\,\hat{z}$ with $\hat{z} \in S^{C-1}$. A function of z is *radial* if it depends on z only through $\|z\|_2$, and *angular* if only through $\hat{z}$. We take the head bias to be zero so that the bias-corrected logits of Equation (14) satisfy $\tilde{z} = z$; a fixed nonzero bias makes $\|\tilde{z}\|_2 = \|z - b\|_2$ and $\mathrm{MSP}(z)$ functions of different vectors, which only weakens the relationship analyzed here.

Proposition B.1 (Head alignment is radial)**.** *For a fixed classifier matrix W and fixed feature norm $\|h\|_2$, L_{head} is a strictly increasing function of $\|z\|_2$, and is invariant under every rotation of z that preserves its norm.*

Proof. By Equation (14), $L_{\mathrm{head}} = 1 + \log_d \|z\|_2 - \log_d \|h\|_2 - \log_d \|W\|_F$. With W and $\|h\|_2$ held fixed, the last two terms are constants, so $L_{\mathrm{head}} = \log_d \|z\|_2 + \mathrm{const}$, and $\log_d$ is strictly increasing on $(0, \infty)$. Replacing z by Oz for orthogonal O leaves $\|z\|_2$ unchanged and hence leaves L_{head} unchanged, whereas it generally changes MSP. □

Proposition B.2 (MSP is shift-invariant and increases along rays)**.** *(i)* $\mathrm{MSP}(z + c\mathbf{1}) = \mathrm{MSP}(z)$ *for every $c \in \mathbb{R}$; consequently MSP is a function of the differences $\{z_c - z_{c'}\}$ alone. (ii) Let $\hat{z} \in S^{C-1}$ have a unique largest entry. Then $t \mapsto \mathrm{MSP}(t\hat{z})$ is strictly increasing on $(0, \infty)$, with $\lim_{t\to 0^+} \mathrm{MSP}(t\hat{z}) = 1/C$ and $\lim_{t\to\infty} \mathrm{MSP}(t\hat{z}) = 1$.*

Proof. (i) For every c', $\mathrm{softmax}(z + c\mathbf{1})_{c'} = e^{z_{c'}+c}/\sum_{c''} e^{z_{c''}+c} = e^c e^{z_{c'}}/\big(e^c \sum_{c''} e^{z_{c''}}\big) = \mathrm{softmax}(z)_{c'}$, so the whole softmax vector, and hence its maximum, is unchanged. Writing $\mathrm{softmax}(z)_{c'} = \big(\sum_{c''} e^{z_{c''} - z_{c'}}\big)^{-1}$ exhibits MSP as a function of differences only.

(ii) Let $\hat{z}_1 = \max_c \hat{z}_c$, unique by assumption, and set $p(t) = \mathrm{softmax}(t\hat{z})$, so that $\mathrm{MSP}(t\hat{z}) = p_1(t)$ for all $t > 0$. Then

$$\log p_1(t) = t\hat{z}_1 - \log \sum_c e^{t\hat{z}_c},$$

and differentiating, using $\frac{d}{dt} \log \sum_c e^{t\hat{z}_c} = \sum_c p_c(t)\hat{z}_c$,

$$\frac{d}{dt} \log p_1(t) = \hat{z}_1 - \sum_c p_c(t)\,\hat{z}_c. \tag{15}$$

Because $p(t)$ is a probability vector with strictly positive entries, $\sum_c p_c(t)\hat{z}_c$ is a strict convex combination of the $\hat{z}_c$, and since $\hat{z}_1$ is the unique maximum the combination is strictly less than $\hat{z}_1$. Hence the derivative in Equation (15) is strictly positive and p_1 is strictly increasing. As $t \to 0^+$ we get $p(t) \to \mathbf{1}/C$, giving the limit $1/C$; as $t \to \infty$, $p_1(t) = \big(1 + \sum_{c\neq 1} e^{-t(\hat{z}_1 - \hat{z}_c)}\big)^{-1} \to 1$ because every exponent $\hat{z}_1 - \hat{z}_c$ is strictly positive. □

Proposition B.2(ii) is the reason the two quantities correlate at all: along a fixed logit direction, increasing the radius increases both L_{head} (Proposition B.1) and MSP. The next two results show that this is the only link between them, and that it is not a functional one.

Lemma B.3 (Range of confidence at fixed logit norm)**.** *Let $C \geq 2$ and $s > 0$. On the sphere $\{z : \|z\|_2 = s\}$, MSP attains exactly the interval*

$$\left[\frac{1}{C},\ \left(1 + (C-1)\,e^{-s\sqrt{C/(C-1)}}\right)^{-1}\right], \tag{16}$$

the lower endpoint at $z = (s/\sqrt{C})\mathbf{1}$ and the upper at $z_1 = s\sqrt{(C-1)/C}$ with $z_c = -s/\sqrt{C(C-1)}$ for $c \neq 1$ (and its coordinate permutations).

Proof. The sphere is compact and connected for $C \geq 2$, and MSP is continuous, so its image is a compact interval and it suffices to compute the two extrema.

Minimum. For any probability vector p we have $\max_c p_c \geq \frac{1}{C}\sum_c p_c = \frac{1}{C}$, with equality iff all entries are equal, which for a softmax happens iff all z_c coincide. The point $z = (s/\sqrt{C})\mathbf{1}$ has norm s and attains it.

Maximum. By permutation symmetry of both MSP and the constraint we may restrict to the closed region where $z_1 = \max_c z_c$ and write

$$\mathrm{MSP}(z) = \left(1 + \textstyle\sum_{c\neq 1} e^{-(z_1 - z_c)}\right)^{-1} = \left(1 + e^{-z_1}\textstyle\sum_{c\neq 1} e^{z_c}\right)^{-1}, \tag{17}$$

so maximizing MSP at fixed z_1 is equivalent to minimizing $F(z_2, \ldots, z_C) = \sum_{c\neq 1} e^{z_c}$ subject to $\sum_{c\neq 1} z_c^2 = R^2$ with $R^2 = s^2 - z_1^2$.

We claim the minimizer of F on that sphere has all coordinates equal. The minimum exists by compactness, and at a minimizer the Lagrange condition $e^{z_c} = 2\mu z_c$ holds for all $c \neq 1$ with a common multiplier μ. If $\mu < 0$, then for each c the value z_c is a root of $\varphi(t) = e^t - 2\mu t$, and $\varphi'(t) = e^t - 2\mu > 0$ everywhere, so φ is strictly increasing and has at most one root: all z_c coincide. If instead $\mu \geq 0$ then $e^{z_c} = 2\mu z_c$ forces $z_c > 0$ for every c, whence $F > C - 1$; but the feasible point with all coordinates equal to $-R/\sqrt{C-1}$ gives $F = (C-1)e^{-R/\sqrt{C-1}} < C - 1$, so such a critical point is not the minimum. Hence at the minimizer $z_c = t$ for all $c \neq 1$.

Substituting into Equation (17) gives $\mathrm{MSP} = \left(1 + (C-1)e^{-(z_1 - t)}\right)^{-1}$, which is strictly increasing in the gap $z_1 - t$, so it remains to maximize $z_1 - t$ subject to $z_1^2 + (C-1)t^2 = s^2$. Writing $a = (z_1, \sqrt{C-1}\,t)$ and $b = (1, -1/\sqrt{C-1})$ we have $z_1 - t = \langle a, b\rangle$ and $\|a\|_2 = s$, so Cauchy–Schwarz gives

$$z_1 - t \;\leq\; \|a\|_2\|b\|_2 = s\sqrt{1 + \tfrac{1}{C-1}} = s\sqrt{\tfrac{C}{C-1}}, \tag{18}$$

with equality iff $a \parallel b$, i.e. $z_1 = \kappa$ and $t = -\kappa/(C-1)$ for some $\kappa > 0$. The norm constraint then gives $\kappa^2\big(1 + \frac{1}{C-1}\big) = s^2$, so $\kappa = s\sqrt{(C-1)/C}$ and $t = -s/\sqrt{C(C-1)}$. Substituting the maximal gap Equation (18) yields the upper endpoint of Equation (16). □

Two checks: at $C = 2$ the upper endpoint is the logistic $\left(1 + e^{-s\sqrt{2}}\right)^{-1}$, and as $s \to 0$ it tends to $1/C$, so the interval Equation (16) collapses to a point, as it must.

Proposition B.4 (Neither determines the other)**.** *For every $m \in (1/C, 1)$ the level set $\{z : \mathrm{MSP}(z) = m\}$ contains points of arbitrarily large norm and its norms are bounded away from zero. Consequently MSP is not a function of L_{head}, and L_{head} is not a function of MSP.*

Proof. Fix $m \in (1/C, 1)$ and let $\hat{z} \in S^{C-1}$ have a unique largest entry. By Proposition B.2(ii), $t \mapsto \mathrm{MSP}(t\hat{z})$ is a continuous, strictly increasing bijection from $(0, \infty)$ onto $(1/C, 1)$, so there is exactly one $t(\hat{z}) > 0$ with $\mathrm{MSP}(t(\hat{z})\hat{z}) = m$, and the corresponding point has norm $t(\hat{z})$.

Unbounded norms. Consider directions $\hat{z}^{(\epsilon)} \propto (1 + \epsilon, 1, \ldots, 1)$, normalized, whose largest gap tends to 0 as $\epsilon \to 0^+$. By Lemma B.3 applied at norm s, every point of norm s has $\mathrm{MSP} \leq \left(1 + (C-1)e^{-s\sqrt{C/(C-1)}}\right)^{-1}$; conversely, for a fixed direction with maximal gap $\gamma > 0$, $\mathrm{MSP}(t\hat{z}) \leq \left(1 + (C-1)e^{-t\gamma}\right)^{-1}$, so attaining the level m requires $t \geq \gamma^{-1}\log\big((C-1)m/(1-m)\big)$, which diverges as $\gamma \to 0^+$. Hence $t(\hat{z}^{(\epsilon)}) \to \infty$.

Norms bounded below. By Lemma B.3, any z with $\mathrm{MSP}(z) = m > 1/C$ must satisfy $\left(1 + (C-1)e^{-\|z\|_2\sqrt{C/(C-1)}}\right)^{-1} \geq m$, i.e. $\|z\|_2 \geq \sqrt{\frac{C-1}{C}}\,\log\big((C-1)m/(1-m)\big) > 0$.

Now suppose MSP were a function of L_{head}. Fixing W and $\|h\|_2$, Proposition B.1 makes L_{head} a bijective function of $\|z\|_2$, so MSP would be a function of $\|z\|_2$; but the range in Lemma B.3 shows that a single norm $s > 0$ carries an interval of MSP values of positive width $\left(1 + (C-1)e^{-s\sqrt{C/(C-1)}}\right)^{-1} - 1/C$, so two points of the same norm have different MSP, a contradiction. Conversely, suppose L_{head} were a function of MSP. Then all points with a given MSP would share

one norm; but the unbounded-norm construction above exhibits points with the same MSP and different norms, again a contradiction. □

The practical consequence is that the empirical correlation between head alignment and confidence is governed by the split of across-input variance in z between its radial and angular parts, which is a property of the trained model rather than of the definitions. Appendix G.1 and Figure 4 report both signs across training regimes, which is why the head is excluded from the structural detector.

C The Randomized-Direction Control

To test the role of learned directions, we keep the network and its predictions fixed and randomize the *measurement*: each layer's right-singular basis is replaced by an independent random one,

$$W_R = U\Sigma R^\top, \qquad R \sim \mathrm{Haar}\big(O(d)\big), \tag{19}$$

which preserves the singular values, the Frobenius norm and the shape of W while destroying the correspondence between activation directions and the basis. By Proposition A.1 this is *not* a change of coordinates: the activations are not rotated with the basis. This appendix computes the resulting null exactly.

Theorem C.1 (Random-direction null)**.** *Fix $h \neq 0$ and let R be Haar-uniform on $O(d)$, independent of h. Write $r_{\mathrm{eff}} = (\sum_i p_i^2)^{-1}$ for the participation-ratio effective rank of the squared spectrum. Then*

$$\mathop{\mathbb{E}}_{R}\big[S(W_R,h)\big] = \frac{1}{d}, \qquad \mathrm{Var}_R\big[S(W_R,h)\big] = \frac{2\,(d/r_{\mathrm{eff}}-1)}{d^2(d+2)}, \tag{20}$$

and consequently $\mathbb{E}_R[L(W_R,h)] \le \frac{1}{2}$, with equality if and only if $S(W_R,h)$ is almost surely constant, i.e. iff the spectrum is isotropic.

