ATTRIBUTION EVALUATION SCORES DEPEND ON THE NETWORK BEING EXPLAINED

Anonymous authors
Paper under double-blind review

ABSTRACT

Attribution methods are often compared using faithfulness scores based on interventions and localisation scores based on human annotation. We show that both can depend strongly on the network being explained, even when the attribution method is unchanged. In fixed-width CIFAR-10 ResNets, rank-based faithfulness tracks how concentrated the measured behaviour is across channels (within-seed Spearman -0.83). An extreme-weight-decay model with one effective channel reaches only 18% accuracy yet receives a score of 1.000, because almost all channel–image pairs are tied at zero attribution and zero ablation effect. On ImageNet, using the same attribution and the same images, faithfulness rises from 0.595 to 0.880 across the first three stages of ResNet-18 and reaches 1.000 at the final stage, where the linear readout makes the metric exact by construction; activation magnitude alone scores 0.887 there. In a separate localisation study, two publicly available diabetic-retinopathy classifiers with the same ViT-B/16 architecture differ by $3.9\times$ in lesion localisation, more than the variation across five attribution methods within either classifier; one falls below a matched random baseline under every method. Gain-over-random normalisation, reproducibility normalisation, and support restriction do not remove the concentration effect. We therefore recommend reporting the explained layer and its distance from the readout, the concentration of the measured effect, and results across more than one layer and model.

1 INTRODUCTION

Consider a trained ResNet-18, a fixed attribution method, and a fixed set of images. Rank-based faithfulness is 0.595 at the first residual stage, 0.880 at the third, and 1.000 at the last. The attribution method and the images do not change; only the explained layer does. At the final stage the result is fixed in advance: the head is linear, so zeroing a channel changes the target logit by exactly its gradient-times-activation attribution, and the rank correlation is 1 by construction.

This example illustrates the central problem. A faithfulness score can reflect properties of the explained layer and network as much as properties of the attribution method. In three ImageNet CNNs, the score rises toward the output. In fixed-width CIFAR networks, it rises as the behaviour becomes concentrated in fewer channels. In two clinical classifiers with the same architecture, whether saliency maps localise lesions depends strongly on which checkpoint is explained.

The evaluation metrics studied here share a common structure: rank model components by their attributed scores, intervene on them in that order, and summarise the resulting response. Rank correlation with per-component ablation effects, deletion and insertion area under the curve (Petsiuk et al., 2018), and the gains of those curves over a random ordering all have this form.

Such metrics inherit the difficulty of the ranking problem defined by the network. If one component carries the behaviour and the rest carry almost none, the true effects consist of one large value and many near-zeros. Any attribution that places the large component first can then obtain a nearly perfect ranking. The score improves because the network has made the ranking problem easier, not necessarily because the attribution has become more faithful.

This effect is large even at fixed width. Across CIFAR-10 ResNets varied only in training length, weight decay, learning rate, and an interference penalty, rank-based faithfulness tracks the number

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.

of channels carrying the behaviour at within-seed Spearman -0.83. Under extreme weight decay, a network with one effective channel and 18% accuracy receives a score of 1.000 on all three seeds, compared with 0.690 for the fully trained network. In the degenerate network, 120 of 128 channel–image pairs have both zero attribution and zero measured effect. The metric therefore assigns a perfect score largely to tied zero-effect components.

Our main findings are:

- **The score rises toward the readout** (§6). With weights, images, and attribution fixed, rank faithfulness increases toward the output in three ImageNet CNNs. At the last ResNet stage, the linear readout makes the attribution equal to the zero-ablation effect, so the score is 1 by construction; activation magnitude alone scores 0.887 there (Appendix E).
- **At fixed width, concentration strongly predicts the score** (§4–5). Across a 21-network CIFAR-10 sweep, rank faithfulness correlates with the number of effective channels at -0.83 within seed. The association remains inside a 5.5-point accuracy band, while accuracy itself is only weakly associated with the score.
- **The explained checkpoint strongly affects localisation** (§7). Two publicly available retinopathy classifiers with the same ViT-B/16 architecture differ by $3.88\times$ in lesion localisation, more than the variation across five attribution methods within either classifier. One remains below a matched random baseline under all five methods.
- **Simple scalar corrections do not remove the effect** (§8). Retraining-based evaluation also shows no attribution signal on the layer studied (§9). We conclude with a reporting protocol in §10.

