# WHERE NETWORKS ADAPT AND WHERE THEY FAIL: LAYER-LOCALISED DIAGNOSTICS FOR CONTINUAL LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We ask two questions that continual learning methods leave unanswered: *which* layer is currently the active site of adaptation, and *which* individual predictions are unreliable after a task shift? We show both admit diagnostic answers from a single principle — *activation-subspace consistency* — measured at training time and inference time respectively.

**Gradient Direction Coherence (GDC)**, the mean pairwise cosine similarity of per-sample gradient directions within a mini-batch, identifies the adaptation boundary at the layer with minimum coherence. In plain feedforward CNNs this argmin is seed-stable ($\ell^* = \texttt{block2}$, 5/5 seeds); under pervasive residual connections the gradient-space signal collapses predictably, a boundary condition the theory predicts. An exploratory activation-space variant shows a depth-correlated signal in ResNet-18 and ViT-small, reported as a preliminary finding rather than a resolution of the architectural restriction (see Introduction for caveats).

**Activation Projection Score (APS)**, computed in a single forward pass without label access, measures how well a sample's activations fit the training subspace at each layer. Across four configurations spanning plain CNNs, ResNet-18, and ViT-small ($n$=5 seeds each), APS at the deepest stage predicts per-sample correctness ($\rho = 0.253$–$0.286$; below softmax confidence by $0.14$–$0.19$ AUROC) and detects distribution shift (AUROC-OOD up to $0.990$), and localises OOD detection to the final stage on a standard ImageNet-1k benchmark (ResNet-50, no fine-tuning; AUROC-OOD $= 0.866$). Against the Mahalanobis detector (Lee et al., 2018) APS wins at the deepest layer in CNN and ResNet architectures ($\Delta = +0.109$, $+0.160$, $p<0.001$, label-free) — where Mahalanobis degrades under ill-conditioned high-dimensional covariance estimation — and ties it at ceiling in ViT-small. We additionally document that naive lr-scaling from GDC yields no reliable retention improvement ($n$=5, $d = 0.02$) and that the largest previously-reported single-seed effect was a 5-SD outlier baseline artefact.

## 1 INTRODUCTION

A neural network trained sequentially on multiple tasks does not forget uniformly. Probe accuracy, weight displacement, and activation statistics all degrade in a pattern that is strongly depth-dependent and layer-specific: the deepest representation stage degrades catastrophically while early feature-extraction layers remain largely intact. Yet the dominant framing of continual learning treats forgetting as a global phenomenon — something to be penalised or replayed away across all parameters — without asking *which* layer is currently reorganising its representations, or *which* individual predictions are unreliable after a shift has occurred. This paper argues that both questions have diagnostic answers sharing a common underlying signal.

Regularisation-based methods such as EWC (Kirkpatrick et al., 2017) and Synaptic Intelligence (Zenke et al., 2017) answer the question "which parameters matter" but not "which layer is the active adaptation site." Separately, feature-space OOD detectors such as the Mahalanobis distance score (Lee et al., 2018), energy-based scoring (Liu et al., 2020), and subspace-residual methods (Ming et al., 2022) detect that a prediction may be wrong without resolving *where in the*



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.

*network* the mismatch originates, and require task-label access to evaluate the correct output head — unavailable after a shift.

We propose that both questions can be answered by measuring *activation-subspace consistency*: the degree to which a layer's representations remain within the subspace established during training. This principle manifests as two signals. **Gradient Direction Coherence (GDC)**, measured during training, captures the cosine consistency of per-sample gradient directions at each layer across a mini-batch; the layer with minimum GDC identifies the adaptation boundary. In plain feedforward CNNs this boundary is seed-stable ($\ell^* = \texttt{block2}$, 5/5 seeds); under pervasive residual connections the gradient-space signal collapses — a predicted, not incidental, failure. An exploratory activation-space variant shows a depth-correlated signal in ResNet-18 and ViT-small, but it tests a structurally different hypothesis and its boundary coincides with APS's own deepest-layer signal rather than offering independent localisation; we treat it as a preliminary finding, not a resolution of the restriction (Appendix D). **Activation Projection Score (APS)**, measured at inference time, captures how well a sample's activations at each layer fit the PCA subspace of training activations. Across four configurations spanning plain CNNs, ResNet-18, and ViT-small ($n$=5 seeds each), APS at the deepest stage predicts per-sample correctness and detects distribution shift, with shallow layers held at chance as negative controls (§4). It outperforms the label-supervised Mahalanobis detector (Lee et al., 2018) at the deepest, highest-dimensional layer in CNN and ResNet architectures — where Mahalanobis degrades under ill-conditioned covariance estimation — and ties it at ceiling in ViT-small, all without class labels. On a standard ImageNet-1k OOD benchmark (ResNet-50, pretrained, no fine-tuning) APS localises OOD detection to the final stage, validating it as a general detector at scale, though not the continual-learning claims at that scale. APS's core statistic — a residual against the training principal subspace — is deliberately in the same family as ViM (Wang et al., 2022) and other subspace OOD scores; our contribution is not detection performance but the use of this quantity as a label-free, per-layer *localiser* of the adaptation frontier, linked to the train-time GDC signal by a single principle. Where task identity is available, softmax confidence still predicts correctness better than APS, and we report that openly.

We also document that naive learning-rate scaling from GDC produces no reliable retention improvement ($n$=5, $d = 0.02$, $p = 0.483$), and that the dominant originally-observed effect ($3.6\times$) was a 5-SD outlier baseline artefact.

**Contributions.**

1. **GDC**, a train-time adaptation-boundary signal, with a principled account of when the gradient-space form holds (no residual connections) and when it collapses (residual streams, forced-uniform MoE routing); plus a preliminary activation-space variant reported with explicit caveats (Appendix D).
2. **APS**, an inference-time, label-free, per-layer diagnostic for per-sample correctness and distribution shift, replicated across four configurations spanning CNNs, ResNet-18, and ViT-small (with block0/embed saturation as negative controls), outperforming label-supervised Mahalanobis at the deepest layer and validated as a general OOD detector at ImageNet scale (OOD sub-claim only).
3. **A unifying principle** — activation-subspace consistency as the common mechanism underlying both signals — that predicts from architecture alone where each signal works and where it fails.
4. **A documented null result** for GDC-derived lr-scaling with a seed-variance analysis characterising the retention noise floor that makes single-seed continual-learning intervention claims unreliable.

## 2 Related Work

**Continual learning and forgetting diagnostics.** The continual learning literature offers three families of methods for managing catastrophic forgetting (Parisi et al., 2019; van de Ven & Tolias, 2019): regularisation (Kirkpatrick et al., 2017; Zenke et al., 2017; Li & Hoiem, 2017), rehearsal (Rebuffi et al., 2017; Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019), and architectural isolation (Rusu et al., 2016; Mallya & Lazebnik, 2018). All operate on the full parameter vector without resolving

which layer is the active adaptation site. Closest to our diagnostic aims, Ramasesh et al. (2022) characterise the anatomy of forgetting through probing and find depth-structured degradation; Davari et al. (2022) show that representation forgetting precedes output accuracy degradation. Our contribution is a *gradient-based* boundary signal (GDC) that identifies the adaptation layer online, and an *activation-based* signal (APS) that localises post-shift degradation at inference without labels — both complementing the probe-based literature that requires labelled evaluation sets.

**Out-of-distribution detection and predictive uncertainty.** Softmax max-probability (Hendrycks & Gimpel, 2017) and its refinements — ODIN (Liang et al., 2018), temperature scaling (Guo et al., 2017), and deep ensembles (Lakshminarayanan et al., 2017) — provide output-space uncertainty signals that require task-identity to select the correct head. Representation-space detectors remove this dependency by scoring intermediate features: the Mahalanobis detector (Lee et al., 2018) fits a class-conditional Gaussian, deep-nearest-neighbour scoring (Sun et al., 2022) is non-parametric, and — closest to APS — ViM (Wang et al., 2022) derives an OOD score from the norm of a feature's *residual against the training principal subspace*, the same projection-residual quantity APS computes (Eq. 5). We therefore state the relationship plainly: APS is not proposed as a new OOD detector competing on benchmark AUROC, and its core statistic is in the ViM/subspace-residual family. Its novelty is in *how the quantity is used*: (i) where ViM, Mahalanobis, and KNN produce one image-level score at a single (penultimate) layer, APS is evaluated at *every* layer to yield a depth-resolved degradation *profile* whose argmax localises the adaptation frontier — a "which-layer" answer these detectors are not built to give; (ii) APS is label- and logit-free (ViM fuses class logits, Mahalanobis needs labels), so it applies immediately after a shift without task identity; (iii) the same score predicts per-sample *correctness*, not only ID/OOD membership; and (iv) most importantly, APS is one half of a principle that ties this inference-time residual to a *train-time* gradient signal (GDC) — the contribution is that link and the localisation it enables, not detection performance, and softmax confidence still beats APS on correctness when task identity is available (Table 10).