Proof. Reduction to the sphere. By Theorem 3.3 applied to W_R,

$$S(W_R,h) = \frac{\|U\Sigma R^\top h\|_2^2}{\|U\Sigma R^\top\|_F^2\,\|h\|_2^2} = \frac{\|\Sigma R^\top h\|_2^2}{\|\Sigma\|_F^2\,\|h\|_2^2},$$

since U and R are orthogonal and both norms are invariant under them. Put $u = R^\top h/\|h\|_2$. Because R is Haar-distributed and independent of h, and Haar measure is invariant under left multiplication by any fixed orthogonal matrix, u is uniformly distributed on S^{d-1}. Writing $\Sigma^\top\Sigma = \mathrm{diag}(\sigma_1^2,\dots,\sigma_d^2)$,

$$S(W_R,h) = \frac{\sum_i \sigma_i^2 u_i^2}{\sum_j \sigma_j^2} = \sum_{i=1}^{d} p_i\,u_i^2. \tag{21}$$

So the null depends on h only through the fact that u is uniform: it is the same for every fixed input.

Moments of u. Represent $u = g/\|g\|_2$ with $g \sim \mathcal{N}(0, I_d)$, which is uniform on the sphere. The vector $(u_1^2,\dots,u_d^2)$ is then $\mathrm{Dirichlet}(\frac{1}{2},\dots,\frac{1}{2})$, whose moments give

$$\mathbb{E}[u_i^2] = \frac{1}{d}, \quad \mathbb{E}[u_i^4] = \frac{3}{d(d+2)}, \quad \mathbb{E}[u_i^2u_j^2] = \frac{1}{d(d+2)}\ (i \neq j). \tag{22}$$

(Equivalently: $\mathbb{E}[u_i^2] = 1/d$ by symmetry and $\sum_i u_i^2 = 1$; and $\mathbb{E}[u_i^4], \mathbb{E}[u_i^2u_j^2]$ follow from $\mathbb{E}[g_i^4] = 3$, $\mathbb{E}[g_i^2g_j^2] = 1$ together with the independence of $g/\|g\|$ and $\|g\|$, which gives $\mathbb{E}[g_i^4] = \mathbb{E}[u_i^4]\ \mathbb{E}[\|g\|^4]$ and $\mathbb{E}[\|g\|^4] = d(d+2)$.)

Mean. Taking expectations in Equation (21), $\mathbb{E}[S] = \sum_i p_i\,\mathbb{E}[u_i^2] = \frac{1}{d}\sum_i p_i = \frac{1}{d}$.

Variance. Squaring Equation (21) and using Equation (22),

$$\begin{aligned}\mathbb{E}[S^2] &= \sum_i p_i^2\ \mathbb{E}[u_i^4] + \sum_{i\neq j} p_ip_j\ \mathbb{E}[u_i^2u_j^2] \\ &= \frac{3\sum_i p_i^2 + \big(1-\sum_i p_i^2\big)}{d(d+2)} = \frac{1+2\sum_i p_i^2}{d(d+2)},\end{aligned}$$

where we used $\sum_{i\neq j} p_i p_j = \big(\sum_i p_i\big)^2 - \sum_i p_i^2 = 1 - \sum_i p_i^2$. Subtracting $(\mathbb{E}\,S)^2 = 1/d^2$,

$$\mathrm{Var}(S) = \frac{1 + 2\sum_i p_i^2}{d(d+2)} - \frac{1}{d^2} = \frac{2\big(d\sum_i p_i^2 - 1\big)}{d^2(d+2)},$$

which is Equation (20) after substituting $\sum_i p_i^2 = 1/r_{\mathrm{eff}}$.

Jensen. $L = 1 + \frac{1}{2}\log_d S$ is a strictly concave function of S on $(0,\infty)$, so $\mathbb{E}[L] \le 1 + \frac{1}{2}\log_d \mathbb{E}[S] = 1 + \frac{1}{2}\log_d(1/d) = \frac{1}{2}$, with equality iff S is degenerate. By Equation (20), $\mathrm{Var}(S) = 0$ iff $\sum_i p_i^2 = 1/d$, which by Cauchy–Schwarz holds iff $p_i = 1/d$ for all i. □

The bound $\mathbb{E}[L_R] \le \frac{1}{2}$ is not tight, and the gap is computable.

Proposition C.2 (Exact null level)**.** *With $Q = \sum_i p_i g_i^2$ for $g \sim \mathcal{N}(0, I_d)$,*

$$\mathop{\mathbb{E}}_{R}[L(W_R, h)] = 1 + \frac{\mathbb{E}[\ln Q] - \ln 2 - \psi(d/2)}{2\ln d}, \tag{23}$$

where ψ is the digamma function and

$$\mathbb{E}[\ln Q] = \int_0^\infty t^{-1}\Big(e^{-t} - \prod_i (1 + 2p_i t)^{-1/2}\Big)\, dt. \tag{24}$$

To second order,

$$\mathop{\mathbb{E}}_{R}[L(W_R, h)] = \frac{1}{2} - \frac{d/r_{\mathrm{eff}} - 1}{2(d+2)\ln d} + O(\kappa_3), \tag{25}$$

where κ_3 is the third central moment of S; the deficit below $\frac{1}{2}$ is therefore controlled entirely by spectral anisotropy, and is $\approx 1/(2r_{\mathrm{eff}}\ln d)$ when $r_{\mathrm{eff}} \ll d$.

Proof. Separating scale from direction. Let $g \sim \mathcal{N}(0, I_d)$ and $u = g/\|g\|_2$. The distribution of g is rotation invariant, so u is uniform on S^{d-1} and is *independent* of $\|g\|_2^2 \sim \chi_d^2$. Since $Q = \sum_i p_i g_i^2 = \|g\|_2^2 \sum_i p_i u_i^2 = \|g\|_2^2\, S$ with S as in Equation (21), taking logarithms and expectations of a product of independent positive variables gives

$$\mathbb{E}[\ln Q] = \mathbb{E}[\ln S] + \mathbb{E}[\ln \chi_d^2] = \mathbb{E}[\ln S] + \ln 2 + \psi(d/2),$$

using the standard identity $\mathbb{E}[\ln \chi_d^2] = \ln 2 + \psi(d/2)$. Solving for $\mathbb{E}[\ln S]$ and substituting into $\mathbb{E}[L] = 1 + \frac{1}{2\ln d}\,\mathbb{E}[\ln S]$ yields Equation (23).

The integral. For $x > 0$ the Frullani representation $\ln x = \int_0^\infty t^{-1}(e^{-t} - e^{-xt})\, dt$ holds. Applying it to $x = Q$ and taking expectations,

$$\mathbb{E}[\ln Q] = \mathbb{E}\int_0^\infty \frac{e^{-t} - e^{-Qt}}{t}\, dt = \int_0^\infty \frac{e^{-t} - \mathbb{E}[e^{-Qt}]}{t}\, dt,$$

where the exchange is justified by Fubini–Tonelli: the integrand is absolutely integrable because $|e^{-t} - e^{-Qt}|/t$ is bounded by $\max(1, Q)$ near $t = 0$ and by $(e^{-t} + e^{-Qt})/t$ at infinity, and $\mathbb{E}[\max(1, Q)] < \infty$. Since the g_i are independent and $\mathbb{E}[e^{-tp_i g_i^2}] = (1 + 2p_i t)^{-1/2}$ for $t \ge 0$, the Laplace transform factorizes as $\mathbb{E}[e^{-Qt}] = \prod_i (1 + 2p_i t)^{-1/2}$, giving Equation (24).

Second-order form. Expanding $\ln S$ about $\mathbb{E}[S] = 1/d$,

$$\ln S = \ln \mathbb{E}[S] + \frac{S - \mathbb{E}\,S}{\mathbb{E}\,S} - \frac{(S - \mathbb{E}\,S)^2}{2(\mathbb{E}\,S)^2} + O\big((S - \mathbb{E}\,S)^3\big),$$

and taking expectations kills the linear term, leaving $\mathbb{E}[\ln S] = -\ln d - \frac{1}{2}\,\mathrm{Var}(S)/(\mathbb{E}\,S)^2 + O(\kappa_3)$. Substituting $\mathrm{Var}(S)$ from Equation (20) and $(\mathbb{E}\,S)^2 = 1/d^2$,

$$\frac{\mathrm{Var}(S)}{2(\mathbb{E}\,S)^2} = \frac{d^2}{2} \cdot \frac{2(d/r_{\mathrm{eff}} - 1)}{d^2(d+2)} = \frac{d/r_{\mathrm{eff}} - 1}{d+2},$$

and dividing by $2\ln d$ as required by $\mathbb{E}[L] = 1 + \mathbb{E}[\ln S]/(2\ln d)$ gives Equation (25). For $r_{\mathrm{eff}} \ll d$ the numerator is dominated by d/r_{eff} and the denominator by d, leaving $\approx 1/(2r_{\mathrm{eff}}\ln d)$. □

Numerical verification. Equation (23) agrees with Monte-Carlo estimates of $\mathbb{E}_R[L]$ to within 2×10^{-4} across isotropic, power-law ($p_i \propto i^{-1}, i^{-2}$) and exponential spectra at $d = 64$ and $d = 512$. The second-order form Equation (25) is accurate to 10^{-3} except under extreme anisotropy ($r_{\text{eff}} \approx 2.5$), where it overstates the deficit by about 0.01. For the layers probed in this paper r_{eff} ranges from 13 to 579, placing every null within 5.5×10^{-3} of $\frac{1}{2}$; the control may therefore be read against a null of $\frac{1}{2}$, with Equation (23) available when the correction matters.

D Differential Geometry of Alignment

This appendix states and proves Theorem D.1 and derives the gradients of the two penalties used in Section 4.4. Throughout W is fixed, h satisfies $Wh \neq 0$, and gradients are taken with respect to h.

Theorem D.1 (Alignment gradients are directional)**.** *For $Wh \neq 0$,*

$$\nabla_h L(W,h) = \frac{1}{\ln d}\left(\frac{W^\top W h}{\|Wh\|_2^2} - \frac{h}{\|h\|_2^2}\right), \quad h^\top \nabla_h L = 0. \tag{26}$$

Right-singular vectors are stationary points; the maximum is attained at a top right-singular direction. Likewise, for the subspace indexed by K, with projector $P_K = \sum_{i\in K} v_i v_i^\top$, the occupancy regularizer $Q_K(h) = \sum_{i\in K} q_i(h) = \|P_K h\|_2^2/\|h\|_2^2$ has $\nabla_h Q_K = \frac{2}{\|h\|^2}(P_K h - Q_K h)$ and $h^\top \nabla_h Q_K = 0$.

Proof of Theorem D.1. The gradient of L. Converting the base of the logarithm in Equation (5),

$$L(W,h) = 1 + \frac{1}{\ln d}\Big(\ln \|Wh\|_2 - \ln \|W\|_F - \ln \|h\|_2\Big), \tag{27}$$

in which only the first and third terms depend on h. For the first, the chain rule applied to $\ln \sqrt{h^\top W^\top W h}$ gives

$$\nabla_h \ln \|Wh\|_2 = \frac{1}{2\|Wh\|_2^2}\,\nabla_h\big(h^\top W^\top W h\big) = \frac{W^\top W h}{\|Wh\|_2^2},$$

using $\nabla_h(h^\top A h) = 2Ah$ for symmetric $A = W^\top W$. The same computation with $A = I$ gives $\nabla_h \ln \|h\|_2 = h/\|h\|_2^2$. Substituting both into Equation (27) yields Equation (26).

Orthogonality to h. Contracting Equation (26) with h,

$$h^\top \nabla_h L = \frac{1}{\ln d}\left(\frac{h^\top W^\top W h}{\|Wh\|_2^2} - \frac{h^\top h}{\|h\|_2^2}\right) = \frac{1}{\ln d}\big(1-1\big) = 0,$$

since $h^\top W^\top W h = \|Wh\|_2^2$ and $h^\top h = \|h\|_2^2$. Thus $\nabla_h L$ is tangent to the sphere through h: to first order, gradient steps on L rotate h and leave $\|h\|_2$ unchanged. This is the exact counterpart of the scale invariance $L(W, ch) = L(W,h)$ of Proposition A.1, of which it is the infinitesimal form: the directional derivative of L along the radial direction h vanishes.

Stationary points. Suppose $h = v_i$ with $\sigma_i > 0$. Then $W^\top W h = \sigma_i^2 h$ and $\|Wh\|_2^2 = \sigma_i^2 \|h\|_2^2$, so the two terms of Equation (26) coincide and $\nabla_h L = 0$: every right-singular direction with nonzero singular value is a critical point. By Proposition A.2, L attains its maximum on its domain $Wh \neq 0$ exactly on the top singular subspace. If W has full column rank, its minimum is attained on the smallest singular subspace. If W has a nontrivial nullspace, L is unbounded below as a unit activation approaches a nullspace direction through points with $Wh \neq 0$, so no minimum is attained on the full domain. In that case, the smallest positive singular subspace minimizes L only when h is restricted to $\ker(W)^\perp \setminus \{0\}$. The critical values are ordered by their singular values.