2 SETUP

Networks. We train CIFAR ResNets (He et al., 2016) with two residual blocks per stage (six blocks total) and **fixed width 16** on CIFAR-10. We vary accuracy only through training length (1–8 epochs), weight decay (5×10^{-4} to 10^{-1}), learning rate (0.05–0.20), and an optional interference penalty. The penalty is the training-time term $\lambda \,\mathrm{mean}_j(\log M_j - \log|A_j|)$, where $M_j = \Sigma_p|A_p|$ sums the magnitudes of the individual input–output paths through channel j, and A_j is their signed sum. It rewards agreement among a channel's paths and is scale-invariant by construction. We include it because it changes concentration through a mechanism distinct from ordinary optimisation quality. Width is fixed throughout: deletion and insertion curves are defined over the channel index, and at width 4 a 35% removal fraction leaves only one curve point, making the gain undefined.

Attribution. At the first residual block, the per-channel contribution is $A_j = \sum_p h_{jp}\,\partial y/\partial h_{jp}$, the standard hidden-unit attribution obtained by multiplying a unit's activation by its output gradient (Melander et al., 2026); Dhamdhere et al. (2019) use the same quantity as a baseline for conductance. Except in §7 and Appendix E, our conclusions concern how this attribution is evaluated rather than the merits of the attribution itself.

Effective channel count. `n_eff` is the smallest number of channels whose absolute attributions account for 90% of $\Sigma_j|A_j|$. Thus, $n_{\mathrm{eff}} = 1$ means that essentially all measured influence is concentrated in one channel.

Metrics.

- `rank_spearman`: Spearman correlation between $|A_j|$ and the measured $|\Delta\mathrm{logit}|$ from ablating channel j individually. We sample 64 channels per image and pool the resulting channel–image pairs.
- `del_gain`: area under the normalised target-logit curve while cumulatively ablating channels in descending $|A|$ order, minus the area under the corresponding random-order curve.
- `ins_gain`: the analogous insertion score, obtained by retaining only the top k channels, again relative to a random ordering.
- `roar`: accuracy after permanently masking the top k channels and retraining from scratch, compared with masking k random channels (Hooker et al., 2019).

Deletion and insertion curves are normalised by the target-logit drop produced by ablating the entire layer, $L_0 - L_{\mathrm{abl}}$. If ablating the whole layer does not move the target logit, the normalised curve is