**Gradient statistics, activation analysis, and layer-wise adaptation.** Yosinski et al. (2014) establish the early-generic / late-specific hierarchy that APS quantifies exactly (block0 saturated at 1.000; block3 the largest OOD gap and post-shift drop). Representation-similarity methods — SVCCA (Raghu et al., 2017), CKA (Kornblith et al., 2019), Morcos et al. (2018) — characterise layer geometry but do not localise adaptation boundaries or predict per-sample correctness, while Neyshabur et al. (2020) and Achille et al. (2019) document depth-structured plasticity consistent with our GDC interpretation. Layer-wise adaptive-rate methods (LARS (You et al., 2017)) resemble our tested lr-scaling but target large-batch stability, so our null result does not bear on them. GDC measures *within-batch* gradient-direction coherence across samples at each layer — distinct from the across-step consistency of Fort & Scherlis (2019) — and it is this within-batch quantity that reflects layer-level input-distribution consistency.

# 3 METHODS

We introduce two signals — Gradient Direction Coherence (GDC) and Activation Projection Score (APS) — and derive both from a single principle stated formally in Section 3.4. Notation is introduced at first use throughout.

## 3.1 GRADIENT DIRECTION COHERENCE (GDC)

**Definition.** Let $\ell(\mathbf{x}_i; \mathbf{W})$ denote the per-sample loss for sample $i$. For a mini-batch of size $B$, the *per-sample gradient* at layer $\ell$ is

$$\mathbf{g}_\ell^{(i)} = \mathrm{vec}\left( \frac{\partial\, \ell(\mathbf{x}_i; \mathbf{W})}{\partial\, \mathbf{W}_\ell} \right) \in \mathbb{R}^{p_\ell}, \tag{1}$$

where $\mathrm{vec}(\cdot)$ flattens the weight-gradient tensor and $p_\ell$ is the number of parameters in layer $\ell$. The unit-normalised gradient direction is $\hat{\mathbf{g}}_\ell^{(i)} = \mathbf{g}_\ell^{(i)} / \|\mathbf{g}_\ell^{(i)}\|_2$.

**Gradient Direction Coherence** at layer $\ell$ for a mini-batch $\mathcal{B}$ is the mean pairwise cosine similarity of the unit-normalised gradient directions:

$$\mathrm{GDC}_\ell = \frac{2}{B(B-1)} \sum_{i<j} \hat{\mathbf{g}}_\ell^{(i)} \cdot \hat{\mathbf{g}}_\ell^{(j)}. \tag{2}$$

$\mathrm{GDC}_\ell \in [-1, 1]$. High values indicate that samples in the mini-batch produce gradient directions pointing in similar directions at layer $\ell$ — the layer is receiving a consistent update signal. Low or negative values indicate that gradient directions are dispersed across the unit sphere — the layer is receiving conflicting update signals from different samples.

An $O(B)$ equivalent formulation and implementation notes (per-sample gradients via `vmap`) are in Appendix H.

**Adaptation boundary identification.** We define the *adaptation boundary* as the layer with the minimum GDC value at a given training step:

$$\ell^* = \underset{\ell \in \{1,\ldots,L\}}{\arg\min}\ \mathrm{GDC}_\ell. \tag{3}$$

This layer is receiving the most dispersed gradient signal — its input distribution is the most variable across samples — i.e. the deepest stage still actively reorganising. (An earlier rank-based scheme converting GDC into per-layer learning-rate multipliers gave no reliable benefit, Section 4.5; the diagnostic use of $\ell^*$ is supported independently.)

### 3.2 Activation-Space GDC (Exploratory)

Gradient-space GDC requires gradients to attenuate monotonically with depth (Section 3.4); residual connections violate this by giving every layer near-direct loss access. We explored whether an activation-space analogue can stratify depth where the gradient-space signal fails. The variant we tested measures *across-time* drift of the batch-mean activation (defined in Appendix D, Eq. 10) – a structurally different quantity from gradient-space GDC's *within-batch, across-sample* dispersion; a closer per-sample analogue remains untested. It yields a depth-correlated boundary in ResNet-18 and ViT-small, but that boundary coincides with the deepest layer APS already flags as most degraded, adding no spatial information beyond APS in these architectures (unlike the plain-CNN case, where GDC and APS localise to adjacent but distinct layers). We therefore report it as a preliminary finding, not a resolution of the architectural restriction; full results and caveats are in Appendix D.

### 3.3 Activation Projection Score (APS)

**Training-time basis construction.** After training on task $\mathcal{T}$, for each layer $\ell$ we collect activations over $\mathcal{D}_{\text{train}}$:

$$\mathbf{A}_\ell = \left[\mathbf{a}_\ell(\mathbf{x}_1),\, \mathbf{a}_\ell(\mathbf{x}_2),\, \ldots,\, \mathbf{a}_\ell(\mathbf{x}_N)\right]^\top \in \mathbb{R}^{N \times d_\ell}. \tag{4}$$

We mean-centre $\mathbf{A}_\ell$ and compute its top-$K$ right singular vectors via truncated SVD, yielding the basis matrix $\mathbf{V}_\ell \in \mathbb{R}^{d_\ell \times K}$. The columns of $\mathbf{V}_\ell$ span the $K$-dimensional subspace that captures the most activation variance for task $\mathcal{T}$ at layer $\ell$. This computation is performed once after training and requires $O(Nd_\ell)$ storage; only $\mathbf{V}_\ell$ (size $d_\ell \times K$) is retained.

**Definition.** For an incoming test sample $\mathbf{x}$ and its activation $\mathbf{a}_\ell(\mathbf{x}) \in \mathbb{R}^{d_\ell}$, the **Activation Projection Score** at layer $\ell$ is the fraction of activation energy captured by the training subspace:

$$\mathrm{APS}_\ell(\mathbf{x}) = \frac{\left\|\mathbf{V}_\ell^\top \left(\mathbf{a}_\ell(\mathbf{x}) - \bar{\mathbf{a}}_\ell\right)\right\|_2^2}{\left\|\mathbf{a}_\ell(\mathbf{x}) - \bar{\mathbf{a}}_\ell\right\|_2^2 + \epsilon}, \tag{5}$$

where $\bar{\mathbf{a}}_\ell$ is the training-set mean activation at layer $\ell$ and $\epsilon > 0$ is a small constant for numerical stability. $\mathrm{APS}_\ell(\mathbf{x}) \in [0, 1]$. A value near 1 indicates that the sample's activation lies almost entirely within the training subspace — its representation is consistent with the training distribution. A value near 0 indicates that the activation energy lies mostly in directions not captured by training — the representation is anomalous with respect to the learned task.

**Derived population-level measures.** From per-sample APS values we derive three population-level diagnostic measures used in our experiments:

1. **OOD gap** at layer $\ell$:

$$\Delta_\ell^{\text{OOD}} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{train}}}\big[\text{APS}_\ell(\mathbf{x})\big] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{OOD}}}\big[\text{APS}_\ell(\mathbf{x})\big]. \tag{6}$$

   A large positive gap indicates that in-distribution samples are well-captured by the training subspace while OOD samples are not, making $\Delta_\ell^{\text{OOD}}$ a layer-localised distributional discriminator.

2. **Post-shift drop** at layer $\ell$:

$$\Delta_\ell^{\text{shift}} = \mathbb{E}_{\mathcal{D}'_{\text{train}}}\big[\text{APS}_\ell(\mathbf{x})\big] - \mathbb{E}_{\mathcal{D}_{\text{train}}}\big[\text{APS}_\ell(\mathbf{x})\big], \tag{7}$$

   where $\mathcal{D}'_{\text{train}}$ denotes the original task-1 evaluation set after the model has been trained on task 2. A large negative $\Delta_\ell^{\text{shift}}$ indicates that the layer's representation of the original task has degraded.

3. **Correctness correlation** at layer $\ell$:

$$\rho_\ell = \text{Spearman}\big(\{\text{APS}_\ell(\mathbf{x}_i)\},\ \{\mathbb{1}[\hat{y}_i = y_i]\}\big)\,, \tag{8}$$

   the Spearman rank correlation between per-sample APS and binary prediction correctness. Area under the ROC curve (AUROC) for correctness prediction and OOD detection is computed using per-sample $\text{APS}_\ell$ as the ranking signal.

**Computational cost.** APS adds one matrix-vector product per layer at inference ($O(Kd_\ell)$ alongside the standard forward pass), stores $K \times d_\ell$ floats per layer after training, and requires no backward pass, no label access, and no task-identity knowledge.