The subspace regularizer. Let $K \subseteq \{1, \ldots, d\}$ and $P_K = \sum_{i\in K} v_i v_i^\top$, an orthogonal projector. By orthonormality of $\{v_i\}$,

$$Q_K(h) = \sum_{i\in K} q_i(h) = \frac{\sum_{i\in K}(v_i^\top h)^2}{\|h\|_2^2} = \frac{\|P_K h\|_2^2}{\|h\|_2^2} = \frac{h^\top P_K h}{h^\top h}, \tag{28}$$

using $P_K^\top P_K = P_K$. So Q_K is exactly the fraction of activation energy lying in the chosen subspace. Differentiating the Rayleigh quotient Equation (28) by the quotient rule,

$$\nabla_h Q_K = \frac{2P_K h\,(h^\top h) - 2h\,(h^\top P_K h)}{(h^\top h)^2} = \frac{2}{\|h\|_2^2}\big(P_K h - Q_K(h)\,h\big),$$

and contracting with h,

$$h^\top \nabla_h Q_K = \frac{2}{\|h\|_2^2}\big(h^\top P_K h - Q_K(h)\|h\|_2^2\big) = 0$$

by the definition of Q_K. So Q_K is scale invariant and its gradient is also tangent to the sphere. □

For fixed weights, direct gradient steps rotate h without changing its norm to first order; parameter updates need not preserve the norm.

What the penalties do. The target penalty $R_L = (L - L^\star)^2$ has

$$\nabla_h R_L = \frac{2(L - L^\star)}{\ln d}\left(\frac{W^\top W h}{\|Wh\|_2^2} - \frac{h}{\|h\|_2^2}\right), \tag{29}$$

so a gradient step on R_L moves h along $-\nabla_h L$ when $L > L^\star$ and along $+\nabla_h L$ when $L < L^\star$: the penalty rotates activation occupancy toward higher-gain learned directions when alignment is below target and away from them when above, in both cases without first-order change in $\|h\|_2$. The occupancy penalty $R_q = -Q_K$ has

$$\nabla_h R_q = -\frac{2}{\|h\|_2^2}\big(P_K h - Q_K(h)h\big), \tag{30}$$

whose descent direction $P_K h - Q_K h$ is the component of $P_K h$ orthogonal to h; it therefore rotates h toward the subspace K and vanishes exactly when $h \in K$ or $h \perp K$.

Two consequences are used in Section 5.4. First, R_q does not mechanically reduce the rank of the representation: it applies rotational pressure toward K and, if $K = \{1, \ldots, d\}$, then $P_K = I, Q_K \equiv 1$ and the gradient vanishes identically, so the penalty is inert rather than maximally constraining. Second, because Q_K depends on K only through the subspace it spans, two penalties with $|K|$ equal but different direction identity—the learned top-k singular directions, a random k-dimensional subspace, or the bottom k—impose the same rank pressure and differ only in *which* directions they favor. Comparing them therefore isolates direction identity from dimension, which is what the sweep in Appendix G.11 does.

E Profile Geometry and Calibration

This appendix states and proves the basis invariance used in Section 4.3 and gives the distributional reading of the score Equation (9).

Three properties motivate the choice of Mahalanobis distance. It is *two-sided*: shift can raise alignment as well as lower it (Appendix G.2), and a distance flags both while leaving the profile shape for diagnosis. It is *coordinate-invariant*: an invertible affine change of profile coordinates, with statistics re-estimated, leaves the distance unchanged (Proposition E.1). This changes how the same measurements are expressed; randomizing a weight matrix's learned directions changes the measurements themselves. And it *scales*: per-class means with one pooled covariance remain estimable at 1000 classes, and under a shared-covariance Gaussian model the score is twice a plug-in negative log-likelihood at the predicted class, up to an additive constant (Proposition E.3); Appendix G.4 compares the calibration variants.

Proposition E.1 (Basis invariance)**.** *Let a profile X have mean μ and covariance Σ estimated from calibration data, and let T be any invertible affine map. If the statistics of TX are estimated the same way, then $(TX - \mu_{TX})^\top \Sigma_{TX}^{-1}(TX - \mu_{TX}) = (X - \mu)^\top \Sigma^{-1}(X - \mu)$ for every input.*

Lemma E.2 (Equivariance of the plug-in statistics)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^k$ be calibration profiles with sample mean $\hat{\mu}$ and sample covariance $\hat{\Sigma}$, and let $Y_i = TX_i + c$ for an invertible $T \in \mathbb{R}^{k\times k}$ and $c \in \mathbb{R}^k$. Then the statistics computed the same way from $Y_1, \ldots, Y_n$ satisfy $\hat{\mu}_Y = T\hat{\mu} + c$ and $\hat{\Sigma}_Y = T\hat{\Sigma}T^\top$.*

Proof. By linearity of the average, $\hat{\mu}_Y = \frac{1}{n}\sum_i (TX_i + c) = T\hat{\mu} + c$. Then $Y_i - \hat{\mu}_Y = T(X_i - \hat{\mu})$, so

$$\hat{\Sigma}_Y = \frac{1}{n-1}\sum_i T(X_i - \hat{\mu})(X_i - \hat{\mu})^\top T^\top = T\hat{\Sigma}T^\top,$$

and $\hat{\Sigma}_Y$ is invertible whenever $\hat{\Sigma}$ is, since T is. The same argument applies verbatim to class-conditional means and to a covariance pooled over within-class residuals, because both are averages of the same centered outer products. □

Proof of Proposition E.1. Invariance. Let X be a profile with statistics $\hat{\mu}, \hat{\Sigma}$ and let $Y = TX + c$ with T invertible. Using Lemma E.2 and $(T\hat{\Sigma}T^\top)^{-1} = T^{-\top}\hat{\Sigma}^{-1}T^{-1}$,

$$\begin{aligned}(Y - \hat{\mu}_Y)^\top \hat{\Sigma}_Y^{-1}(Y - \hat{\mu}_Y) &= \big[T(X - \hat{\mu})\big]^\top T^{-\top}\hat{\Sigma}^{-1}T^{-1}\big[T(X - \hat{\mu})\big] \\ &= (X - \hat{\mu})^\top T^\top T^{-\top}\hat{\Sigma}^{-1}T^{-1}T(X - \hat{\mu}) \\ &= (X - \hat{\mu})^\top \hat{\Sigma}^{-1}(X - \hat{\mu}),\end{aligned}$$

which is the claimed identity. This uses only the fact that the statistics are estimated from the transformed data in the same way; the score is *not* invariant if $\hat{\mu}, \hat{\Sigma}$ are held fixed while the data are transformed.

The map of Equation (7). Per layer,

$$\begin{pmatrix} g_\ell \\ r_\ell \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}}_{T_1} \begin{pmatrix} a_\ell \\ b_\ell \end{pmatrix}, \qquad \det T_1 = 1,$$

with inverse $T_1^{-1} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Stacking over the P probed depths gives the block-diagonal map $T = \mathrm{diag}(T_1, \ldots, T_1) \in \mathbb{R}^{2P \times 2P}$, which is invertible with $\det T = 1$. Applying the first part with this T and $c = 0$ shows that the raw log-norm profile $[a_1, b_1, \ldots, a_P, b_P]$ and the profile $[g_1, r_1, \ldots, g_P, r_P]$ produce identical Mahalanobis scores.

Standardization. The per-coordinate standardization $z_j = (\Phi_j - m_j)/s_j$ preceding Equation (9) is itself an invertible affine map with diagonal T, so it too leaves the score unchanged; it is applied for numerical conditioning of $\hat{\Sigma}$, not because the score depends on it. □

Proposition E.3 (Gaussian plug-in reading)**.** *Suppose the calibrated profile satisfies $Z \mid y = c \sim \mathcal{N}(\mu_c, \Sigma)$ with a covariance shared across classes. Then for each class,*

$$-\log p(Z \mid y = c) = \tfrac{1}{2}(Z - \mu_c)^\top \Sigma^{-1}(Z - \mu_c) + \textit{const},$$

with the constant $\frac{1}{2}\log\big((2\pi)^k \det\Sigma\big)$ independent of c and of the input. Hence Equation (9) *is twice the plug-in negative log-likelihood at the predicted class, and if the model holds with the correct class known, $U_{\mathrm{LAP}} \sim \chi^2_k$ where $k = 2P$.*

Proof. The density of $\mathcal{N}(\mu_c, \Sigma)$ is $p(Z \mid y = c) = (2\pi)^{-k/2}(\det\Sigma)^{-1/2}\exp\big(-\frac{1}{2}(Z - \mu_c)^\top\Sigma^{-1}(Z - \mu_c)\big)$; taking negative logarithms gives the first claim, and the additive constant does not depend on c because Σ is shared. For the second, write $\Sigma^{-1} = A^\top A$ (e.g. $A = \Sigma^{-1/2}$); then under the model $A(Z - \mu_c) \sim \mathcal{N}(0, I_k)$ and $U_{\mathrm{LAP}} = \|A(Z - \mu_c)\|_2^2$ is a sum of k independent squared standard normals, i.e. χ^2_k. □

The shared-covariance Gaussian model motivates class-specific means and a pooled covariance: classes differ in location while sharing a scale, keeping the parameter count manageable as the number of classes grows (Section 4.3). The Gaussian reading above uses the true class. In Equation (9), we evaluate the plug-in statistic at the predicted class $\hat{y}$. We set alarm thresholds using empirical ID quantiles to account for departures from the Gaussian model.

F Efficient Convolutional Form

Setup. Consider a convolution with weight tensor $\mathcal{W}$ of shape $(m, c_{\text{in}}, k_1, k_2)$ acting on an input feature map x. Let $W \in \mathbb{R}^{m \times d}$, $d = c_{\text{in}} k_1 k_2$, be its matrix form, and for each output location $l \in \{1, \dots, n\}$ let $x_l \in \mathbb{R}^d$ be the corresponding receptive-field patch, so that the layer computes $y_l = W x_l$ (plus a bias, treated below). Collecting the patches as columns of $X = [x_1, \dots, x_n]$ gives the im2col identity $Y = WX$ with $Y = [y_1, \dots, y_n]$.

The sample-level quantity. Applying Equation (5) to the whole set of patches produced by one input, the overlap is

$$S_{\text{conv}} = \frac{\|WX\|_F^2}{\|W\|_F^2 \, \|X\|_F^2} = \frac{\sum_l \|W x_l\|_2^2}{\|W\|_F^2 \sum_l \|x_l\|_2^2}, \tag{31}$$

which is the $\|x_l\|_2^2$-weighted average of the per-patch overlaps, in the same sense as Definition 3.1. Both sums can be obtained without forming X.

Proposition F.1 (No patch matrix is needed)**.** *Let y be the layer's output feature map computed with zero bias. Then*

$$\sum_l \|W x_l\|_2^2 = \|y\|_F^2, \qquad \sum_l \|x_l\|_2^2 = \big\| \operatorname{conv}(x^{\odot 2}; \mathbf{1}) \big\|_1, \tag{32}$$

where $x^{\odot 2}$ is the elementwise square of the input and $\operatorname{conv}(\cdot; \mathbf{1})$ is a convolution with an all-ones kernel of the same spatial size, stride, padding and dilation as the layer, summed over its $c_{\text{in}} k_1 k_2$ taps. For a layer with bias b, replace y by $y - b$ (broadcast over spatial positions).

Proof. The first identity is immediate: the output at location l is exactly $y_l = W x_l$, so $\sum_l \|W x_l\|_2^2 = \sum_l \|y_l\|_2^2 = \|y\|_F^2$, the sum of squares of all entries of the output map.

For the second, write the patch extraction explicitly. If $\pi(l, j)$ denotes the input index contributing to the j-th coordinate of patch l (with $x_{l,j} = 0$ for taps falling in the padded region), then

$$\sum_l \|x_l\|_2^2 = \sum_l \sum_{j=1}^{d} x_{\pi(l,j)}^2,$$

i.e. each input entry is counted once for every patch in which it participates, with multiplicity given by the layer's overlap pattern. A convolution of $x^{\odot 2}$ with an all-ones kernel of identical geometry computes exactly $\sum_j x_{\pi(l,j)}^2$ at each output location l; summing those values over l gives the double sum above. Padding is handled consistently because zero-padded taps contribute 0 to both expressions, and dilation and stride only relabel π.

For a biased layer the output is $y_l = W x_l + b$, so $W x_l = y_l - b$ and the first identity holds with y replaced by $y - b$; the second is unaffected since it involves only the input. □

Cost. Each probed layer therefore requires one squaring of the input, one all-ones convolution (with the same access pattern as the layer itself, and no learned parameters), and two reductions. No $d \times n$ patch matrix is materialized, and no SVD is computed at test time: by Theorem 3.3 the singular basis never appears in the norm-ratio form. The measured overhead is reported in Appendix J; it is under 10% of a forward pass at batch 200 for both architectures studied, including the per-sample bookkeeping that a deployment would also pay.

G Additional Results and Experimental Details

Vision models and comparisons. Our vision experiments include 30 ResNet-18 models trained on CIFAR-10 subsets of $10{,}000$ images, with ten seeds at each label-noise rate of 0, 0.2, and 0.4. We repeat experiments with CIFAR-100, TinyImageNet, and ViT-Tiny. At larger scale, we evaluate five ResNet-50 models trained from scratch on ImageNet-200 (Zhang et al., 2024a) and pretrained ResNet-50 and ViT-B/16 models on ImageNet-1K. Interior probes span each network's depth. Feature-norm and globally average-pooled feature Mahalanobis profiles use matched depths and identical calibration to test whether depth coverage alone accounts for the profile's performance.