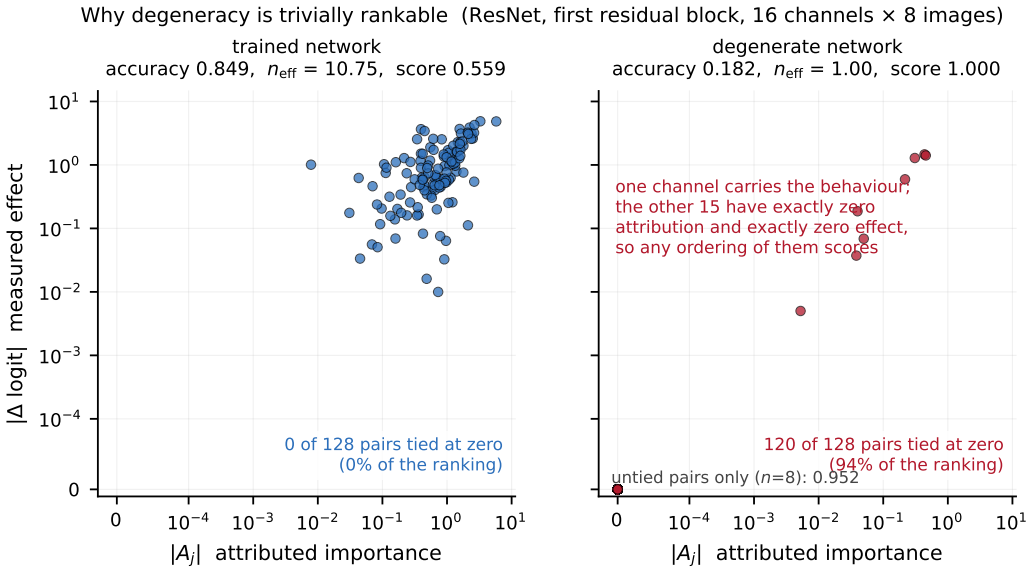


Figure 1: **Degenerate representations are easy to rank.** Attributed importance versus measured ablation effect for every channel–image pair at the first residual block of two networks trained from the same initialisation and differing only in weight decay. Both axes use a symmetric-log scale so that exact zeros remain visible. *Left:* the fully trained network, where the relation is diffuse and Spearman correlation is 0.559. *Right:* the degenerate network, where 120 of 128 pairs lie at the origin and the eight remaining pairs are nearly monotone (Spearman 0.952).

undefined; we exclude those images and report their count. Curve metrics are reported for a model only when at least half of its evaluation images are usable, a criterion fixed before analysis and applied uniformly. Thirty-six of 41 records satisfy this rule (39 distinct networks; the accuracy-band analysis reuses two sweep networks). The five failures are two of the three extreme-weight-decay networks and three interference-penalty networks, with 3 usable images of 8, compared with 8 of 8 for the best-trained network. `rank_spearman` is defined on the full evaluation set and therefore includes all 41 records. Whenever attribution and intervention targets enter the same reported statistic, we compute them on **disjoint** image sets, as motivated in Appendix C.

3 Why degeneracy inflates rank faithfulness

The effect is not a small-sample artefact. It follows directly from applying a rank correlation when most entries are tied. Holding architecture, initialisation, and data fixed, increasing weight decay until the network routes its behaviour through a single channel produces the following comparison:

	trained	degenerate
test accuracy	0.849	0.182
`n_eff`	10.75	1.00
pairs with $\lvert A\rvert = 0$ **and** $\lvert\Delta\text{logit}\rvert = 0$	0 of 128	**120 of 128**
rank faithfulness, all pairs	0.559	**1.000**
rank faithfulness, untied pairs only	0.559	0.952

In the degenerate network, 94% of the pairs contributing to the perfect score have neither attribution nor measured effect. Because rank correlation treats these ties as agreement, the score increases as the number of genuinely rankable components shrinks. This also explains why gain-over-random normalisation does not remove the effect (§8): both the attribution ordering and the random ordering become easier as the representation collapses. The support-restricted construction in §8 fails in the

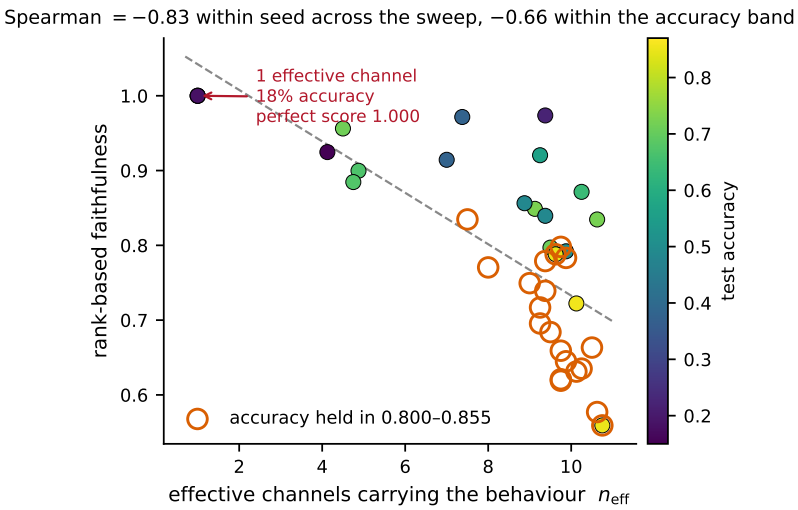


Figure 2: **Rank-based faithfulness increases as influence becomes concentrated.** Rank faithfulness versus the number of effective channels for all 41 records (39 distinct networks), coloured by test accuracy. Orange rings mark the 20 networks in the 5.5-point accuracy band. The perfect score belongs to a network with one effective channel and 18% accuracy.

opposite direction because concentrated networks contribute only a few channels to the correlation, making the estimate noisier.