### 3.4 The Unifying Principle

Both GDC and APS are instances of a single underlying quantity: *input-distribution consistency at a layer*.

**Informal statement.** A layer $\ell$ is *consistent* with respect to an input distribution $p(\mathbf{x})$ when: (a) the gradient directions induced by different samples from $p$ are similar at layer $\ell$ (high GDC), and (b) the activations of new samples from $p$ lie within the subspace established during training on $p$ (high APS). Both conditions fail when $p$ shifts: the gradient signal becomes dispersed as the layer reorganises (low GDC), and the activation subspace no longer captures the new samples (low APS).

**Architectural conditions for GDC.** GDC stratifies by depth — producing a monotone coherence gradient from shallow (high GDC) to deep (low GDC) — under a specific condition: gradients must be attenuated monotonically with depth. This holds in plain CNNs but fails under **residual connections**, which give every layer near-direct loss access: in ResNet-18 the GDC range compresses from $0.032$ to $0.004$ ($8\times$) and $\ell^*$ is unstable; in transformers $\rho(\text{GDC},\ \text{depth}) \approx -0.25$ (not significant) versus $-1.0$ in plain CNNs. It also fails under **forced-uniform MoE routing**: uncoerced routing gives $\rho(\text{GDC},\ \text{freq}) = -0.881$; forcing uniformity drives it to $-0.024$ ($p$=0.96), destroying the signal by design.

Table 1 summarises these boundary conditions; the MoE routing sweep (Appendix Figure 5) — the monotone decay of $\rho(\text{GDC},\ \text{freq})$ as routing is forced uniform — is the sharpest confirmation that GDC tracks input-distribution consistency rather than any architecture-specific property.

**GDC and APS as dual signals.** GDC identifies where the network is actively changing (low coherence) and APS where those changes are most visible (large subspace drop); in plain CNNs these are adjacent layers — `block2` (cause) and `block3` (consequence) — connected by the same mechanism (Section 5).

Table 1: Architectural conditions for *gradient-space* GDC stratification. GDC is informative wherever layers receive differentiated, depth-attenuated gradients; uninformative wherever architecture or routing policy equalises gradient access. An exploratory activation-space variant for the ResNet-18/ViT failures is reported in Appendix D, with caveats.

| **Architecture** | **Mechanism** | **GDC range** | $\ell^*$ **stable?** |
|---|---|---|---|
| Plain CNN (LayeredNet) | No skip connections | 0.032 | ✓ (5/5 seeds) |
| Plain CNN + attention stages (HybridNet) | Conv stages: no skips; Attn stages: residual stream | 0.032 (conv) $\approx 0$ (attn) | ✓ (conv only) |
| ResNet-18 | Pervasive residuals | 0.004 | × (unstable) |
| Vision Transformer (ViT) | Residual stream | $\approx 0$ | × |
| MoE (uncoerced routing) | Differentiated expert inputs | – | ✓ ($\rho = -0.881$) |
| MoE (forced-uniform routing) | Equalised expert inputs | – | × ($\rho = -0.024$) |

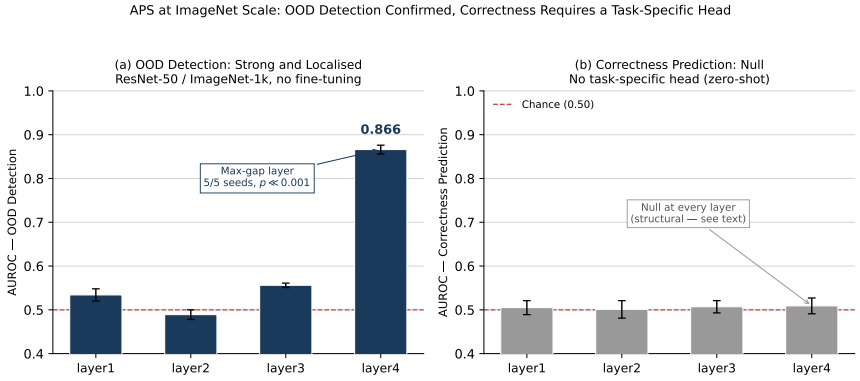



Figure 1: **GDC spread collapses** $8\times$ **under residual connections.** *Left:* LayeredNet (plain CNN): range $= 0.032$, seed-stable boundary ($\ell^* =$ `block2`, 5/5 seeds). *Right:* ResNet-18: range $= 0.004$ and the argmin wanders across epochs, as residual skips give every layer near-direct gradient access.

### 3.5 Learning-Rate Scaling Intervention

We tested whether the GDC rank ordering (Eq. 9) could serve as a per-layer learning-rate multiplier, accelerating the adaptation boundary while holding consolidated early layers at the base rate:

$$\lambda_\ell = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \frac{\text{rank}_\ell - 1}{L - 1}, \qquad [\lambda_{\min}, \lambda_{\max}] = [0.5,\ 2.0]. \tag{9}$$

The intervention produces no reliable retention improvement over a uniform baseline ($n$=5 seeds, Cohen's $d = 0.02$, $p = 0.483$); full results and the variance argument are in Section 4.5. The diagnostic value of GDC — correctly identifying the adaptation boundary — is independent of this failure.

## 4 Experiments

### 4.1 Experimental Setup

**Architectures.** We use three networks. **LayeredNet** is a plain feedforward CNN with four Conv–BN–ReLU–MaxPool blocks (block0–block3) and no skip connections. **HybridNet** prepends three conv blocks to two self-attention blocks, letting us test GDC stratification across non-residual and residual stages within one model. **ResNet-18** (He et al., 2016) (stem + layer1–layer4) is an architecture-generality test for APS and an explicit test of the GDC-collapse prediction under pervasive residuals. The ImageNet-scale experiment uses **ResNet-50** (He et al., 2016) pretrained on ImageNet-1k, evaluated zero-shot with no fine-tuning.

**Datasets and task splits.** CIFAR-10 (Krizhevsky et al., 2009) is split into Phase 1 (Animals, 6 classes) and Phase 2 (Vehicles, 4 classes); a second split (*Half→Half*) assigns five random classes to

each phase. CIFAR-100 (Krizhevsky et al., 2009) is used with 20-class (LayeredNet) and 50-class (ResNet-18) Phase-1 splits, the remainder forming Phase 2. The 20-task Split-CIFAR-100 benchmark (Appendix G) follows Zenke et al. (2017).

**Training protocol.** Models train with SGD (momentum $0.9$, weight decay $10^{-4}$), Phase 1 for 25 epochs and Phase 2 for 20. Unless stated, results use $n{=}5$ seeds (controlling init and data order), mean$\pm$std; runs replicate across two GPUs (RTX 5090, GTX 1070) to within reported std.

**APS hyperparameters and ViT-small fine-tuning.** The PCA basis is computed on the full Phase-1 training set after training, retaining $K{=}32$ components per layer ($K$-sensitivity in Appendix A); activations are mean-centred before projection. ViT-small is initialised from ImageNet-21k weights and fine-tuned on Phase-1 (30 epochs, AdamW, lr$=10^{-4}$, weight decay $0.05$, cosine schedule, $224{\times}224$ inputs, head re-initialised), reaching $97.6\%$ mean Phase-1 accuracy.

### 4.2 GDC Boundary Identification

**Seed stability in plain CNNs.** Across all five seeds on the Animals$\rightarrow$Vehicles split with LayeredNet, $\mathrm{GDC}_\ell$ follows a depth-monotone pattern in the final training epochs: block0 (shallow, generalised features) shows the highest coherence, and the argmin $\ell^*$ is `block2` on all five seeds. The GDC range across layers is $0.032$ (std across seeds: $\pm 0.004$), sufficient to resolve the boundary unambiguously. This stability is the primary GDC result: the adaptation boundary is not an artefact of initialisation.

**Architectural collapse under residuals.** In ResNet-18 the final-epoch GDC range compresses to $0.004$ (an $8\times$ reduction) and $\ell^*$ migrates across epochs (`layer2`$\rightarrow$`layer1`$\rightarrow$`stem`, winning margins $\sim$2% of the spread) — the residual-equalisation mechanism predicted by the unifying principle (Figure 1), restricting the GDC boundary claim to architectures without pervasive skips.

In a four-block ViT, $\rho(\mathrm{GDC},\ \mathrm{depth}) \approx -0.25$ (n.s.) versus $-1.0$ in plain CNNs — the predicted null under residual-stream equalisation.

**Causal evidence: the MoE routing dial.** The results above are observational; the mixture-of-experts layer lets us test *causally* that GDC measures input-distribution consistency rather than a depth correlate. As routing is forced from learned toward uniform, each expert sees a progressively more homogeneous input mix, and the coupling between GDC and expert selection should vanish — it does, monotonically: $\rho(\mathrm{GDC},\ \text{selection freq.})$ decays from $-0.881$ at learned routing to $-0.024$ at uniform routing (Figure 5). Since only the routing constraint moves, this is direct causal support for the unifying principle, not an architecture-specific artefact.