Probe coverage. The probe count P ranges from 6 to 48 across our testbeds, with every block probed in the language models. The architecture-specific setups below and Appendix I specify the probe placements.

G.1 Depth coupling and training dynamics

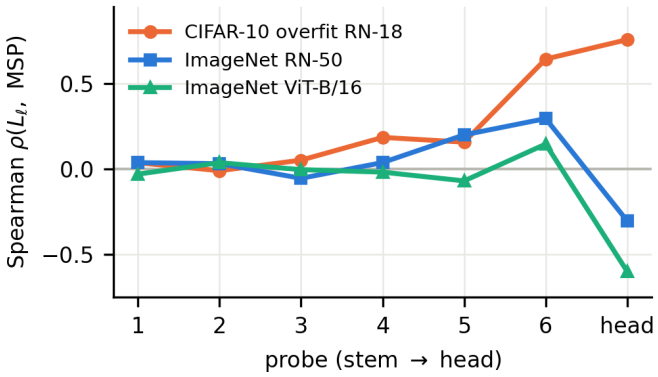


Figure 4: Interior alignment is near-orthogonal to confidence in every regime; only the head couples, with a sign that flips between regimes. Spearman correlation between per-layer alignment and MSP on ID data, per probe, for the overfit CIFAR-10 ResNet-18 models (mean over ten seeds) and the pretrained ImageNet-1K ResNet-50 and ViT-B/16.

Depth coupling. Interior-layer alignment is close to orthogonal to output confidence across four training regimes, two architecture families, and models from 10k-image CIFAR subsets to pretrained ImageNet-1K (Figure 4). On CIFAR-10 ResNet-18 models the Spearman correlation between per-layer alignment and MSP has magnitude at or below 0.19 at five of six interior sensors and reaches 0.76 at the head; from ViT-Tiny to pretrained ImageNet-1K models $|\rho| \leq 0.30$ at every interior probe, while the head coupling inverts in sign relative to the small overfit regime (-0.30/-0.60 vs. $+0.76$), as Equation (14) allows. Head LAR is therefore not a stable diagnostic, and both design choices of Section 4.1 follow.

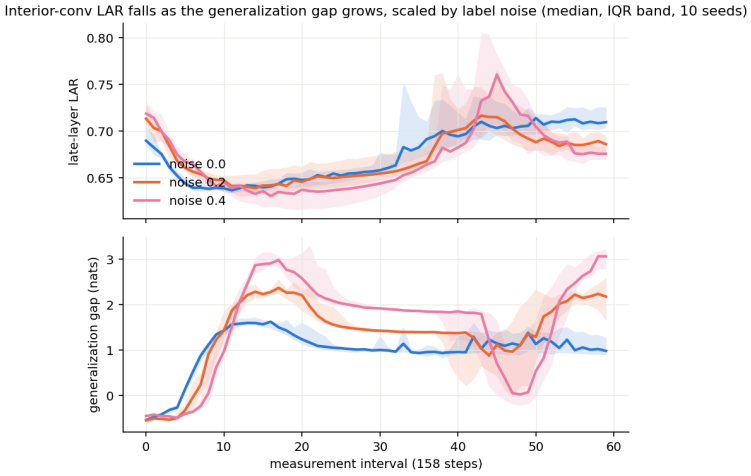


Figure 5: Late-layer alignment (top) and the generalization gap (bottom) over training; median and IQR over ten seeds per noise level. Alignment falls while the gap opens, and the decline deepens with label noise.

Training dynamics. This is the setting behind the training-time relationship summarized in Section 4.1. Thirty ResNet-18 runs (CIFAR-10 10k subsets, label noise $\{0, 0.2, 0.4\} \times 10$ seeds) are measured at 60 intervals of 158 steps each. At every interval the monitor evaluates alignment on a

fixed probe pool of 2,048 training images drawn once at run start, at two sensors: the classifier input ("head") and the late convolution `layer4.1.conv2`.

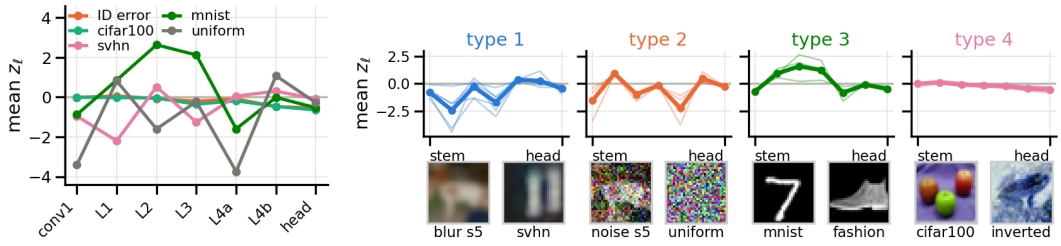
Loss-stratified gap. At each interval the per-example training loss is computed over the pool, and alignment is measured separately on the lowest and highest loss quartiles. The stratified gap is

$$\Delta L_t \;=\; L_t^{\text{high loss}} - L_t^{\text{low loss}},$$

i.e. how much less aligned the examples the network is still struggling with are, relative to the examples it has fitted. *Generalization gap.* $G_t = \mathcal{L}_{\text{val},t} - \mathcal{L}_{\text{train},t}$ in nats (the logs carry losses, not accuracies). The quoted values ($-0.34, -0.53, -0.66$ at noise $0, 0.2, 0.4$) are within-run Spearman correlations of ΔL_t against G_t at the late sensor, after dropping a three-interval burn-in, reported as mean over the ten seeds of each noise level; per-run values are in the supplementary CSVs. These within-run correlations describe co-movement of two trending series. The interventions in Section 5.4 test the effect of changing alignment on learning.

The training-trajectory analysis uses the CIFAR-10 runs, which record sixty checkpoints per run with the stratified monitor attached. For the ImageNet-scale models, we evaluate depth coupling at a single fixed checkpoint (Figure 4).

G.2 Localizing atypicality across depth: profiles and archetypes



(a) (b)

Figure 6: **Mean alignment profiles reveal distinct depth responses to distribution shift.** (a) Mean calibrated profiles per set (CIFAR-10 ID, noise-0.2 ensemble): SVHN shows a first-block deficit, MNIST a mid-network overshoot, novelty and ID errors a small late deficit. (b) Four archetypes organized by the depth of the deviation, not by dataset provenance (thin: members, thick: centroids, images: example members).

Different shifts produce distinct mean depth profiles (Figure 6a). A deficit denotes below-reference alignment; hyper-alignment denotes above-reference alignment. On CIFAR-10-ID models, CIFAR-100 novelty remains close to the reference at early layers and develops a late deficit. Misclassified ID images follow a similar mean direction at smaller amplitude (cosine similarity 0.97). SVHN produces a sharp first-block deficit that recovers toward an ID-typical head, while MNIST produces an intermediate-layer overshoot.

Clustering the 24 per-set mean profiles yields four archetypes: early deficit, early hyper-alignment, intermediate-layer hyper-alignment and late deficit (Figure 6b). Corresponding patterns also appear with CIFAR-100 as the reference and on ImageNet-1K classifiers. The clusters group shifts with similar depth responses: for example, SVHN groups with blur and contrast reduction rather than with the other digit datasets. The profile distinguishes early, intermediate and late departures that an image-level anomaly score compresses into a single number.

We describe the members and profile shapes of types 1–4 below.

Type 1: smoothing (blur s3/s5, contrast s3/s5, JPEG q10, SVHN). Deep occupancy deficit at the first block (-4.4 at L1 for the strongest members), recovering toward an ID-typical head. Every member removes high-frequency content, leaving a low-frequency-dominated input; early layers occupy edge-tuned directions, and an input without edges cannot fill them. SVHN belongs here for the same reason—its smooth, flat-background statistics—despite being a digit dataset by provenance.

Type 2: high-frequency injection (Gaussian noise s3/s5, impulse noise s3/s5, uniform). The mirror image of type 1 at the first block: L1 *hyper*-aligns ($+1.0$ vs. type 1's -4.4) because injected noise

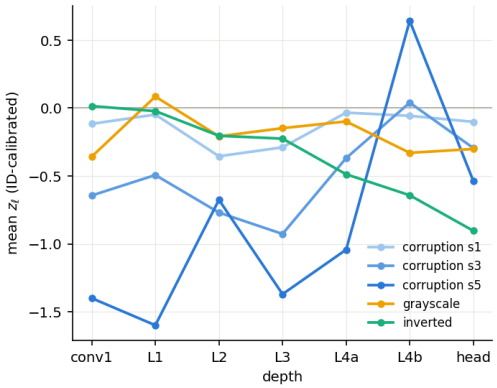


Figure 7: Depth signatures under shifts that alter image statistics but not content (CIFAR-10 ID, noise-0.2 ensemble; companion to Figure 6a). *Corruption sN* is the mean profile over four corruptions (Gaussian noise, impulse noise, Gaussian blur, contrast reduction) at severity N of a five-level ramp. Severity deepens an early deficit; grayscale is mild; inversion deviates across depth.

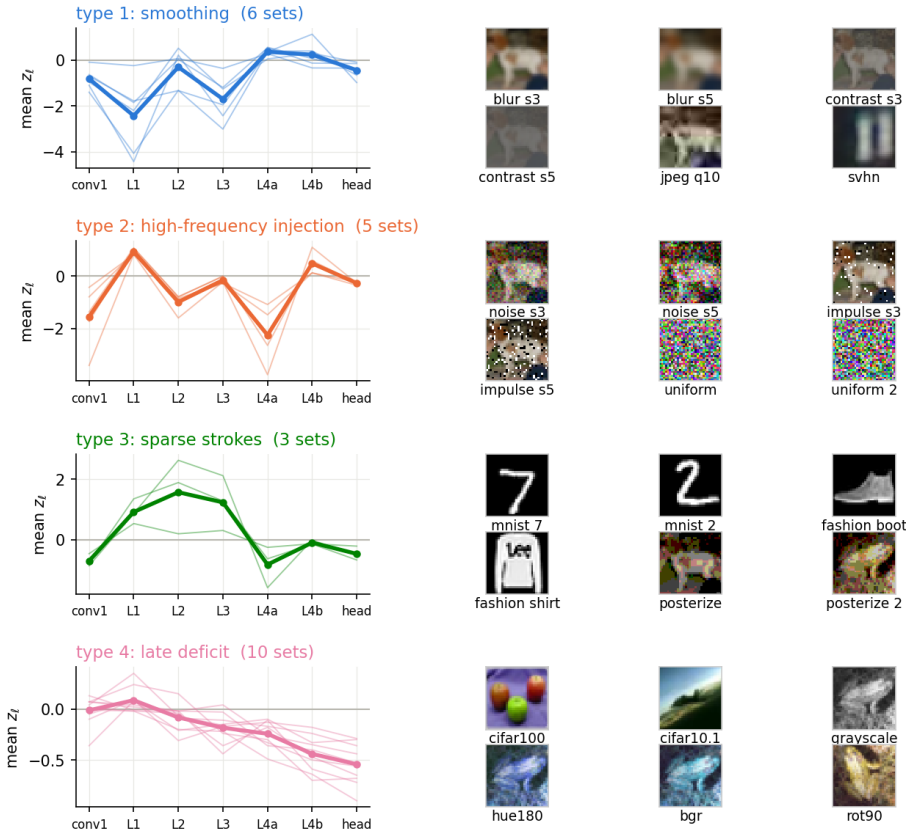


Figure 8: Extended version of Figure 6b: types 1–4 with all member profiles (thin lines; centroid thick; count per cluster in each title) and six example inputs per archetype, constructed exactly as the evaluation sets are. Axes are scaled per row—type 4 is an order of magnitude gentler than types 1–3, which is why it is the shape shared with in-distribution errors rather than a gross corruption signature.

saturates the edge-tuned directions, and the deficit instead appears at L4a, where ID inputs have settled into object-level structure that noise cannot supply.

Type 3: sparse strokes (MNIST, Fashion-MNIST, posterize-2). Mid-network hyper-alignment (L2–L3): sparse images with flat regions and sharp strokes concentrate activation energy in few mid-level shape directions, more than natural images ever do.

Type 4: late deficit (all semantic and appearance sets: CIFAR-100 near/far, CIFAR-10.1, grayscale, inversion, hue rotations, BGR, rot90). Early layers ID-typical, then a monotone decline into the late blocks. These inputs have natural low-level statistics, so early processing proceeds normally; the deviation appears where representations become semantic. This is the family whose direction ID errors share, and its amplitude is an order of magnitude smaller than types 1–3 (note the per-row axes in Figure 8).