4 Rank faithfulness tracks concentration

We evaluate 21 network conditions with three seeds per condition. Values below are reported as mean ± spread, where spread is the root-mean-square deviation from the seed mean (half the range when only two seeds are available). Curve metrics include only networks for which the curves are defined under the criterion in §2.

condition	accuracy	`n_eff`	rank	del gain	ins gain
full training	0.848 ± 0.005	10.17 ± 0.46	0.690 ± 0.096	0.145 ± 0.122	0.061 ± 0.197
short training	0.719 ± 0.007	9.75 ± 0.64	0.827 ± 0.022	0.260 ± 0.101	0.368 ± 0.059
very short	0.490 ± 0.005	9.38 ± 0.41	0.829 ± 0.027	0.272 ± 0.017	0.106 ± 0.189
interference penalty λ = 0.03	0.526 ± 0.110	8.83 ± 1.36	0.902 ± 0.022	1.680	1.262
interference penalty λ = 0.10	0.243 ± 0.099	6.96 ± 2.16	0.957 ± 0.023	0.628 ± 0.524	0.872 ± 0.444
high weight decay	0.685 ± 0.019	4.71 ± 0.16	0.914 ± 0.031	0.167 ± 0.071	0.506 ± 0.324
extreme weight decay	**0.181 ± 0.001**	**1.000 ± 0.000**	**1.000 ± 0.000**	0.467	0.467

The score moves broadly in the opposite direction from accuracy. The table below reports correlations computed within each seed and then averaged across seeds.

metric	vs accuracy	vs `n_eff`	spread across seeds
rank	−0.810	**−0.833**	0.034
del gain	−0.538	−0.260	0.252
ins gain	−0.233	−0.445	0.175

Appendix D plots all three metrics against n_eff. Deletion and insertion gains are already measured relative to a random ordering, yet both remain associated with concentration.

For rank faithfulness, the association with concentration is large relative to seed-to-seed variation: the within-seed correlation with n_eff is -0.83, while the spread across seeds is 0.034. The extreme condition is exactly reproducible. With weight decay high enough to collapse the network to one effective channel, n_eff=1.000 ± 0.000 and rank faithfulness is 1.000 ± 0.000 on all three seeds, despite 18.1% accuracy on a ten-class task.

The interference penalty from §2 provides a second route to the same pattern. It rewards agreement among a channel's paths but does not optimise the ranking metric directly. Increasing the penalty from 0.03 to 0.10 reduces n_eff from 8.8 to 7.0 and raises rank faithfulness from 0.902 to 0.957. Two different training interventions therefore increase concentration and increase the score.

On the best-trained networks, ins_gain is 0.061 ± 0.197 and is negative on two of the three seeds, so attribution-ordered insertion performs no better than random. At roughly 85% test accuracy, channel-level attribution ranking at this layer carries little usable insertion signal. This is an empirical counterpart to, rather than an instance of, the impossibility results of Bilodeau et al. (2024) for complete and linear methods.

5 The association persists at comparable accuracy

The full sweep ranges from 18% to 85% accuracy, so concentration and competence vary together. To separate them, we restrict attention to a 5.5-point accuracy band and vary only weight decay, training length, and learning rate. This subset contains 20 networks, two seeds per condition, with accuracy 0.800–0.855 and n_eff 7.5–10.8.

Per-condition values are reported in Appendix B.

metric	vs n_eff	partial, controlling accuracy	vs accuracy
rank	$-$**0.662** (p = 1.5×10^{-3})	$-$**0.595**	-0.363 (p = 0.12)
del gain	-0.236 (p = 0.32)	-0.313	+0.068 (p = 0.77)
ins gain	-0.269 (p = 0.25)	-0.263	-0.083 (p = 0.73)

Within this narrow accuracy range, concentration remains associated with rank faithfulness. The correlation with n_eff is -0.662 ($p = 1.5 \times 10^{-3}$), and the partial correlation controlling for accuracy is -0.595. By contrast, accuracy is only weakly associated with the score (Spearman -0.36, $p = 0.12$). The least accurate condition in the band (mean 0.8013) has the highest mean faithfulness, 0.803, whereas a condition at 0.8386 accuracy scores 0.611. Thus, two networks separated by only 3.7 accuracy points can differ by 0.19 in measured faithfulness in the direction predicted by concentration.