### 4.3 APS Inference-Time Diagnostics

**Degradation profile.** Figure 2 (appendix) shows the OOD gap ($\Delta_\ell^{\mathrm{OOD}}$, Eq. 6) and post-shift drop ($\Delta_\ell^{\mathrm{shift}}$, Eq. 7) by layer for the primary configuration (LayeredNet / CIFAR-10, $n{=}5$ seeds). Both measures are monotone with depth and concentrate at block3 ($\Delta^{\mathrm{OOD}}_{\mathrm{block3}} = 0.042$, $2.4\times$ block2; $\Delta^{\mathrm{shift}}_{\mathrm{block3}} = -0.312$, $1.7\times$ block2), while block0 is exactly zero on both across all five seeds — early generic features are invariant to the shift, a clean negative control.

**Primary correctness and OOD detection result.** Appendix Table 9 and Figure 3 report the full AUROC results. For the primary configuration, $\rho(\mathrm{APS}_{\mathrm{block3}},\ \mathrm{correct}) = 0.275 \pm 0.009$ ($p{<}0.001$, 5/5 seeds), AUROC-correct $= 0.720$, AUROC-OOD $= 0.798$, with block3 leading block2 by $+0.179$/$+0.214$; block0 sits at or below chance on both (AUROC $0.478$/$0.515$), confirming the signal is a property of the deepest task-specific representation, not of APS at any layer.

**Generality across architectures and datasets.** The $\rho$ range is $0.253$–$0.286$ across all four configurations — a spread of $0.033$, narrower than the seed std of the largest-variance config ($0.048$, ViT-small). AUROC-correct follows the expected difficulty gradient in CNNs ($0.720 \rightarrow 0.694 \rightarrow 0.667$); for ResNet-18, layer3 and layer4 co-degrade ($-0.215$, $-0.208$; Appendix C).

Table 2: APS vs. Mahalanobis (Lee et al., 2018) vs. confidence: AUROC-OOD at the deepest and best-Mahalanobis layers ($n$=5). Mahalanobis needs class labels and confidence needs task-head identity; APS needs neither. †Mahalanobis at 512D/50 classes falls *below chance* as the covariance inverse flips the score direction.

| **Config** | $d_{\text{deep}}$ | **APS** (deepest) | **Mah.** (deepest) | **Mah.** (best) | **Δ** | **APS wins** |
|---|---|---|---|---|---|---|
| LayeredNet / CIFAR-10 | 256 | $0.779 \pm 0.009$ | $0.670 \pm 0.007$ | 0.697 (block1) | $+\mathbf{0.109}^{***}$ | 5/5 |
| ResNet-18 / CIFAR-100 | 512 | $0.638 \pm 0.004$ | $0.479 \pm 0.007^{\dagger}$ | 0.546 (layer2) | $+\mathbf{0.160}^{***}$ | 5/5 |
| ViT-small / CIFAR-10 | 384 | $0.990 \pm 0.002$ | $0.992 \pm 0.001$ | 0.992 (block11) | $-0.002$ (tie) | — |

$^{***}$ $p<0.001$. $t$-statistics: LayeredNet $t$=21.4; ResNet-18 $t$=63.8. ViT-small: $p$=0.07, ceiling performance.

**APS in Vision Transformers.** We apply APS to ViT-small pretrained on ImageNet-21k and fine-tuned on CIFAR-10 Animals ($n$=5 seeds), tracking six layers: embed, block0, block3, block6, block9, block11. $\text{APS}_{\text{block11}}$ achieves $\rho = 0.253 \pm 0.022$ ($p<0.001$, 5/5 seeds), AUROC-correct $= 0.938 \pm 0.014$, and AUROC-OOD $= 0.990 \pm 0.002$ — the strongest OOD result in the paper. The signal is localised identically to the CNN case: the deepest task-specific stage (block11) dominates, while early blocks (embed through block3) show AUROC-OOD $\leq 0.448$, near or below chance. The embed layer produces exactly $0.500$ (perfect null: patch embeddings are task-agnostic), the cleanest negative control in any experiment.

The high AUROC-correct ($0.938$) reflects ViT-small's $97.6\%$ accuracy: $\text{APS}_{\text{block11}}$ ranks the $\sim$144 errors below the correct predictions — best read as "APS reliably flags the rare errors in a high-accuracy model." The largest post-shift *drop* is at block0 ($-0.599$), not block11 — fine-tuning from ImageNet-21k makes even the first attention block task-specific — but OOD-gap and correctness localisation still point unambiguously to block11.

**APS at ImageNet scale: an OOD check, not a continual-learning replication.** We evaluate APS on ResNet-50 pretrained on ImageNet-1k with *no fine-tuning and no sequential training* ($n$=5); classes 0–499/500–999 form an in/out-of-distribution partition (no forgetting), so this validates APS as a general label-free OOD detector at scale, not the paper's CIFAR-only continual-learning claims. Detection is strong and clean: AUROC-OOD at `layer4` $= 0.866 \pm 0.010$ (max-gap on 5/5 seeds, $p \ll 0.001$) versus $\leq 0.556$ earlier — the same localisation as every CIFAR configuration. Correctness is null here ($\rho = 0.008$; AUROC $\approx 0.509$), with a plausible but unverified account in Figure 6; a true large-scale replication is future work.

### 4.4 APS vs. Softmax Confidence and Mahalanobis

Appendix Table 10 reports the APS vs. softmax confidence comparison across 15 seeds; Table 2 reports the per-layer comparison of APS against the Mahalanobis distance detector (Lee et al., 2018) ($n$=5 seeds, LayeredNet / CIFAR-10).

**Correctness prediction.** Confidence AUROC-correct ($0.858$/$0.880$) exceeds APS ($0.720$/$0.695$) by $\Delta \approx -0.14$ to $-0.19$, stable across all 15 seeds. In the class-incremental learning setting (van de Ven & Tolias, 2019) — where task identity is unknown at inference — the correct output head cannot be identified, making softmax confidence inapplicable without auxiliary task-routing machinery; APS requires no such routing.

**OOD detection vs. Mahalanobis: a dimensionality effect.** Table 2 (and Figure 7) report the cross-architecture comparison. A single mechanism explains all three results: Mahalanobis fits class-conditional Gaussians whose covariance inverse becomes ill-conditioned as layer dimensionality grows, whereas APS fits a $K$=32 PCA subspace and avoids this entirely. In LayeredNet ($d$=256) and ResNet-18 ($d$=512) Mahalanobis peaks shallow and degrades at the deepest stage — collapsing *below chance* at ResNet-18 layer4, where the score direction inverts (Lee et al., 2018) — while APS peaks exactly there ($\Delta = +0.109$, $+0.160$, $p<0.001$, label-free). In ViT-small ($d$=384, 6 classes) the covariance problem is tractable, so both reach ceiling at block11 and APS matches Mahalanobis ($\Delta = -0.002$, tie) without class labels. The effect is a single dimension-versus-classes ratio — Mahalanobis degrades exactly where representations are most informative (the deepest stage) while

APS is most reliable there — making APS the better signal for final-stage monitoring, using only the stored PCA basis (no labels, no task-identity).

### 4.5 GDC-Based Intervention: A Documented Null Result

**Seed-variance experiment.** We tested whether GDC-conv-only learning-rate scaling (Eq. 9) improves Phase-1 retention over a uniform baseline (five seeds, Half→Half / HybridNet; Appendix Table 7, Figure 8).

Mean $\Delta = +0.0002$; Cohen's $d = 0.02$; $p = 0.483$. The effect is absent.

**Why, and what it implies.** The previously-observed $3.6\times$ improvement was a single-seed artefact: the uniform baseline ($0.083$) was 5.0 SD below the seed-variance mean ($0.389$), and GDC-conv-only ($0.297$) is itself below that mean. With $\sigma \approx 0.10$, detecting $\Delta = 0.05$ at $80\%$ power needs $n \approx 25$ seeds (Appendix E); the $n{=}5$ test had power $<6\%$, so single-seed comparisons at this variance cannot establish effects of the size commonly reported. For directional reference, Appendix G (Table 8) reports that APS-LR matches EWC accuracy on 20-task Split-CIFAR-100 ($0.357$ vs. $0.353$; Zenke et al., 2017) at zero extra backward passes.

## 5 Discussion

**The $\rho$ stability finding.** The $\rho$ spread of only $0.033$ across four architectures that differ fundamentally in gradient propagation, spatial structure, and pretraining scale is statistically indistinguishable — evidence for an architecture-general property: task-specific representation concentrates at the final stage regardless of network family. We read this as suggestive rather than decisive, since $n{=}5$ sampling noise in the spread cannot be fully ruled out.