Benchmark-construction labels, by contrast, do not predict profile shape: in a pre-registered assignment test, held-out SVHN—a digit set by provenance—lands in the statistical-corruption family by profile shape, not with the other digit sets. The profile records where the network stopped processing an input like ID data, which cuts across how the datasets were built. Figure 8 shows all member profiles and six example inputs per archetype; the per-set curves behind the statistical clusters are in Figure 7.

G.3 Evaluation sets, ablations and baseline tables

Evaluation sets. CIFAR-10/100 taxonomy (24 sets): Gaussian and impulse noise, Gaussian blur and contrast at severities 1–5 (pooled as s1/s3/s5), grayscale, inversion, hue rotations $90°$/$180°$, BGR channel swap, rot90, posterize-2, JPEG q10, CIFAR-10.1 (Recht et al., 2018), CIFAR-100 near/far superclass splits, SVHN, MNIST, Fashion-MNIST, and uniform noise. ImageNet-1K: official ImageNet-C cells (Hendrycks & Dietterich, 2019) (contrast s3/s5, defocus, motion blur, pixelate, JPEG at s5), synthetic Gaussian noise/blur/contrast, ImageNet-R (Hendrycks et al., 2021), NINCO (Bitterwolf et al., 2023), iNaturalist, SUN, Places (Huang & Li, 2021), and Textures (Cimpoi et al., 2014). ImageNet-200 uses the OpenOOD class split with ImageNet-C corruptions on identical images. Language shifts are listed in Appendix G.12, robotics conditions in Appendix I.

Output-score baselines. For an input x, let $u_c(x)$ be the classifier's logit for class c among C classes, and let $\pi_c(x) = \exp(u_c(x))/\sum_{a=1}^{C}\exp(u_a(x))$ be its softmax probability. The four output scores in Section 5.1 are

$$\mathrm{MSP}(x) = \max_c \pi_c(x), \qquad H(x) = -\sum_{c=1}^{C}\pi_c(x)\log\pi_c(x),$$

$$E(x) = -\log\sum_{c=1}^{C}\exp(u_c(x)), \qquad M(x) = \max_c u_c(x).$$

Maximum softmax probability (MSP) retains the probability of the predicted class, whereas entropy H measures how broadly probability is spread across classes. Energy E aggregates the unnormalized logits through a negative log-sum-exp; maximum logit M retains only the largest logit. We use natural logarithms and temperature 1 for both softmax and energy.

For shift-detection AUROC, larger scores indicate more anomalous inputs: we reverse the direction of MSP and maximum logit, while using entropy and energy as written. For ensemble MSP and entropy, we first average the members' softmax probability vectors and then apply the definitions above to that mean vector.

The ablations in Table 5 separate the contributions of gain and activation scale. Gain-only and radius-only profiles measure how much each coordinate contributes to detection. A second control replaces each layer's right-singular measurement basis with a Haar-random orthogonal basis (randomized V), preserving the input activations and weight spectrum while changing their directional correspondence. The network and its predictions remain fixed. Appendix C derives the random-direction null.

Randomizing V weakens detection across the suite, dramatically on ViT-B/16 (Gaussian noise $0.971 \to 0.562$, defocus $0.990 \to 0.822$) and on ResNet-50 by 0.05 (noise) to 0.28 (contrast s3). Learned directional correspondence contributes information beyond spectrum and norms.

Gram-matrix baseline. We reimplement Sastry & Oore (2020) following the official recipe: at the stem and every block output (17 layers for ResNet-50, 13 for ViT-B/16 with tokens as positions) and every power $p = 1, \ldots, 10$, the row sums of the Gram matrix of the elementwise p-th power of the feature map, with a sign-preserving p-th root ($151{,}680$ and $99{,}840$ numbers per sample); per-entry minima and maxima over the calibration split, conditioned on the predicted class (*class-cond.*, the

Score	con. s3	con. s5	noise	C-con.	C-def.	DTD	NINCO	iNat	IN-R
ResNet-50									
gain (LAR) only	0.569	0.737	0.980	0.983	0.991	0.915	0.736	0.810	0.730
radius only	0.840	0.958	0.963	0.992	0.963	0.915	0.731	0.847	0.743
gain + radius ($= U_{\mathrm{LAP}}$)	0.822	**0.983**	**1.000**	**1.000**	**0.998**	0.938	0.758	0.872	0.761
randomized V	0.542	0.717	0.952	0.900	0.906	0.852	0.693	0.806	0.661
Gram, class-cond. (Sastry & Oore, 2020)	**0.846**	0.951	0.970	0.976	0.970	0.891	0.732	0.750	0.770
Gram, global min/max	0.550	0.958	0.971	0.995	0.990	0.884	0.556	0.685	0.606
7-layer feat. Maha	0.482	0.722	0.998	0.859	0.957	0.921	0.610	0.736	0.680
RMDS	0.490	0.508	0.898	0.889	0.958	0.853	**0.858**	0.931	0.852
ReAct	0.491	0.474	0.148	0.156	0.136	0.425	0.292	0.311	0.162
ASH-S	0.496	0.475	0.291	0.426	0.474	0.595	0.441	0.516	0.313
MSP	0.502	0.502	0.773	0.824	0.907	0.749	0.768	0.845	0.779
entropy	0.503	0.498	0.711	0.784	0.875	0.676	0.703	0.775	0.745
energy	0.501	0.490	0.497	0.617	0.636	0.485	0.548	0.508	0.592
max-logit	0.504	0.499	0.729	0.794	0.884	0.692	0.729	0.799	0.751
feat. norm	0.500	0.492	0.206	0.259	0.121	0.186	0.352	0.175	0.214
penult. feat. Maha	0.486	0.504	0.914	0.919	0.986	0.925	0.824	**0.951**	**0.864**
KNN	0.497	0.499	0.730	0.797	0.903	0.878	0.594	0.724	0.672
ViM	0.486	0.500	0.878	0.899	0.984	**0.945**	0.741	0.917	0.857
ViT-B/16									
gain (LAR) only	0.718	0.984	0.984	**1.000**	**0.995**	0.910	0.767	0.896	0.790
radius only	0.858	0.975	0.922	0.997	0.974	0.896	0.762	0.872	0.743
gain + radius ($= U_{\mathrm{LAP}}$)	**0.897**	**0.998**	0.971	**1.000**	0.990	0.943	0.779	0.887	0.791
randomized V	0.578	0.827	0.562	0.924	0.822	0.792	0.630	0.631	0.627
Gram, class-cond. (Sastry & Oore, 2020)	0.688	0.919	0.750	0.964	0.914	0.780	0.634	0.605	0.714
Gram, global min/max	0.646	0.988	0.669	0.999	0.975	0.868	0.579	0.552	0.620
7-layer feat. Maha	0.560	0.972	**0.986**	0.997	0.974	**0.948**	0.697	0.779	0.699
RMDS	0.509	0.700	0.755	0.962	0.920	0.878	**0.870**	0.960	0.846
ReAct	0.576	0.929	0.763	0.997	0.838	0.848	0.758	0.862	0.812
ASH-S	0.452	0.176	0.525	0.049	0.563	0.479	0.527	0.499	0.514
MSP	0.550	0.863	0.756	0.997	0.837	0.830	0.782	0.882	0.776
entropy	0.567	0.921	0.789	0.999	0.845	0.845	0.780	0.898	0.785
energy	0.587	0.940	0.775	0.999	0.767	0.793	0.663	0.794	0.728
max-logit	0.576	0.917	0.776	0.998	0.806	0.817	0.725	0.853	0.751
feat. norm	0.489	0.379	0.370	0.348	0.139	0.176	0.213	0.138	0.153
penult. feat. Maha	0.509	0.686	0.733	0.924	0.932	0.877	0.865	0.959	0.864
KNN	0.505	0.621	0.585	0.831	0.743	0.674	0.508	0.631	0.678
ViM	0.519	0.837	0.758	0.980	0.954	0.884	0.857	**0.965**	**0.868**

Table 5: Ablation and depth-resolved baselines on ImageNet-1K (AUROC; extends Table 1 to the semantic sets and the single-coordinate profiles). The gain+radius row *is* U_{LAP} (Equation (9); the raw and gain–radius bases are equivalent by Proposition E.1); the two rows above it drop one coordinate; randomized V replaces the learned basis while preserving spectrum and norms; the Gram rows are the detector of Sastry & Oore (2020) in its published class-conditional form and with class-agnostic min/max; the bottom block of each model lists the eight standard detectors of Section 5 on the same images and split. Best per column in bold within each block. Full per-set tables in the supplementary CSVs.

published form, here with ≈ 22 calibration images per class) or pooled (*global*); a sample's score is its relative excursion outside $[\min, \max]$ summed per layer, divided by the layer's mean excursion on a held-out tenth of the calibration split, and summed over layers. Splits, subsets and corruption seeds are those of the profile. Gram wins one ResNet-50 structural cell (contrast s3, $+0.02$), trails the profile by 0.02–0.03 on the other four, and trails it clearly on the unsaturated ViT-B/16 cells and seven of the eight semantic sets (Table 5). Its calibration tables occupy $1.2\,\mathrm{GB}$ (ResNet-50) and $0.8\,\mathrm{GB}$ (ViT-B/16) against the profile's $28\,\mathrm{KB}$ (Appendix J).

More specifically, Gram trails the profile by 0.21–0.22 on ViT-B/16 contrast at severity 3 and Gaussian noise, and by 0.03–0.28 on seven of eight semantic sets. Its per-input statistic is approximately 10^4 times larger than the profile.

Other baselines. DICE uses the clean calibration mean and 70% pruning; SCALE uses percentile 85 on the unmodified features, including signed ViT features; GEN uses $\gamma = 0.1$ and the top 100 probabilities. All three use their standard anomaly-score direction.

The single-layer activation-shaping and residual baselines evaluated on the same cells (RMDS, ReAct, ASH-S; the last three rows of the upper block of Table 5) sit well below the profile on these structural shifts—ReAct and ASH-S fall to or below chance on the ResNet-50 cells—consistent with their design target of far-OOD semantic detection rather than covariate structure. At the same time, radius profiles are strong, beating gain alone on ResNet-50 contrast (0.958 vs. 0.737), while gain carries the ViT noise and blur signal. The relative performance of the two seven-dimensional variants depends on the model and shift. We use the joint $[\text{gain}, \text{radius}]$ profile as the default detector to retain both measurements (Proposition E.1). Gain summarizes directional alignment, whose contribution to detection is tested by the random-basis control.

Per-set ablation and baseline tables for every model and shift family are in the supplementary CSVs.

G.4 Shift detection at ImageNet scale

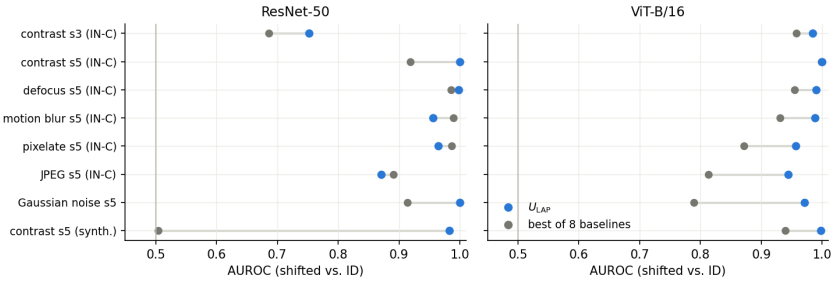


Figure 9: Structural shift detection on ImageNet-1K: for each evaluation set, U_{LAP} (blue) versus the best of eight baselines for that set (gray; MSP, entropy, energy, max-logit, feature norm, feature Mahalanobis, KNN, ViM), AUROC for separating shifted inputs from held-out ID validation images. Synthetic linear contrast reduction (bottom row) is the diagnostic case: every baseline sits at chance on ResNet-50 while the profile detects it. The semantic and far-OOD sets, where feature-density methods lead as the attribution predicts, are in Figure 10.

We evaluate the profile on standard shift detection. We score each input using ID-only calibration and measure the AUROC of separating each shifted evaluation set from held-out ID validation images. If the attribution of Section 5.1 is right, the profile should be strongest exactly where the shift is structural and weakest where it is purely semantic—detection performance becomes a prediction of the attribution story.

U_{LAP} transfers to ImageNet-1K without retraining or shift-specific tuning (Figure 9). On official ImageNet-C cells with ViT-B/16 it reaches 0.985 (contrast s3) vs. 0.957 for the strongest baseline, 0.990 (defocus) vs. 0.954, and 0.944 (JPEG) vs. 0.813; on ResNet-50 it leads on contrast and defocus and is near parity on motion blur, pixelation and JPEG.

Severe corruption also damages prediction and displaces features, allowing predictive and feature-density scores to detect those shifts. Synthetic linear contrast reduction barely moves ResNet-50's predictions: every baseline in Figure 9 falls to chance (0.504), while the profile detects the shift at 0.98. Feature-density methods are often stronger on natural semantic and far-OOD sets (Figure 10), consistent with the attribution results.