6 At ImageNet scale, the score rises toward the readout

The CIFAR sweep fixes the explained width at 16 channels, isolating concentration but not showing how the metric behaves across the larger layers used in standard vision models. We therefore repeat the measurement on ImageNet-pretrained networks with the same attribution and intervention protocol. The controlled experiment evaluates every stage of ResNet-18, ResNet-50, and ConvNeXt-Tiny on **100 ImageNet validation images per layer**, all usable at every stage. We also evaluate one mid-network layer from each of fourteen pretrained models on six images (eight for ResNet-18). A preliminary six-image version of the within-model sweep gave 0.477 rather than 0.314 for the 96-channel ConvNeXt stage, showing that small evaluation sets can substantially inflate early-layer estimates.

Across architectures, the score is strongly family-dependent. ResNet stages score 0.85–0.95, ConvNeXt stages 0.35 and 0.47, and transformer blocks 0.08–0.32, with tokens rather than channels as the ablated units in the transformer models. Across all fourteen models, n_eff correlates with

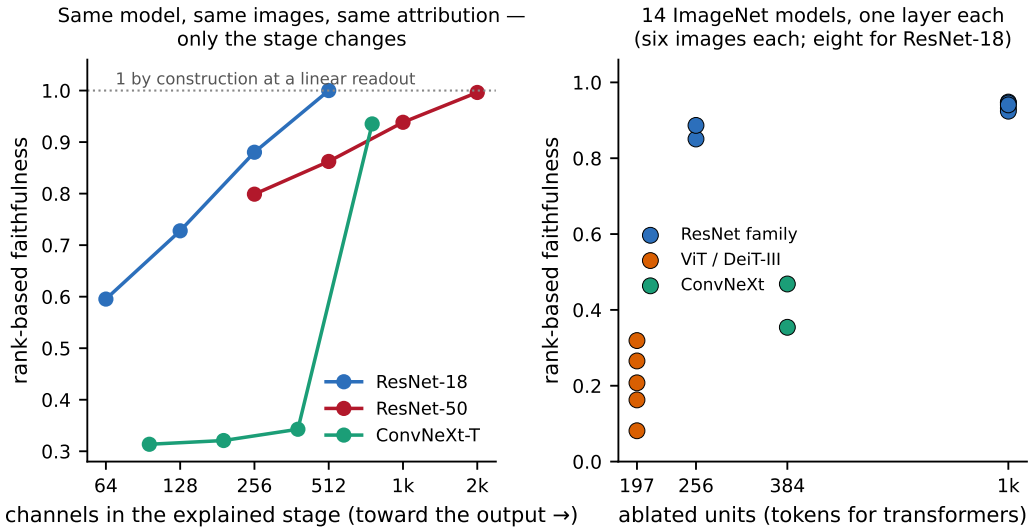


Figure 3: **Rank faithfulness rises toward the readout.** *Left:* the same model, images, and attribution method, with only the explained stage changed; each point uses 100 ImageNet validation images. The final ResNet stage sits immediately before a linear readout, where the score is 1 by construction. *Right:* one layer from each of fourteen ImageNet-pretrained models, coloured by family; for transformers, the ablated units are the 197 tokens rather than channels.

the score at Spearman +0.820, the opposite sign from the fixed-width CIFAR sweep. Within family, however, the association is not significant (ResNet, $n = 7$: +0.32, $p = 0.48$; ViT/DeiT, $n = 5$: -0.30, $p = 0.62$). Because architecture, width, and depth all vary in this comparison, we next hold the model fixed and change only the explained stage.

model	64	96	128	192	256	384	512	768	1024	2048
ResNet-18	0.595		0.728		0.880		**1.000**			
ResNet-50					0.799		0.862		0.938	**0.996**
ConvNeXt-T		0.314		0.321		0.343		0.935		

Rank faithfulness increases stage by stage in both ResNets. ConvNeXt-Tiny is nearly flat over its first three stages (0.314–0.343) and then rises to 0.935 at the final stage. Within each architecture, the widest stage is also the deepest, so these data identify position in the network rather than width alone. At the final ResNet stage, the head consists of global average pooling followed by a linear layer. Zeroing channel j therefore changes the target logit by exactly A_j, making the rank correlation 1 by construction (0.999999 for ResNet-18 and 0.996 for ResNet-50). ConvNeXt's last stage is followed by LayerNorm and scores 0.935. Thus, the layer most commonly used for explanation is also the layer at which this metric is structurally easiest to satisfy.