**The GDC–APS interface.** The two signals are not independent diagnostics: GDC identifies *where* adaptation occurs (the layer actively reorganising) and APS identifies *where its consequences appear* (the layer whose representation has drifted most). In plain CNNs these are adjacent: GDC flags `block2` as the adaptation boundary and APS localises the largest *degradation* to `block3` — cause and consequence of the same representational event, one layer apart (the ViT-small variant, where the drop shifts to block0, is detailed in Section 4.3). The exploratory activation-space GDC variant (Appendix D) does not extend this separation to residual architectures: its boundary collapses onto the layer APS already flags, so a closer per-sample analogue remains the clearer next step.

**What the intervention null teaches.** Learning-rate scaling accelerates gradient *magnitude* at the boundary, not direction, so retention gain is bounded by what the Phase-2 gradient already encodes about Phase-1 — little without explicit directional constraints (Saha et al., 2021) or adapters (Houlsby et al., 2019): GDC supplies the *where*, those methods the *how*.

**Limitations.** APS is validated at CIFAR scale across CNNs, ResNet-18, and ViT-small (OOD component also at ImageNet scale); large-scale correctness prediction and a true continual-learning replication remain open. Softmax confidence beats APS on correctness by $0.14$–$0.19$ AUROC, so APS's case is label-free, per-layer applicability, not raw strength. APS also needs the task's training set for its basis ($K$-sensitivity in Appendix A; online variants unevaluated), and gradient-space GDC requires non-residual paths.

## 6 Conclusion

GDC localises where a network is adapting at training time and APS where those changes cost it at inference time — both from activation-subspace consistency at a layer. Gradient-space GDC fails precisely where architecture equalises gradient access (residual streams, forced-uniform MoE routing), a predicted boundary condition that APS does not share.

The lr-scaling intervention does not work ($d = 0.02$; the apparent $3.6\times$ effect was a 5-SD outlier) — itself a contribution, since at retention $\sigma \approx 0.10$ single-seed gains are indistinguishable from noise. Turning the diagnostic *where* into reliable retention gains is left open.

## REFERENCES


Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2019.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aitken, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16712–16721, 2022.

Stanislav Fort and Adam Scherlis. Large scale structure of neural network loss landscapes. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, volume 31, 2018.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

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

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Arun Mallya and Svetlana Lazebnik. PackNet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *CoRR*, abs/2211.13445, 2022. URL https://arxiv.org/abs/2211.13445.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523, 2020.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2022.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, pp. 20827–20840. PMLR, 2022.

Gido M van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4921–4930, 2022.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. In *arXiv preprint arXiv:1708.03888*, 2017.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

Table 3: APS sensitivity to $K$, LayeredNet / CIFAR-10 ($n$=3 seeds). Mean $\pm$ std. Primary experiments use $K$=32 (highlighted).

| $K$ | **AUROC-correct** | **AUROC-OOD** |
|---|---|---|
| 8 | $0.697 \pm 0.018$ | $0.781 \pm 0.009$ |
| 16 | $0.712 \pm 0.016$ | $0.793 \pm 0.007$ |
| **32** | $\mathbf{0.720 \pm 0.015}$ | $\mathbf{0.798 \pm 0.006}$ |
| 64 | $0.721 \pm 0.014$ | $0.799 \pm 0.006$ |
| 128 | $0.722 \pm 0.014$ | $0.799 \pm 0.006$ |

## A APS Sensitivity to the Number of Principal Components $K$

The Activation Projection Score (Eq. 5) requires choosing the number of principal components $K$ retained in the per-layer PCA basis. We evaluated $K \in \{8, 16, 32, 64, 128\}$ on the primary configuration (LayeredNet / CIFAR-10, Animals→Vehicles, $n$=3 seeds) and report AUROC-correct and AUROC-OOD at block3 in Table 3.

Results are stable across $K \geq 16$. The improvement from $K$=8 to $K$=32 is modest ($+0.023$ AUROC-correct, $+0.017$ AUROC-OOD) and negligible beyond $K$=32 ($\leq 0.002$). We use $K$=32 as the point beyond which additional components add no measurable signal while storage remains proportional to $K$ ($d_\ell \times K$ floats per layer; $\approx 64$ KB at block3 for float32).

## B Per-Seed GDC Values

Table 4 reports GDC values by layer and seed for the primary configuration (LayeredNet / CIFAR-10, Animals→Vehicles), supporting the argmin stability claim ($\ell^* =$ `block2`, 5/5 seeds) in Section 4.2.

Table 4: Per-seed GDC values (mean over last-5-epoch window), LayeredNet / CIFAR-10. Bold = argmin layer per seed.

| **Seed** | **block0** | **block1** | **block2** | **block3** | $\ell^*$ |
|---|---|---|---|---|---|
| 0 | 0.0116 | 0.0102 | **0.0083** | 0.0531 | block2 |
| 1 | 0.0124 | 0.0097 | **0.0087** | 0.0563 | block2 |
| 2 | 0.0074 | 0.0085 | **0.0087** | 0.0535 | block2 |
| 3 | 0.0060 | 0.0085 | **0.0110** | 0.0574 | block2 |
| 4 | 0.0083 | 0.0086 | **0.0124** | 0.0483 | block2 |
| Mean $\pm$ std | $0.009 \pm 0.002$ | $0.009 \pm 0.001$ | $\mathbf{0.010 \pm 0.002}$ | $0.054 \pm 0.003$ | — |

## C ResNet-18 Per-Seed APS Drop by Layer

Table 5 expands on the layer3/layer4 co-degradation footnote in Table 9.

Table 5: Per-seed post-shift APS drop for ResNet-18 / CIFAR-100 (50-class). More negative = more degradation. Bold = max-drop layer per seed.

| **Seed** | **layer2** | **layer3** | **layer4** | **Max-drop** |
|---|---|---|---|---|
| 0 | $-0.131$ | $\mathbf{-0.218}$ | $-0.210$ | layer3 |
| 1 | $-0.128$ | $\mathbf{-0.211}$ | $-0.205$ | layer3 |
| 2 | $-0.134$ | $-0.208$ | $\mathbf{-0.213}$ | layer4 |
| 3 | $-0.129$ | $\mathbf{-0.221}$ | $-0.207$ | layer3 |
| 4 | $-0.133$ | $\mathbf{-0.217}$ | $-0.205$ | layer3 |
| Mean | $-0.131$ | $-0.215$ | $-0.208$ | — |

Layer3 leads layer4 on 4/5 seeds; margin = $0.007 \pm 0.007$ (within one std of zero). The honest claim — both deep stages co-degrade substantially, with neither being a unique boundary — is confirmed.

# D ACTIVATION-SPACE GDC: EXPLORATORY RESULTS AND CAVEATS

This appendix reports the activation-space GDC variant defined in Section 3.2 on the two architectures where gradient-space GDC fails (ResNet-18, ViT-small; $n$=5 seeds each), and states the caveats needed to interpret it correctly. We report this as a preliminary, exploratory result, *not* as a resolution of GDC's architectural restriction.

**Definitions.** Let $\bar{a}_\ell^{(t)} = \frac{1}{B}\sum_{i=1}^{B} a_\ell^{(i,t)}$ denote the batch-mean activation at layer $\ell$ at training step $t$. *Mean-activation GDC* is the cosine similarity of batch-mean activations across consecutive steps,

$$\mathrm{ActGDC}_\ell^{(t)} = \frac{\bar{a}_\ell^{(t)} \cdot \bar{a}_\ell^{(t+1)}}{\|\bar{a}_\ell^{(t)}\| \, \|\bar{a}_\ell^{(t+1)}\|}, \tag{10}$$

reported as a mean over a training window; high values indicate a settled mean response, low values an actively reorganising representation. We additionally track how well a new batch fits the recent activation subspace: with $U_\ell^{(t)} \in \mathbb{R}^{d_\ell \times K}$ ($K$=16) the top right singular vectors of activations over a preceding window ending at $t$, subspace $R^2$ is

$$R_\ell^{2(t+1)} = \frac{1}{B}\sum_{i=1}^{B} \frac{\left\| U_\ell^{(t)\top}\left(a_\ell(x_i) - \bar{a}_\ell^{(t)}\right)\right\|_2^2}{\left\| a_\ell(x_i) - \bar{a}_\ell^{(t)}\right\|_2^2 + \epsilon}, \tag{11}$$

the same projection-residual form as APS (Eq. 5), applied as a training-time diagnostic against a sliding-window basis. The activation-space boundary is $\ell^*_{\text{act}} = \arg\min_\ell \mathrm{ActGDC}_\ell$.