ImageNet-200 provides a designed contrast at scale. Against the 200 ID classes, the other 800 ImageNet classes (new content, same photographic domain) favor entropy (0.84 vs. 0.63 for U_{LAP} at $M = 5$), while official contrast s5 (0.98 vs. 0.92), defocus s5 (0.99 vs. 0.81) and Gaussian noise (0.98 vs. 0.95) favor the profile. ImageNet-R is subtle: it preserves class identity by construction, yet its renditions change how content reads to the classifier and entropy remains stronger. The profile's response follows changes in internal processing, which can vary among shifts that preserve class identity.

G.5 Statistical procedures and the attribution regression

The observational analysis of Section 5.1 regresses each signal on the two model-independent distances of Section 5 over ≈ 80k samples, with standardized coefficients and 95% cluster-bootstrap

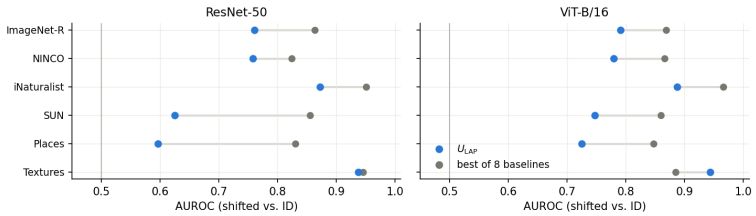


Figure 10: Companion to Figure 9: detection on the semantic and far-OOD sets, same construction. Feature-density methods lead here, as the attribution of Section 5.1 predicts—U_{LAP} measures structural, not semantic, deviation.

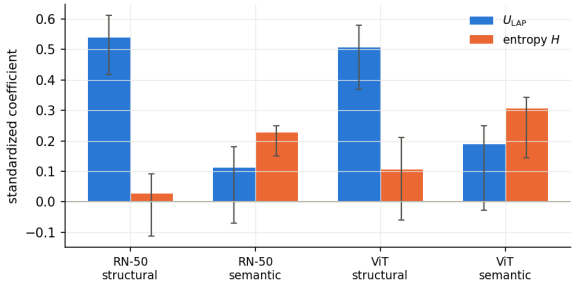


Figure 11: Observational double dissociation on $\approx$ 80k ImageNet samples: standardized coefficients of U_{LAP} and H on the model-independent structural and semantic distances (Appendix H), with 95% cluster-bootstrap intervals over evaluation sets. Companion to Figure 2, which shows the underlying per-sample relationships.

intervals over evaluation sets (Figure 11); rank-based (Spearman partial) checks agree with the linear fits. The factorial fit uses cluster bootstrap over corruption types.

G.6 The corruption-by-novelty experiment

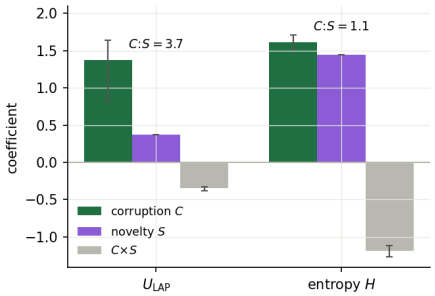


Figure 12: Coefficients of the controlled two-factor experiment on ImageNet-200 (2×2 factorial design), grouped by signal; error bars are cluster-bootstrap intervals over corruption types.

To measure responses to corruption and class novelty directly, we cross two on/off factors on ImageNet-200: an image is either from the 200 ID classes or from 800 unseen ones (class novelty, S), and it is shown either clean or with ImageNet-C corruption applied to the same image (corruption, C). We fit $U_{\mathrm{LAP}} \sim \beta_C C + \beta_S S + \beta_{CS} CS$ and the analogue for H, with cluster bootstrap over corruption types. Because every combination occurs, the fit separates what each signal responds to (Figure 12).

The two scores have different scales, so we compare the corruption and novelty coefficients within each score. For U_{LAP}, the corruption coefficient is $3.7\times$ the novelty coefficient (1.37 vs. 0.37). Entropy responds similarly to both factors (1.61 vs. 1.44, a ratio of 1.1), both of which damage prediction. These results show greater selectivity for corruption in U_{LAP} under this design. Entropy's near-zero structural coefficient in the observational fit partly reflected proxy overlap: corrupted

images also drift in CLIP space; the factorial, which manipulates the factors independently, does not inherit that overlap.

G.7 The confidence-matched split of Figure 1

On the split shown in Figure 1, a penultimate-feature Mahalanobis distance scores 0.34 AUROC. It measures the final hidden representation rather than the alignment profile across depth.

Figure 1 uses the noise-0.2 CIFAR-10 ensemble ($M = 10$). The in-distribution pool is the odd-indexed half of the CIFAR-10 test set (the even half is calibration), the shifted pool the 10k SVHN test subsample. Both pools are histogram-matched on ensemble maximum softmax probability (pooled-quantile bins, the smaller count kept per bin), leaving $4{,}138$ inputs per side with identical confidence distributions; every score that is a function of the output distribution is then at AUROC 0.500 by construction. Interior statistics are not constrained by the matching, so the split also separates *where* to measure from *what* to measure. Table 6 scores the profile against the penultimate-feature Mahalanobis distance of Lee et al. (2018) (class-conditional, training-set statistics, per member) and the feature norm on the same inputs, for all three CIFAR-10 ensembles (the clean and noise-0.4 splits are rebuilt with the same procedure; rebuilding the noise-0.2 split reproduces the figure's profile AUROC within 0.002). Feature Mahalanobis is at or near chance on every ensemble (≤ 0.56): SVHN's low-norm features sit close to every class mean under the shared covariance, the failure mode that motivates relative Mahalanobis (Ren et al., 2021). Probing inside the network is necessary for this split, and the learned-direction profile is what makes it sufficient.

Ensemble	MSP	U_{LAP}	feat. Maha (M=10 / single)	feat. norm
clean (n=3382)	0.500	0.948	0.564 / 0.53 ± 0.09	0.498
noise 0.2 (n=4138)	0.500	0.878	0.344 / 0.40 ± 0.08	0.691
noise 0.4 (n=4531)	0.500	0.820	0.366 / 0.41 ± 0.09	0.685

Table 6: **Interior baselines on the confidence-matched CIFAR-10 vs. SVHN split.** AUROC per ensemble; n is the matched count per side. Feature Mahalanobis is averaged over members (M=10) or scored per member (mean $\pm$ sd over seeds); feature norm is scored with low norm as the shift direction.

G.8 How much does averaging matter?

Averaging over independently trained models is a variance-reduction step, and its value depends on how noisy a single model's alignment is. At CIFAR scale a *single-layer* late-conv score varies substantially across seeds: SVHN AUROC is 0.59 ± 0.24 across seeds at $M = 1$, rising to 0.82 ± 0.09 at $M = 5$ and 0.86 at $M = 10$. The full depth profile is far less fragile because it already pools partially independent measurements across layers, reaching 0.79–0.87 from a single model. At ImageNet-200, where individual models are stronger and less variable, going from $M = 1$ to $M = 5$ moves profile detection by only 0.01–0.05.

We also evaluate checkpoint and posterior averaging. Averaging the last five checkpoints of one constant-learning-rate run tracks the ten-seed profile no better than that run's final checkpoint alone (per-layer Spearman 0.51–0.68 vs. 0.53–0.70), and a diagonal SWAG posterior built from the same trajectory lands at single-model AUROC (Appendix G.9). Within-basin averaging removes optimization noise, while independently initialized solutions provide function diversity. Across the evaluated settings, averaging stabilizes the profile and $M = 1$ remains usable at scale.

G.9 Negative posterior results

Ensemble $\mathrm{Var}[L]$ reaches only 0.42–0.64 AUROC, K-FAC Laplace 0.52–0.61, and single-trajectory approximations stay at single-model level.

G.10 Grokking suite details

For the modular-addition comparison, all interventions resume from the same memorized checkpoint and restore each weight matrix's Frobenius norm after every step. Early-layer targeting holds

alignment near its target without materially changing the delay to generalization. Late-layer targeting accelerates grokking in all eight seeds in the eight-seed comparison (sign test $p = 0.004$); the three-seed comparison in Table 4a reports a 33% reduction. The R_p–R_q comparison acts on the first feed-forward projection in the first transformer block. Theorem D.1 gives the directional-gradient analysis connecting the penalties to activation orientation.

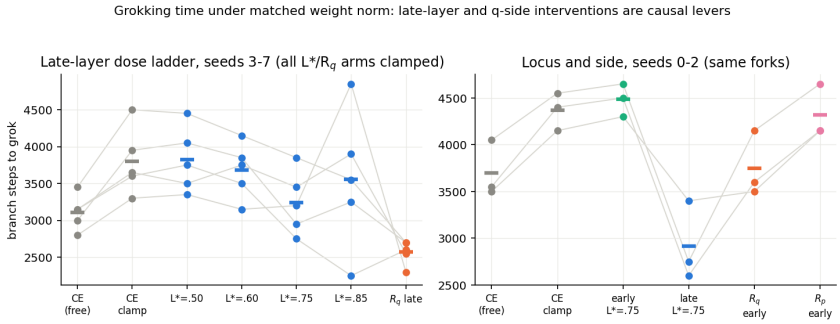


Figure 13: Per-seed grokking trajectories behind Table 4a. Left: late-layer target ladder with per-seed lines. Right: layer and side comparison on matched forks.

The full intervention suite also includes a LAR-triggered controller: every switch arm rescues $12/12$ runs, but the adaptive timing does not beat a matched fixed switch time.

Table 7 repeats the analysis on the full binary-operation suite of Shehper & Vaswani (2026): twelve tasks, of which five never grok; the table lists the seven that do. Two observations from the base runs: early-layer LAR is class-blind—tasks that never grok still reach early alignment as high as grokking tasks—while late-layer alignment separates the two groups at matched checkpoints. On the grokking tasks, forked and norm-clamped exactly as in Section 5.4, soft late-layer targeting accelerates or matches the baseline everywhere, and q-side occupancy accelerates on every task it permits—but *blocks* grokking outright $(0/3)$ on S_5 $x \cdot y \cdot x$, the same high-dimensional task on which the top-subspace intervention prevents grokking in the subspace sweep below: pulling activations into the fork-time top directions is helpful when those directions suffice for the task and harmful when they do not.

Task	groks?	baseline	late $L^\star$=0.75	q-side
$x + y$	yes	4367	2917	2570[c]
$x - y$	yes	2050	1983	1750
x/y	yes	1733	1667	1633
$x^2 + y^2$	yes[a]	50	50	50
S_5: $x \cdot y$	yes	1733	1583	1417
S_5: xyx^{-1}	partial[b]	3975	5125	3200
S_5: xyx	yes	9950	9133	blocked $(0/3)$

Table 7: The seven grokking tasks among the twelve binary operations of Shehper & Vaswani (2026) ($p = 97$ unless S_5); the other five never grok in base runs. Grokking tasks are forked at memorization and norm-clamped; steps to grok, mean over seeds, for the baseline and the two interventions of Section 5.4. [a]immediate: groks within one 50-step logging interval of the fork. [b]fragile task: groks in $1/2$ base seeds; fork arms grok in $2/3$ (baseline, late) and $1/3$ (q). [c]modular-addition q-side value is from the five-seed ladder arm, not the three-seed R_q arm of Table 4a.

G.11 Subspace dimension and direction identity

For S_5 xyx, the model receives two permutations x and y of five elements and predicts $x \circ y \circ x$. Each intervention resumes from a memorized checkpoint and runs for at most 12,000 further steps. A run counts as grokked when validation accuracy first reaches 99%, evaluated every 50 steps. The top-8 and bottom-8 subspaces span the late feed-forward matrix's right-singular vectors with the eight largest and smallest singular values, respectively. The R_q penalty encourages activation energy in the selected subspace. We recompute directions at every step and stop gradients through the SVD.

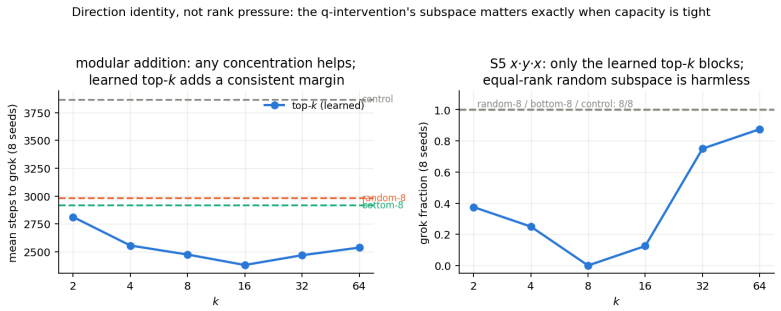


Figure 14: Subspace sweep. Left: on modular addition every concentration pressure accelerates, with a consistent margin for learned directions. Right: on S_5 $x \cdot y \cdot x$ the learned top-k blocks grokking for small k while an equal-rank random subspace is harmless.