Across the twelve model–layer pairs, the partial correlation between stage width (which is collinear with depth) and the score, controlling for `n_eff`, is +0.696. The partial correlation between `n_eff` and the score, controlling for width, is -0.446, matching the sign of the fixed-width effect. Neither estimate is significant with only twelve layers from three models. ConvNeXt's final transition helps separate the factors: `n_eff` rises from 201.5 to 377.6 while the score rises from 0.343 to 0.935, so the change follows network position rather than concentration.

The rise is carried by the activation term. Repeating the ResNet-18 sweep on the same 100 images with four alternative rankings (Appendix E) shows where the effect enters. Integrated Gradients applied in activation space rises from 0.467 at the first stage to 1.000 at the last, closely tracking gradient-times-activation. Activation magnitude alone, which is identical for every output and therefore carries no information about which output is being explained, rises from 0.465 to 0.887. Gradient magnitude alone remains between 0.33 and 0.41 at every stage, while a random

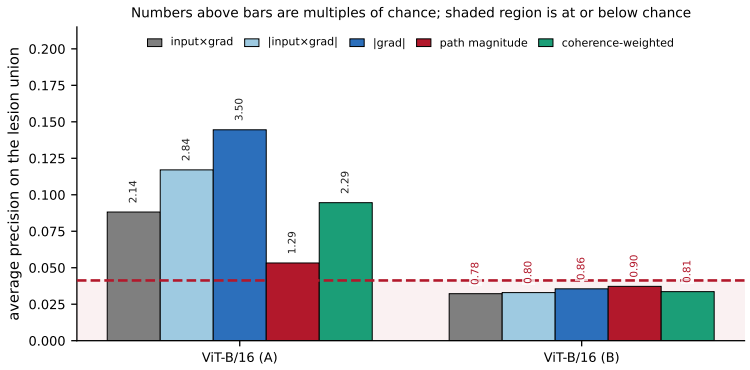


Figure 4: **Checkpoint choice strongly changes lesion localisation.** Mean lesion average precision over 81 annotated IDRiD images for five attribution variants on two ViT-B/16 classifiers with identical architecture and preprocessing. Numbers above the bars give multiples of the matched random baseline (0.0413); the shaded region is at or below that baseline.

ranking stays within 0.03 of zero. Near the readout, channel activation magnitude already predicts the effect of ablation, and the metric rewards any attribution that inherits this structure.

Interpretation. The fixed-width experiments show that concentration moves the score when architecture and layer width are held fixed (§4–5). The within-model ImageNet sweep shows that network position can move it even when weights, images, and attribution method are fixed. In ResNet-18, changing only the explained stage moves rank faithfulness from 0.595 to 1.000. A single-layer faithfulness number therefore mixes attribution quality with properties of the layer on which it is measured.

7 Localisation depends on the explained checkpoint

The preceding experiments concern faithfulness, where attributions are scored against interventions on the network. A second common evaluation family compares saliency maps with human annotation. We test whether the same dependence on the explained model appears there. We evaluate two publicly available diabetic-retinopathy classifiers from the Hugging Face Hub (`Kontawat/vit-diabetic-retinopathy-classification` and `rafalosa/diabetic-retinopathy-224-procnorm-vit`, denoted A and B) using five attribution variants. Localisation is measured against expert pixel masks for microaneurysms, haemorrhages, and hard and soft exudates on all 81 annotated IDRiD fundus images (Porwal et al., 2018). Both classifiers use the ViT-B/16 architecture and resize the full image to 224×224 without cropping, so the masks remain aligned with the model input. We report average precision of the saliency map on the union of lesion pixels. A uniformly random map on the same 81 images gives 0.0413 ± 0.0001 over ten seeds.

The best localisation score differs by $3.88\times$ between the two classifiers. Within a classifier, the five attribution methods span $2.71\times$ for A and $1.16\times$ for B. Classifier A reaches $3.50\times$ the random baseline under gradient magnitude. Classifier B remains below the matched random baseline under all five variants ($0.78\times$–$0.90\times$), despite sharing A's architecture and preprocessing.