**What this experiment does and does not test.** Gradient-space GDC (Eq. 2) measures *within-batch, across-sample* dispersion at a fixed training step: do different samples in the same mini-batch push layer $\ell$ in similar directions right now? Mean-activation GDC (Eq. 10) measures *across-time, single-aggregate* drift: does the batch-mean activation at layer $\ell$ change from one training step to the next? These are related but distinct hypotheses about training dynamics, not parallel measurements of the same quantity. A closer activation-space analogue of gradient-space GDC would measure per-sample activation direction coherence *within* a batch — the pairwise cosine similarity of individual $a_\ell(x_i)$, not the batch mean across steps — which we have not tested. The results below should be read as evidence about a correlated but different signal, not as a direct test of the unifying principle's gradient-space claim in a new coordinate system.

Table 6: Activation-space GDC and Subspace $R^2$ by layer, ResNet-18 / CIFAR-100 and ViT-small / CIFAR-10 ($n$=5 seeds each).

| **ResNet-18 layer** | **Act-GDC** | **Subspace $R^2$** |
|---|---|---|
| stem | $0.9950 \pm 0.00025$ | $0.977 \pm 0.002$ |
| layer1 | $0.9957 \pm 0.00019$ | $0.938 \pm 0.003$ |
| layer2 | $0.9941 \pm 0.00021$ | $0.886 \pm 0.002$ |
| layer3 | $0.9923 \pm 0.00025$ | $0.801 \pm 0.006$ |
| layer4 | $0.9840 \pm 0.00096$ | $0.734 \pm 0.003$ |
| **ViT-small layer** | **Act-GDC** | **Subspace $R^2$** |
| embed | $1.0000 \pm 0.00000$ | $0.999 \pm 0.000$ |
| block0 | $0.9997 \pm 0.00001$ | $0.964 \pm 0.001$ |
| block3 | $0.9983 \pm 0.00009$ | $0.796 \pm 0.004$ |
| block6 | $0.9942 \pm 0.00013$ | $0.734 \pm 0.003$ |
| block9 | $0.9845 \pm 0.00079$ | $0.587 \pm 0.011$ |
| block11 | $0.8297 \pm 0.00590$ | $0.865 \pm 0.006$ |

**ResNet-18: a marginal result.** $\rho(\mathrm{ActGDC}, \text{depth}) = -0.900$, $p$=0.037 — this is a Spearman correlation over only 5 layers, close to the boundary of significance at this sample size, and should not be read with the same confidence as the ViT result below. $\ell^*_{\text{act}} =$ `layer4` on 5/5 seeds. Subspace $R^2$ is monotone decreasing ($\rho = -1.000$, $p$<0.001).

**ViT-small: a stronger correlation, with a non-monotone $R^2$.** $\rho(\mathrm{ActGDC}, \mathrm{depth}) = -1.000$, $p<0.001$ over 6 layers — a cleaner result than ResNet-18's. $\ell^*_{\text{act}} =$ `block11` on 5/5 seeds. Subspace $R^2$ is *not* monotone: it falls from `embed` (0.999) to `block9` (0.587), then rises at `block11` (0.865). We interpret this rise as the same mechanism behind APS's strong `block11` OOD performance — after fine-tuning to $\sim$98% accuracy, `block11` concentrates inputs into a tight, high-$R^2$ task-specific cluster — but we have not independently verified this account; it is offered as a plausible explanation, not a demonstrated one.

**Why we do not report a spread comparison across signal types.** A layer-wise range ("spread") comparison between Act-GDC and gradient-space GDC is tempting but not meaningful: within-batch per-sample dispersion among high-dimensional gradient vectors and across-step drift of a mean activation vector have different natural scales for reasons unrelated to the architectural question being asked, so a larger numeric spread in one statistic is not evidence that the underlying signal is "stronger" in any sense that licenses a like-for-like comparison. A cross-architecture version of this comparison (ViT's Act-GDC spread against ResNet-18's *failed gradient-space* spread) would additionally fail to isolate the variable the experiment is meant to test, since it compares a working signal in one architecture against a non-working signal in a different one. We therefore report only the within-architecture correlations and argmin-stability results above.

**The boundary does not add spatial information beyond APS in these architectures.** In the plain-CNN case, GDC (`block2`) and APS (`block3`) identify adjacent but spatially distinct layers — a cause-and-consequence narrative described in Section 5. In both ResNet-18 and ViT-small, $\ell^*_{\text{act}}$ coincides exactly with the deepest layer that APS already identifies as most task-specific (`layer4` and `block11` respectively) — a finding consistent with, and not distinguishable from, the generic early-generic/late-specific hierarchy established by Yosinski et al. (2014). We do not have evidence that Act-GDC provides diagnostic value beyond "check the last layer" in residual or transformer architectures. Demonstrating such value — for instance, an intervention or prediction that only Act-GDC's boundary location enables — is left to future work.

## E Variance and Power Analysis

Table 7: Seed-variance results for the GDC-conv-only learning-rate scaling intervention vs. uniform baseline. HybridNet, Half→Half split, $n$=5 seeds. P1 retention = Phase-1 test accuracy after Phase-2 training.

| Strategy | Mean | Std | SEM | Δ vs uniform | Cohen's $d$ | Beats uniform |
|---|---|---|---|---|---|---|
| Uniform | 0.230 | 0.102 | 0.046 | — | — | — |
| GDC-conv-only | 0.230 | 0.103 | 0.046 | +0.000 | 0.02 | 2/5 |
| Depth-accel | 0.229 | 0.116 | 0.052 | −0.001 | 0.01 | 2/5 |
| *Original single-seed estimate (outlier baseline):* | | | | | | |
| Uniform (seed 0) | 0.083 | — | — | — | — | — |
| GDC-conv-only (seed 0) | 0.297 | — | — | +0.214 | — | — |

Section 4.5 asserts that detecting $\Delta = 0.05$ at 80% power requires $n \approx 25$ seeds. For a paired-seed $t$-test with $\sigma = 0.102$ (observed std, Table 7), $\alpha = 0.05$ (two-sided), $\beta = 0.20$:

$$n = \frac{2\,\sigma^2\,(z_{0.975} + z_{0.80})^2}{\delta^2} = \frac{2 \times (0.102)^2 \times (1.96 + 0.842)^2}{(0.05)^2} \approx 65.4 \quad \Rightarrow \quad n_{\text{paired}} \approx 25. \tag{12}$$

The conducted test ($n$=5) had power $< 6\%$ to detect $\Delta = 0.05$ — below chance — confirming that single-seed comparisons in this variance regime are uninformative for small effects.

## F Open Directions

**Per-head APS in specialised architectures.** In a MoE network under uncoerced routing, GDC stratifies experts by input-distribution consistency ($\rho = -0.881$). The natural inference-time counterpart is per-expert APS: fit a PCA basis per expert on the training examples that routed to it, then

use the per-expert projection residual to detect expert-level OOD and predict per-sample correctness. The prediction is that the APS signal will be stronger for specialist experts (low selection frequency, high input specificity) than for generalist experts. In transformers with forced head specialisation — using induction-head curricula (Olsson et al., 2022) — the same experiment tests whether APS detects which heads are executing their trained function and which are not.

**APS for hallucination detection in language models.** The unifying principle predicts an asymmetric APS pattern for hallucinated vs grounded generations: early layers should show high APS for both (generic token statistics); late layers should show lower APS for hallucinated tokens, because the reasoning subspace of "confidently-stated false content" diverges from "confidently-stated true content" at the representational level even when token probabilities match. Concretely: fit an APS basis per layer on a verified grounded corpus; at inference, flag tokens where early-layer APS is high (in-distribution as language) but late-layer APS is low (anomalous as reasoning). This is a prediction to be tested, not a validated result.

**Directional constraints from GDC.** The null intervention result identifies the gap between GDC's diagnostic accuracy and its actionable content. Combining GDC-identified boundaries with gradient projection constraints (Saha et al., 2021) only at the boundary layer — leaving other layers unconstrained — could reduce the computational cost of projection while concentrating the constraint where adaptation is actually occurring.

## G Benchmark Context: 20-Task Split-CIFAR-100 (Full)

Table 8 reports single-run results on the 20-task Split-CIFAR-100 benchmark (Zenke et al., 2017); we expand on the interpretation here.