To test direction identity at equal subspace dimension, we sweep $k \in \{2, 4, 8, 16, 32, 64\}$ and compare learned top-k, random-k and bottom-k subspaces of identical dimension (Figure 14). Two regimes emerge. On lower-dimensional tasks (modular addition, subtraction, S_5 multiplication) almost any concentration accelerates: top-k arms beat the baseline by 27–38% on addition, while random-8 and bottom-8 also accelerate by 23–25%; learned top-8 still beats random-8 in $8/8$ addition seeds, a consistent but modest direction-specific bonus.

The high-dimensional S_5 $x \cdot y \cdot x$ task behaves differently. Under confinement to the learned top-8 directions, $0/8$ runs grok; small learned k mostly blocks, $k = 32$ mostly groks and $k = 64$ fully groks. The equal-rank *random*-8 subspace, by contrast, groks $8/8$ and is slightly faster than the baseline, as does bottom-8. Thus subspaces of the same dimension produce different grokking outcomes. For the learned top-k subspaces, the fraction of runs that grok increases with k. Together, these comparisons show that the effect of the occupancy intervention depends on both the chosen directions and the subspace dimension.

G.12 Language-model pilot

Models: GPT-2 small (12 blocks, probe `mlp.c_fc`), TinyLlama-1.1B (22 blocks) and the language tower of LLaVA-1.5-7B (32 blocks), probe `mlp.gate_proj`; every block is probed. Per sequence, $a_\ell = \log \sum_t \|W_\ell h_t\|_2^2$ and $b_\ell = \log \sum_t \|h_t\|_2^2$ over the 256 tokens (the token matrix is H, as for ViT). The score is U_{LAP} in its global variant: the z-scored (g_ℓ, r_ℓ) profile under the quadratic form of Equation (9) with a single mean—no class or position conditioning—calibrated on 1,500 ID sequences; evaluation uses 800 disjoint ID sequences against 800 per shift. ID corpus: enwik8. Shifts, built exactly as the ID sequences (tokenize the corpus with the model's tokenizer, sample 256-token windows at random offsets): token-shuffled and token-reversed ID sequences, uniform random tokens, base64 strings, Python source, LaTeX source, and Shakespeare. "Fluent" in Figure 3a means the model predicts the shifted text as well as or better than ID text (the NLL columns of the tables), so increased perplexity does not flag these shifts. One semantic difference from the vision experiments: there, ID is the model's own training distribution; here ID is an operator-chosen reference corpus, enwik8. Calibration on this corpus defines the expected input statistics. The profile detects departures from these statistics even for domains, such as Python, that the model predicts well. Baseline: mean next-token NLL. In Figure 3a–b LLaVA is evaluated text-only through its language tower, an architecture-and-scale replication.

Image-token experiment (Figure 3c). Images are fed with the fixed prompt "`USER: <image>\nDescribe the image. ASSISTANT:`"; at every language-tower block, (a_ℓ, b_ℓ) is accumulated separately over the image-token slice (576 positions) and the trailing prompt-text positions of the same forward. Each slice gets its own globally calibrated U_{LAP} (800 ImageNet calibration images, 400 evaluation, 400 per shift). Shifts alter only the image: Gaussian noise ($\sigma = 0.22$) and Gaussian blur ($\sigma = 3$), the plainest members of the ImageNet-C noise and blur families (Hendrycks & Dietterich, 2019); uniform random images (the classic no-structure control); and DTD textures (Cimpoi et al., 2014), a standard far-OOD set also used in Appendix G.4. Figure 3b plots all four; their ordering reproduces the pooled bars of Figure 3c. The text-token profile is the

leakage control; its elevation on the extreme shifts (uniform 0.94) is attention carrying strongly anomalous image content into text positions.

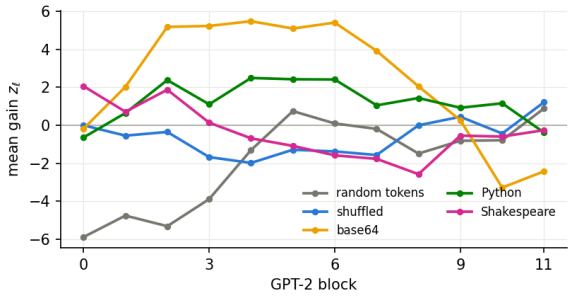


Figure 15: GPT-2 depth signatures (mean gain z_ℓ across the 12 blocks): breakdown depth tracks linguistic level. Random tokens break block 0 (lexical); shuffled tokens break the middle blocks with block 0 normal (unigrams intact, syntax destroyed); base64 and code *hyper*-align mid-network; Shakespeare is early-hyper, late-deficit (style).

The pooling curve of Figure 3b calibrates each (block, position) pair on 600 ID images; for pool size k, the per-layer mean z over k randomly sampled image positions forms a 32-dim vector scored by Mahalanobis distance, whitened on half of the ID evaluation images and evaluated on the other half, averaged over ten position subsets per k. Pooling within layers before combining across them matters: blur's per-block shifts have mixed signs, and averaging blocks first cancels its signal entirely. Per-position AUROC (k=1) is flat near chance across the whole image slice for both shifts. Structural corruptions of text (shuffled, reversed, random tokens, base64) are detected at 0.96–1.00 by profile and perplexity alike; only the fluent shifts separate the two signals. Depth signatures sort shifts by where processing breaks, tracking linguistic level from lexical (block 0) through syntactic (middle blocks) to stylistic (Figure 15). Within the image-token slice, pooling detects because correlated small per-token shifts add while their noise averages away, which rules out the reading that the profile merely catches a few outlier tokens; per-token localization, which works for the ViT classifier (Section 5.1), needs a probe that precedes attention mixing. The difficulty ordering of image shifts *inverts* relative to the vision experiments: pixel corruptions, easy for the ViT classifier, are the hard case for the LLaVA language tower, because the CLIP tower absorbs much of the pixel-level corruption before the projected embeddings arrive. A layer response is considered atypical relative to the responses measured on the calibration data. The language experiments use one ID corpus, one seed per model, and global calibration (Appendix K). Full tables and scripts are in the supplementary material.

G.13 Spatial experiments

We locate atypical processing within an image by evaluating alignment separately for each patch token in ViT-B/16, producing a spatial anomaly map (Table 2b). At each encoder MLP probe (blocks 0/3/6/9 and the block-11 second MLP layer), we estimate a mean and standard deviation for each grid position over 2,000 clean reference images. This position-specific calibration accounts for variation in activation statistics across the image. The measurement uses the same closed form as the global profile, evaluated per patch token.

Figure 16 shows the resulting maps for two selected perturbations. The anomaly is visible as a localized deviation whose sign depends on the perturbation—blur concentrates activation energy into the learned directions of the early block, noise scatters it out of them—and whose spatial coherence decays with depth as attention mixes information across tokens, the depth-resolved picture behind the single-map result.

For quantitative evaluation, we add Gaussian noise within random boxes on 500 separate images and label a patch as corrupted when the box covers at least half its area. The first-block signed alignment deficit (the negative position-calibrated z-score) achieves 0.924 pooled patch-level AUROC, compared with 0.466 for a position-calibrated token-norm baseline. This evaluation uses a single alignment coordinate; the joint gain–radius score is defined in Section 4.3.

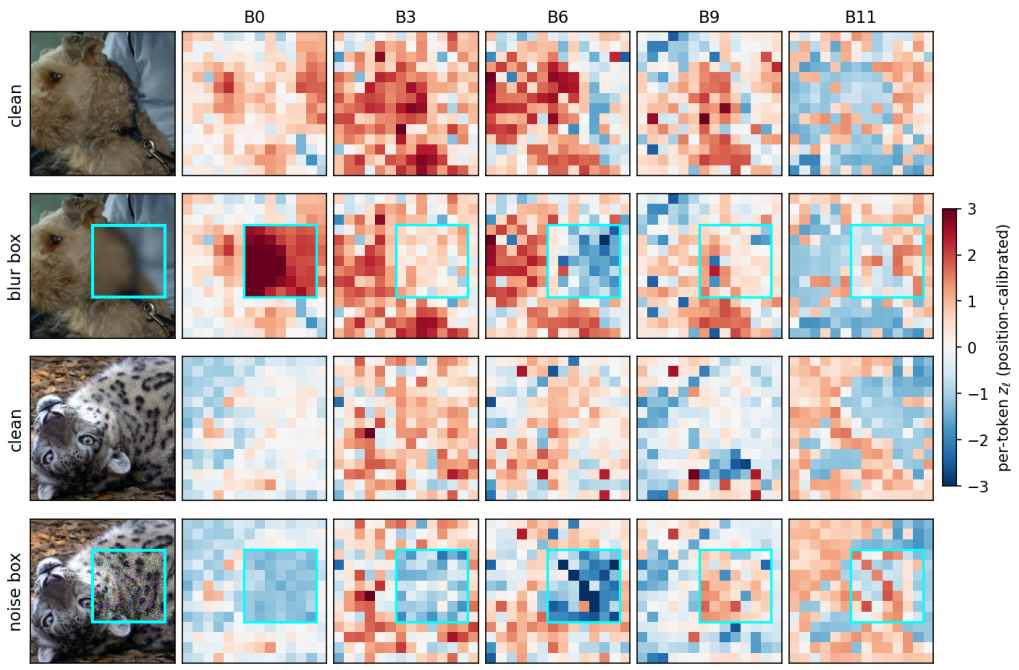


Figure 16: Per-token alignment tracks a spatial anomaly across depth (selected examples). Each pair of rows: a clean image with its five per-depth position-calibrated z-maps, then the same image with one box perturbed (outlined). A blur box *hyper*-aligns at the first encoder block (saturated red at B0) and a noise box under-aligns (blue, deepest at B6); both are crisply localized early and dissolve by the last blocks as attention mixes tokens. The clean rows show no box artifact.

Pastes of natural texture or ID-image crops produce no alignment deviation patch by patch and are localized instead by patch-neighbor semantic signals. This is the division of labor of Section 5.1, reproduced within single images.

G.14 Supervised error prediction

We evaluate error prediction for fixed ImageNet-1K ResNet-50 and ViT-B/16 classifiers. The target is whether the predicted class is incorrect, rather than whether an input is shifted. We calibrate the profile on half of the clean ImageNet validation set. We combine the remaining clean images with an equal number of shifted images sampled from the corruption and ImageNet-R sets, then split this balanced mixture by example into fitting and evaluation halves.

A logistic predictor estimates error probability from standardized entropy, the standardized log-transformed profile score, and their product. The product allows the profile's contribution to depend on the entropy level. Profile calibration uses only ID data; fitting the error predictor uses labeled clean and shifted examples. The underlying classifiers and their predictions remain unchanged. On held-out examples, entropy plus the profile outperforms entropy alone and combinations of entropy with feature Mahalanobis, DICE, SCALE, or GEN on both architectures (Table 2a).

H Model-Independent Distances

The attribution analyses (Section 5.1) use structural and semantic distances computed independently of the evaluated classifier and task labels. Both distances are calibrated on the same ID calibration half used everywhere else and evaluated on the disjoint evaluation half.

Structural distance D_s. For an image x with grayscale channel g, the feature map $\phi(x) \in \mathbb{R}^{18}$ collects:

- *Spectral shape* (6): log fractions of Fourier power in six radial bands. With $P(f) = |\hat{g}(f)|^2$ and band edges $0, 0.02, 0.05, 0.1, 0.2, 0.35, 0.71$ cycles/pixel,

$$\phi_b = \log \frac{\sum_{a_b \le \|f\| < a_{b+1}} P(f)}{\sum_f P(f)}, \qquad b = 1, \dots, 6.$$

- *Color moments* (9): per-channel mean, standard deviation and skewness.
- *Edge statistics* (3): mean and standard deviation of the finite-difference gradient magnitude $\|\nabla g\|$, and the root mean square of the 5-point Laplacian response.

Each coordinate is standardized by its ID calibration mean and standard deviation, giving $z(x)$. With $\hat{\Sigma}$ the calibration covariance of z and shrinkage $\lambda = 0.05$,

$$D_{\mathrm{s}}(x) = \log\big(1 + z(x)^\top (\hat{\Sigma} + \lambda I)^{-1} z(x)\big). \tag{33}$$

The quadratic form is a Mahalanobis distance in low-level statistics space; the shrinkage keeps the 18×18 inverse stable, and the $\log(1 + \cdot)$ compresses the heavy right tail so that extreme corruptions do not dominate a linear regression.

Semantic distance D_{m}. Let $e(x) \in \mathbb{R}^{768}$ be the unit-normalized CLIP ViT-B/16 (Radford et al., 2021) embedding of x (OpenAI weights—a model foreign to every classifier under test), and let $\mathcal{N}_k(x)$ be the $k = 10$ ID calibration images with the highest cosine similarity to x. Then

$$D_{\mathrm{m}}(x) = \log\Big(1 + \big(1 - \tfrac{1}{k} \textstyle\sum_{j \in \mathcal{N}_k(x)} \langle e(x), e_j \rangle\big)\Big), \tag{34}$$

the (compressed) mean cosine distance to the nearest ID content. It is small when calibration data contains images *of similar content*, whatever their low-level statistics, and grows with semantic novelty.