This variation would be invisible in a single-model evaluation. The same attribution method can appear successful on one checkpoint and unsuccessful on another, even when architecture and preprocessing are fixed. Localisation and intervention-based faithfulness use different reference signals, yet both depend strongly on the model being explained.

8 The tested scalar corrections do not remove the effect

Three natural scalar corrections do not remove the effect; Appendix C gives the full analysis. Gain-over-random normalisation leaves deletion and insertion associated with concentration at -0.26 and

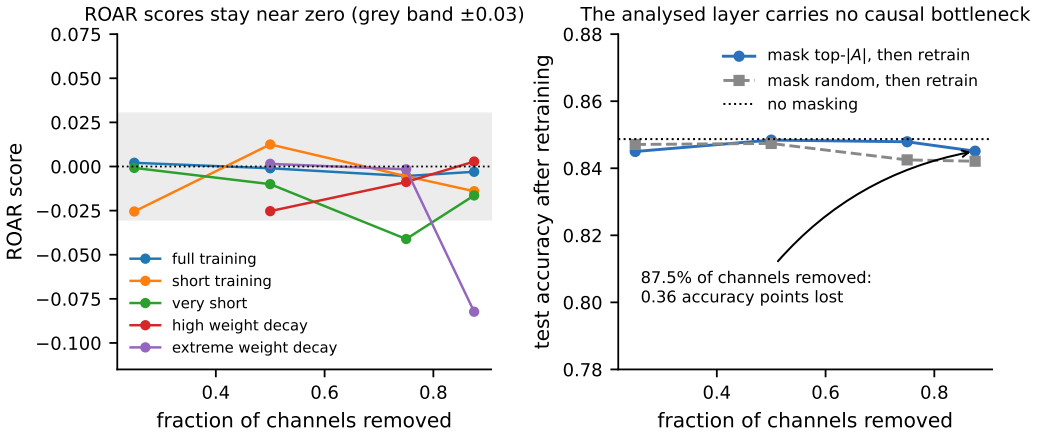


Figure 5: **Retraining-based evaluation shows little attribution signal on this layer.** *Left:* retraining score versus removal fraction for each training condition; the grey band marks ±0.03. *Right:* accuracy after masking k of the 16 channels in the first residual block and retraining, compared with the unmasked baseline. Removing 14 of 16 channels changes accuracy by only 0.36 percentage points.

−0.45, because degeneracy makes both the attribution ordering and the random ordering easier. Normalising by the target's cross-image reproducibility also leaves the association intact, since the reproducibility ceiling itself correlates with concentration at −0.719. Restricting the ranking to channels that carry measured effect reverses the sign, because concentrated networks contribute fewer channels and therefore noisier rank estimates. The useful part of this construction is its support criterion: for sufficiently concentrated networks, the appropriate result is **undefined** rather than a near-perfect score.

9 Retraining-based evaluation shows little signal on this layer

condition	k = 4	k = 8	k = 12	k = 14
full training	+0.0021	−0.0010	−0.0054	−0.0030
short training	−0.0255	+0.0125	−0.0055	−0.0140
very short	−0.0008	−0.0100	−0.0411	−0.0164
high weight decay	—	−0.0253	−0.0088	+0.0028
extreme weight decay	—	+0.0015	−0.0016	−0.0823

Here k of 16 channels are masked before training from scratch. The reported score is accuracy under random masking minus accuracy under attribution-guided masking.

Across five single-seed networks, ROAR shows little attribution signal at any removal fraction from 25% to 87.5%: all scores lie within ±0.09 and most within ±0.03. The underlying retraining result is itself informative. For the best-trained network, masking **14 of 16 channels** in the first residual block before retraining changes accuracy from 0.8487 to 0.8451, a drop of only 0.36 percentage points. Thus, the architecture can recover almost all of its performance without most channels in this layer, even though single-channel ablations in the trained network are not all zero. By contrast, the ranking metrics assign faithfulness scores between 0.690 and 1.000 on the same layer.

The three evaluations therefore answer different questions on the same attributions and networks. Ranking metrics report high faithfulness, retraining-based evaluation finds little separation from random, and retraining after removing 87.5% of the channels shows that the layer is highly redundant. This complements Hooker et al. (2019), who found many saliency methods indistinguishable from random under ROAR, by identifying a network property that can make ranking-based scores large even when retraining-based evaluation is uninformative.