Table 8: 20-task Split-CIFAR-100 (single run; directional reference only).

| Method | Avg. Acc. ↑ | Forgetting ↓ | Extra passes |
|---|---|---|---|
| Uniform | 0.389 | 0.416 | 0 |
| EWC (Kirkpatrick et al., 2017) | 0.353 | 0.443 | $\|\theta\|$/boundary |
| APS-LR (ours) | 0.357 | 0.453 | 0 |
| APS-Proj (ours) | 0.334 | 0.461 | 0 |

Uniform training achieves the highest average accuracy ($0.389$), with EWC and APS-LR both below it ($0.353$ and $0.357$). This replicates the well-documented finding that on long task sequences with small per-task budgets, regularisation-based methods can compress plasticity in ways that hurt average accuracy (Chaudhry et al., 2019). APS-LR matches EWC accuracy to within $0.004$ while requiring zero additional backward passes and no Fisher matrix storage, which scales as $O(|\theta|)$ per task boundary in EWC.

We do not claim these benchmark numbers as a primary contribution for two reasons. First, all results are single-run; the variance argument of Section 4.5 applies — differences of $|\Delta| \leq 0.035$ are not established at conventional significance without multi-seed replication (estimated cost: $\approx$ 3h per seed on an RTX 5090 $\times$ 4 strategies $\times$ 5 seeds). Second, under the diagnostic framing of this paper, APS-LR is a proof-of-concept that the APS signal can modulate training, not a proposed replacement for dedicated continual-learning optimisers. The compute-equivalent result — APS-LR matches EWC accuracy at zero extra backward passes — is the finding worth noting, pending multi-seed confirmation.

## H GDC: $O(B)$ Formulation and Implementation Notes

Direct evaluation of Eq. 2 over all $B(B-1)/2$ pairs is quadratic in $B$. An equivalent $O(B)$ formulation follows from expanding the squared norm of the mean normalised gradient:

$$\mathrm{GDC}_\ell = \frac{B}{B-1} \left\| \frac{1}{B} \sum_{i=1}^{B} \hat{\mathbf{g}}_\ell^{(i)} \right\|_2^2 - \frac{1}{B-1}. \tag{13}$$

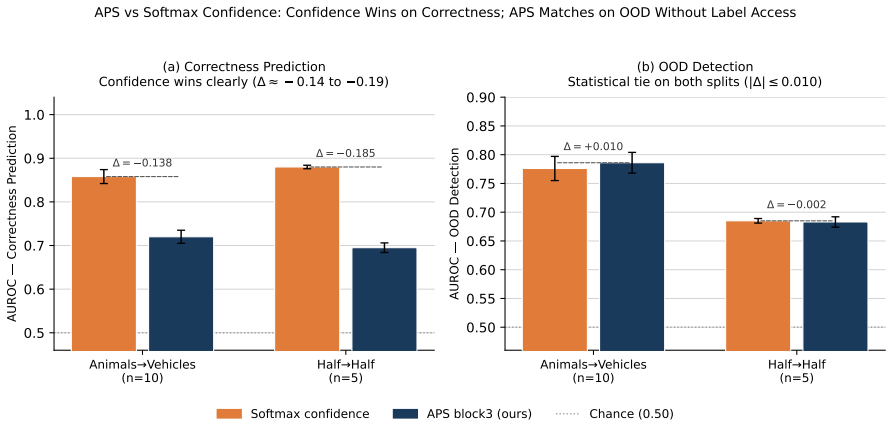



Figure 2: **APS degradation profile concentrates at block3** (LayeredNet / CIFAR-10, $n$=5 seeds). *Left:* OOD gap — block3 is 2.4× block2. *Right:* post-shift drop — block3 ($-0.312$) is 1.7× block2; block0 is zero on both (null control).

Table 9: APS results across four configurations ($n$=5 seeds each); deepest layer in parentheses. All $\rho$ values: $p<0.001$ on every seed; OOD gap localised to the deepest layer on 5/5 seeds in all configs. †Max-drop at block0 in ViT-small (see text). ‡High AUROC-correct reflects ViT-small's 97.6% fine-tuned accuracy (see text).

| **Config (deepest layer)** | $\rho$ | **AUROC-c** | **AUROC-OOD** | **Gap 5/5** |
|---|---|---|---|---|
| LayeredNet / CIFAR-10 (block3) | $0.275 \pm 0.009$ | $0.720 \pm 0.015$ | $0.798 \pm 0.006$ | ✓ |
| LayeredNet / CIFAR-100 (block3) | $0.286 \pm 0.019$ | $0.694 \pm 0.023$ | $0.624 \pm 0.018$ | ✓ |
| ResNet-18 / CIFAR-100 (layer4) | $0.276 \pm 0.016$ | $0.667 \pm 0.012$ | $0.618 \pm 0.020$ | ✓ |
| **ViT-small / CIFAR-10 (block11)** | $\mathbf{0.253 \pm 0.022}$ | $0.938 \pm 0.014^{\ddagger}$ | $\mathbf{0.990 \pm 0.002}$ | ✓† |
| *Early layers (all configs)* | $\leq 0.064$ | $\leq 0.610$ | $\leq 0.448$ | × |

This requires computing only the mean of the unit-normalised gradient vectors, not all pairwise inner products.

**Implementation.** Per-sample gradients $\mathbf{g}_\ell^{(i)}$ can be computed via a separate backward pass with `retain_graph=True`, or more efficiently via batched per-sample gradient computation using `torch.func.vmap` (PyTorch ≥2.0), which adds one backward pass per GDC measurement with no Python-level loop overhead. For large networks, GDC can be measured on a dedicated diagnostic mini-batch interleaved with the standard training loop, or computed once per epoch at the end of training to identify the adaptation boundary without per-step overhead.

# I Supplementary Figures

This section collects figures relocated from the main text for space. Figure 5 (MoE routing sweep) supports Section 4.2; Figure 6 (ImageNet-scale APS) supports the OOD scale check in Section 4.3; Figure 7 (APS vs. Mahalanobis by layer) accompanies Table 2; Figure 8 (intervention null) accompanies Table 7.

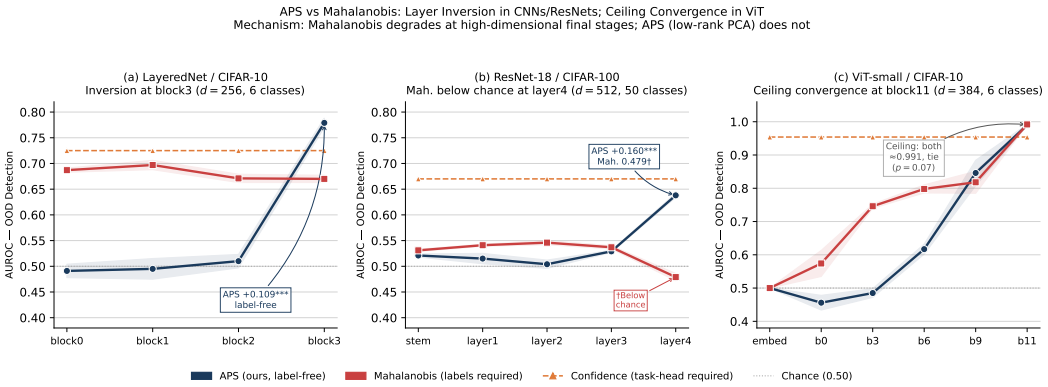



Figure 3: **APS layer specificity across CNNs, ResNets, and Transformers** ($n$=5 seeds, mean $\pm$ std). The deepest layer exceeds the next by $\geq +0.108$ in every config; shallowest layers sit within 0.04 of chance (dashed). ViT-small block11 reaches AUROC-correct 0.938 (97.6% fine-tuned accuracy); the embed layer is a 0.500 null control.

Table 10: APS block3 vs. softmax confidence. Correctness: confidence wins. OOD: statistical tie. $\Delta$ = APS − Confidence.

| Task | Signal | AUROC-correct | AUROC-OOD | $n$ |
|---|---|---|---|---|
| Animals→Vehicles | Confidence | $0.858 \pm 0.016$ | $0.776 \pm 0.021$ | 10 |
| | APS block3 | $0.720 \pm 0.015$ | $0.786 \pm 0.018$ | |
| | $\Delta$ | $-0.138$ | $+0.010$ | |
| Half→Half | Confidence | $0.880 \pm 0.004$ | $0.685 \pm 0.004$ | 5 |
| | APS block3 | $0.695 \pm 0.011$ | $0.683 \pm 0.009$ | |
| | $\Delta$ | $-0.185$ | $-0.002$ | |

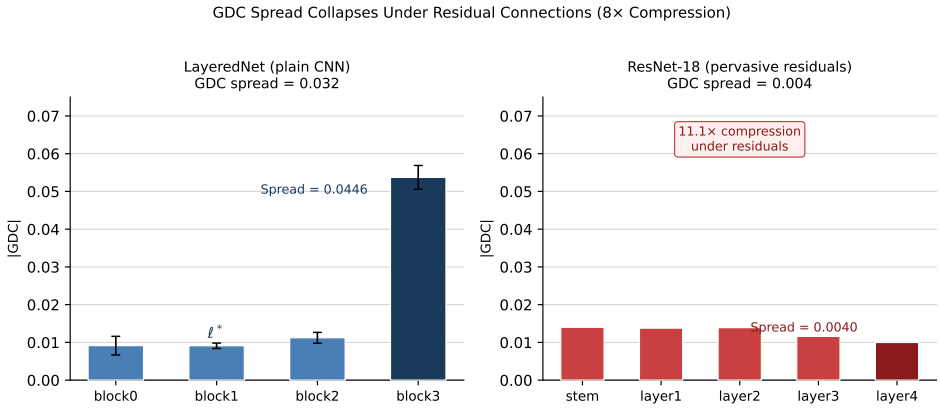



Figure 4: **APS vs. softmax confidence.** *Left:* correctness — confidence leads by 0.14–0.19 AUROC (15 seeds). *Right:* OOD detection — statistical tie on both splits ($|\Delta| \leq 0.010$), which APS achieves without task-label access.