Properties. Neither distance sees the evaluated classifier, its features, or any label. Pooled over all evaluation sets the two are weakly correlated ($r = 0.15$), allowing their associations with each score to be compared in the regression of Section 5.1. CLIP embeddings also move under strong corruption, contributing to the residual overlap. The controlled factorial manipulates corruption and class novelty independently.

I Robotics External Validation

This appendix details the visuomotor-control validation summarized in Section 5.3. We apply the per-layer $(\log \|W_\ell h_\ell\|, \log \|h_\ell\|)$ measurements to four policy classes in three simulators. The coordinates are z-scored on an ID calibration half, and the profile is scored by Mahalanobis distance. For continuous control, calibration uses a single unconditional Gaussian (one mean, one pooled covariance). Episodes are split into calibration and evaluation halves by parity of episode index. We record the profile per policy query (one query per action chunk where policies emit chunks) and per-episode scores aggregate the first ten post-onset queries.

Baselines. Alongside the profile we score every rollout with: a Mahalanobis distance on the policy's penultimate feature (*feat*); a Mahalanobis distance on per-channel image mean and standard deviation (*pix*), the calibrated pixel-statistics control; ensemble action disagreement $\mathrm{tr}\,\mathrm{Cov}_m[a^{(m)}]$ over M independently trained seeds scoring the shared observation off-policy (*disagree*); and, where the policy defines a per-state action distribution, its Monte-Carlo entropy (*entropy*, $K = 16$ samples of a GMM head). All detection numbers are per-step (or per-query) AUROC of ID-evaluation against the shifted condition.

I.1 From-scratch RL policies (ManiSkill3)

PPO policies with a NatureCNN trunk (six profiled layers) are trained from scratch on PickCube and PushT (Tao et al., 2025) (M=3 seeds; PushT analyses use the two seeds whose deterministic policies retain training success). Conditions cover five photometric axes at two severities, an env-level

Trained on	evaluated on		
	Spatial	Object	Goal
Spatial	—	0.95 (.86)	0.86 (.84)
Object	0.99 (.85)	—	0.98 (.86)
Goal	0.81 (.96)	0.98 (.92)	—

Table 8: Cross-suite generalization of the suite-specialized 7B OpenVLA-OFT checkpoints (Section 5.3). Each row is a checkpoint, each column a LIBERO suite it was *not* trained on; entries are the profile's per-query detection AUROC with the feature-Mahalanobis baseline in parentheses. Every checkpoint has 0.00 ID failure. The ordering tracks how much of the scene changes: object swaps are most unfamiliar, same-scene goal changes least, and the one cell where the feature baseline wins (Goal→Spatial) is the pairing with no structural change at all.

camera-pose axis, and, on a separate PickSingleYCB variant, a held-out object-geometry axis: the 74 graspable YCB objects are split $55/19$, policies train on the 55, and the held-out 19 appear only at evaluation—same scene, same camera, same lighting.

Condition	fail	profile	feat	pix	disagree
held-out geometry	0.60	**0.61**	0.60	0.49	0.59
Gaussian noise (s4)	0.42	**0.95**	0.46	0.80	0.50

Table 9: Held-out geometry is the pixel-invariant axis (PickSingleYCB split, acting seed at 0.53 deterministic ID success; 256 episodes per condition). The calibrated pixel-statistics control drops to chance on unseen geometry while the profile detects it; the feature baseline fails on the photometric axis instead. The profile is the only signal above 0.59 on both shift families.

Onset lead time. We measure lead time with mid-episode onset: the corruption switches on at a fixed step and alarm latency is counted from onset, at thresholds set to the ID 95th percentile per signal. Across all 15 onset arms on both tasks the profile alarms at the onset step itself (detection within ten steps $= 1.00$), while ensemble disagreement never crosses its threshold (median latency ∞; within-ten-steps rate 0.02–0.15, indistinguishable from the ID false-alarm rate)—including arms in which a third of episodes subsequently fail. Within mixed-outcome conditions the profile adds information to disagreement for predicting episode failure (Table 3b). Table 3b extends this test to further ManiSkill tasks trained with the same PPO-RGB recipe (three seeds each; 10M steps, PullCube 3M; AnymalC-Reach 20M with γ=0.99 and joint-delta control).

Batch failure-rate tracking. We evaluate fixed PPO checkpoints under various OOD conditions. Two final reset/noise seed sets give 96 batches of 32 episodes per policy, totaling 6,144 evaluation episodes. Training seed 0 supplies the acting policy; three independently trained checkpoints supply the ensemble profile and action disagreement.

All methods use the same first-ten-query observation window; an episode fails if it never achieves success within 50 steps. For each batch, we plot its failure fraction against $\log(1 + \overline{S})$, where $\overline{S}$ averages the score over queries and episodes. We pool both policies' 192 batch points on these log scales and fit one linear regression per method.

I.2 From-scratch imitation policies (LIBERO)

Multitask BC-RNN policies (ResNet-18 encoder per camera, GMM head, 39 profiled layers; M=3 seeds) are trained on LIBERO-Spatial (Liu et al., 2023a) and evaluated on their own suite (ID), the Object and Goal suites, photometric arms, and mid-episode onset arms (20 trials per task). The GMM head defines a true per-state action entropy, the only one among our four policy classes.

At n=600 episodes, the within-condition analysis finds no incremental association between the profile and episode failure beyond ensemble disagreement (likelihood-ratio $p = 0.94$).

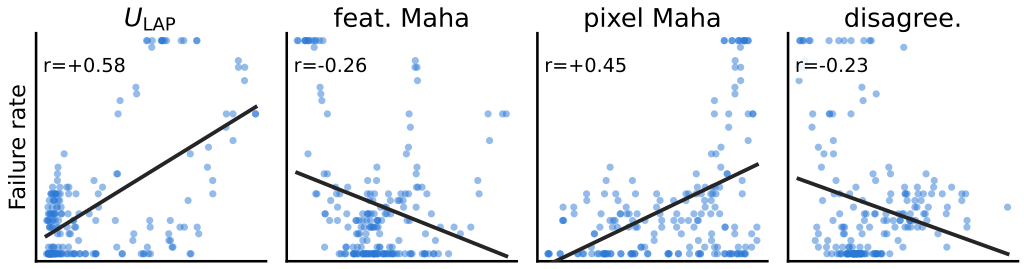


Figure 17: **Profiles have the strongest positive pooled correlation with failure rate.** Each panel contains the same 192 batches. U_{LAP} denotes the ensemble profile. Columns compare U_{LAP}, feature Mahalanobis, pixel Mahalanobis, and ensemble action disagreement. Black lines are pooled linear fits; annotations report Pearson correlations. All methods use identical episodes and observation windows.

Condition	fail	profile	entropy	disagree
contrast (s4)	0.89	**1.00**	0.54	0.70
noise (s4)	1.00	**1.00**	0.44	0.85
suite: Object	1.00	**1.00**	0.58	0.82
suite: Goal	1.00	**1.00**	0.63	0.82
onset arm (latency / rate_{10})				
contrast onset	—	0 / 1.00	16 / 0.32	130 / 0.02
noise onset	—	0 / 1.00	35 / 0.28	36 / 0.10

Table 10: LIBERO BC-RNN, $M{=}3$, 200 episodes per condition. Left: detection AUROC. Right: onset arms—median alarm latency after onset (steps) and the fraction of episodes alarmed within ten steps of onset. The true action entropy is near-blind in detection yet eventually fires after onset; the profile leads it by 16–35 steps and leads ensemble disagreement by 36–130.

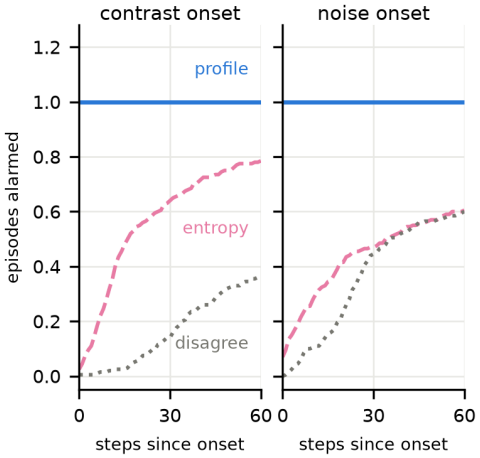


Figure 18: **Profile alarms precede action-uncertainty alarms.** Cumulative fraction of LIBERO BC episodes alarmed against steps since corruption onset, for contrast reduction and Gaussian noise. Curves compare the profile, action entropy, and ensemble disagreement.

I.3 A PRETRAINED 7B VISION-LANGUAGE-ACTION MODEL (LIBERO)

OpenVLA-OFT (Kim et al., 2025) provides 7B-parameter checkpoints fine-tuned separately on each LIBERO suite (Llama-2-7B over DINOv2+SigLIP towers; we profile 48 depths: every LLM block's down-projection, every fourth vision-tower MLP, and the projector). Each checkpoint reproduces its published performance in our harness (0.00 ID failure; 30/30 on the official evaluation script). The continuous action head defines no distribution, so the impact-side baselines are feat and pix only.

Each checkpoint is evaluated on its own suite (ID), the other two suites, and photometric arms (100 episodes per condition).

Table 8 gives the cross-suite matrix. On the corruption arm, Gaussian noise drives the failure rate to 0.81 and the profile detects it at 1.00, with the feature baseline at 0.52 and the pixel control at 0.60 (Table 3). The matrix is graded exactly as structural change predicts: object-identity swaps read most unfamiliar (0.95–0.99), and the Goal↔Spatial pairs—same scenes, same objects—least (0.81–0.86). The one cell where the feature baseline beats the profile (Goal checkpoint on Spatial, 0.96 vs. 0.81) is the cell with the smallest structural difference and a purely task-level one: the semantic signal dominates precisely where structure is preserved.

I.4 A PUBLISHED GENERALIST POLICY (ROBOCASA365)

The released Human300 Diffusion Policy checkpoint (Chi et al., 2023; Nasiriany et al., 2026) (106M parameters; we profile the observation encoder) is evaluated on a scene-by-task factorial: trained tasks in pretraining kitchens (ID), the same tasks with held-out object instances, the same tasks in held-out kitchens, held-out composite tasks, and a noise arm (40 episodes per condition; ID failure 0.82—this baseline family is the weakest on the RoboCasa365 leaderboard, and outcomes remain mixed).

Condition	fail	profile	feat	pix
Gaussian noise (s4)	0.93	**1.00**	0.63	0.41
held-out tasks	1.00	0.96	**1.00**	0.51
held-out kitchens	0.61	**0.85**	0.55	0.69
held-out instances	0.33	**0.82**	0.41	0.42

Table 11: RoboCasa365 factorial. Kitchen-to-kitchen ID variance swamps the pixel footprint of observation noise—the calibrated pixel-statistics control sits at chance while the profile is perfect. The profile is the strongest signal on every condition except held-out tasks, where the feature baseline saturates.

Held-out kitchens and held-out object instances are detected by the profile alone (0.85/0.82; the feature baseline at or near chance, the pixel control at most 0.69), at n=40 episodes per condition.

We extract the profiles from the four policy classes with their trained parameters fixed. Appendix K discusses the scope of these evaluations.

J COMPUTATIONAL COST

Inference overhead. At batch 200 on one RTX 4090, the full seven-probe profile raises a forward pass from 103.5 to 112.4 ms for ResNet-50 (+8.6%) and from 235.2 to 258.7 ms for ViT-B/16 (+10.0%), including feature capture and bookkeeping. The cost per probe is two reductions over tensors the forward pass already produces (Theorem 3.3); no SVD is ever taken at test time, and for convolutions the ones-kernel identity of Appendix F replaces the patch matrix with one auxiliary single-channel convolution. Language models pay the same two reductions per probed block; the robotics deployments record the profile once per policy query, so the relative overhead falls further when policies emit action chunks.

Calibration footprint. Calibration stores per-coordinate means and standard deviations, class means, and one pooled covariance: at 1000 classes with $2P = 14$ coordinates this is $\approx$ 28 KB, against $\approx$ 100 MB feature banks for nearest-neighbor detectors and 0.8–1.2 GB of per-class min/max tables for the Gram-matrix detector of Table 1, and it is estimated in a single pass over the ID calibration split. The global variant used for language models and control drops the class means, shrinking the footprint further; no variant requires access to shifted data, gradients, or a second forward pass.

K LIMITATIONS

Inputs can remain structurally familiar while prediction degrades, as CIFAR-10.1 and several appearance transforms show. The joint profile uses activation scale as well as alignment, and its incremental value for uncertainty estimation depends on the model. The intervention experiments establish effects of the tested regularizers on grokking in small algorithmic transformers. The inference experiments evaluate shift detection and predictive utility in fixed models. The language-model experiments use one reference corpus and one checkpoint per model, spanning 124M to 7B parameters. Probe selection depends on the task, and calibration requires representative ID data. The archetypes describe the profile patterns observed in the evaluated shifts.