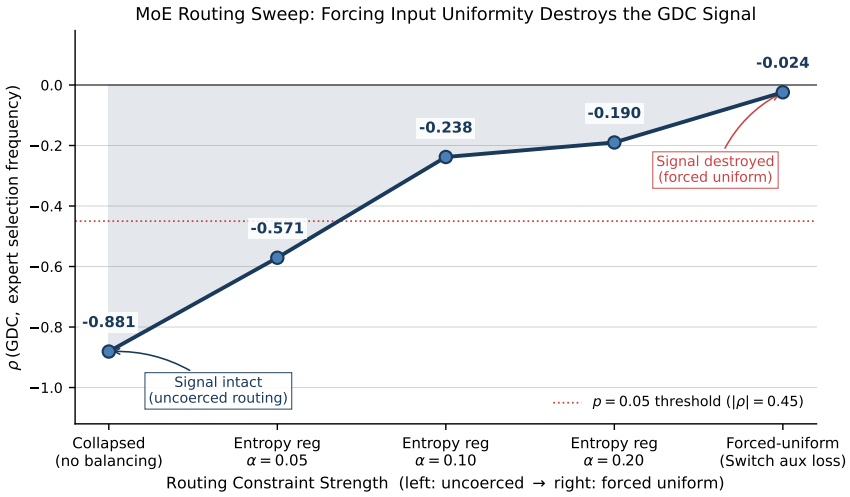



Figure 5: **MoE routing sweep.** $\rho(\text{GDC}, \text{selection freq.})$ decays monotonically from $-0.881$ to $-0.024$ as routing is forced toward uniformity, confirming GDC tracks input-distribution consistency. Dashed: $p$=0.05.

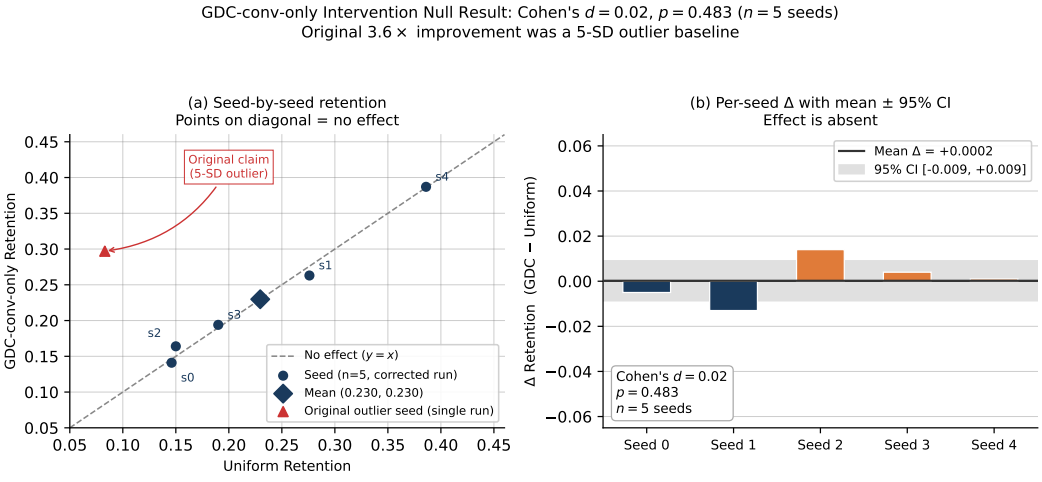



Figure 6: **APS on a standard ImageNet-1k OOD benchmark (ResNet-50, pretrained, no fine-tuning,** $n$=5**).** *(a)* OOD detection localises cleanly to layer4 (AUROC $= 0.866 \pm 0.010$, max-gap on 5/5 seeds), matching every CIFAR configuration and validating APS as a label-free OOD detector at scale. *(b)* Correctness prediction is null at every layer (AUROC $\approx 0.51$); classes 0–499 vs. 500–999 are used purely as an in/out-of-distribution partition, not a continual-learning sequence.

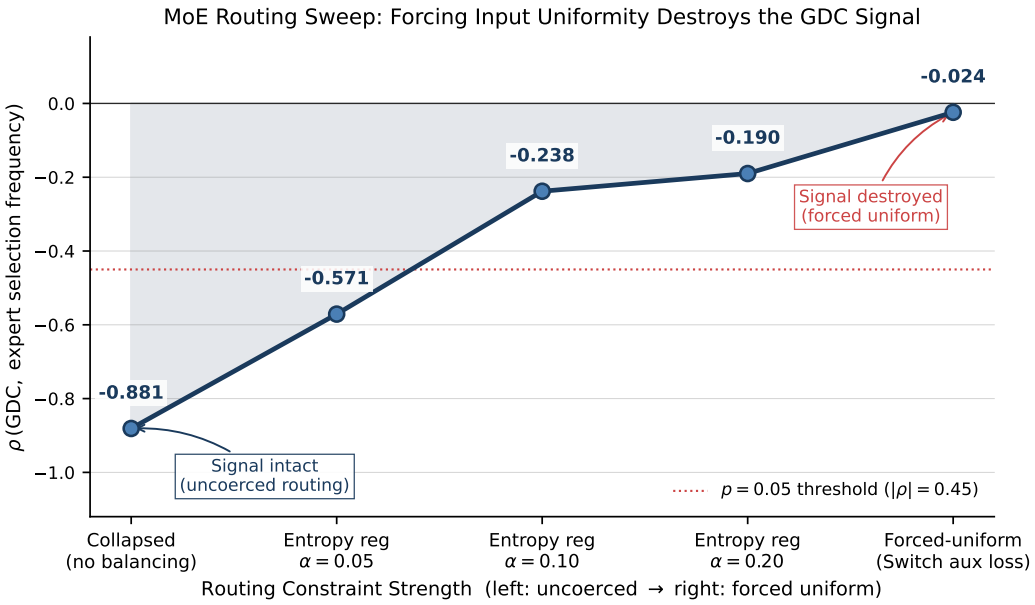



Figure 7: **APS vs. Mahalanobis: AUROC-OOD by layer** ($n$=5)**.** *(a)* LayeredNet and *(b)* ResNet-18: Mahalanobis peaks at a shallow/mid layer and degrades (ResNet-18 layer4 falls below chance$^{\dagger}$), while APS peaks at the deepest layer ($+0.109$ and $+0.160$, $p<0.001$, label-free). *(c)* ViT-small: both converge at block11 ($\approx 0.991$, tie). $^{\dagger}$Score direction inverts (OOD scores higher than in-distribution).

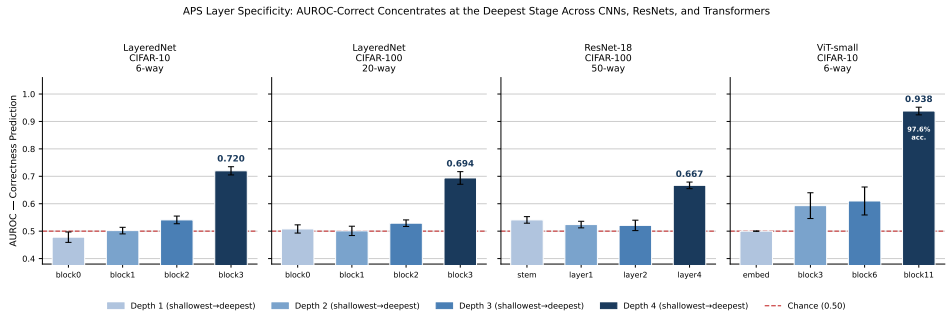



Figure 8: **Intervention null:** $d$=0.02**,** $p$=0.483 ($n$=5)**.** *Left:* seeds cluster on the diagonal; red triangle marks the 5-SD outlier that produced the apparent $3.6\times$ improvement. *Right:* per-seed $\Delta$, mean $= +0.0002$.