10 REPORTING PROTOCOL

1. **Report the explained layer, its distance from the readout, and results from more than one layer.** In ResNet-18, changing only the stage moves rank faithfulness from 0.595 to 1.000; immediately before a linear head, the score is 1 by construction.
2. **Report concentration** (`n_eff` or an equivalent measure) with any ranking-based score.
3. **Use disjoint image sets for attribution and intervention targets** when the same statistic combines both, and compare networks at matched layer position and concentration when possible.
4. **Return "undefined" when too few components carry measurable effect** for a meaningful ranking; the tested scalar normalisations do not provide a general correction (§8).
5. **Pair ranking metrics with retraining-based evaluation** and report the removal fraction.

11 RELATED WORK

Evaluation of attribution. Hooker et al. (2019) introduced RemOve-And-Retrain and found many saliency methods no better than random; Rong et al. (2022) later showed that removal itself can leak information. Randomisation tests have also revealed explanations that are insensitive to model parameters or labels (Adebayo et al., 2018; Sixt et al., 2020), while Bilodeau et al. (2024) establish limits for complete and linear methods. Our result concerns the evaluator: a high score can arise from the structure of the explained network even when the attribution rule is unchanged.

Localisation against annotation. Saliency maps for retinopathy classifiers are increasingly compared with lesion masks (Verma et al., 2026), including the IDRiD annotations used in §7. Related histopathology work finds that measured faithfulness varies strongly with architecture and task (Jamshidi Idaji et al., 2026). Our two ViT-B/16 graders hold architecture fixed, yet one falls below the matched random baseline under every attribution method, isolating checkpoint choice as a substantial source of variation.

Perturbation confounds. Known problems with perturbation-based evaluation often arise from the intervention: masking can create out-of-distribution inputs or leave residual signal (Hsieh et al., 2026; Rong et al., 2022), and rankings can change with the perturbation strategy (Nieradzik et al., 2024; Poolpol et al., 2026). The factor studied here is different. It is a property of the explained layer and persists under a fixed intervention operator. Standard randomisation checks need not detect it because top-down randomisation can preserve the activation scales through which it acts (Binder et al., 2022).

Ordered-intervention metrics. Deletion and insertion originate with RISE (Petsiuk et al., 2018); related evaluations compare attributed importance with activation-patching effects in attribution patching and circuit analysis (Nanda, 2023; Syed et al., 2024; Hanna et al., 2024). Our analysis applies whenever evaluation reduces to ranking components by attribution and comparing that order with intervention effects, including hidden-unit attributions and contribution decompositions (Dhamdhere et al., 2019; Sundararajan et al., 2017; Melander et al., 2026).

12 CONCLUSION

Attribution evaluation scores depend on the model and layer being explained. Rank faithfulness increases as measured influence becomes concentrated and as the explained layer approaches the readout; immediately before a linear readout, the score is 1 by construction. Localisation scores also vary substantially across checkpoints with the same architecture. Reporting layer position, concentration, and results across multiple models and layers makes these dependencies explicit instead of folding them into one number.

AI USE STATEMENT

GPT was used only to aid or polish the writing of this manuscript.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31*, 2018. arXiv:1810.03292.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024. doi: 10.1073/pnas.2304406120.

Alexander Binder, Leander Weber, Sebastian Lapuschkin, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. *arXiv preprint arXiv:2211.12486*, 2022.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron? In *International Conference on Learning Representations*, 2019. arXiv:1805.12233.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems 32*, 2019. arXiv:1806.10758.

Chia-Ying Hsieh, Hsin-Yuan Fang, and Chun-Shu Wei. Aim: Adversarial information masking for faithfulness evaluation of saliency maps. *arXiv preprint arXiv:2605.16905*, 2026.

Mina Jamshidi Idaji, Julius Hense, Tom Neuhäuser, Augustin Krause, Yanqing Luo, Oliver Eberle, Thomas Schnake, Laure Ciernik, Farnoush Rezaei Jafari, Reza Vahidimajd, Jonas Dippel, Christoph Walz, Frederick Klauschen, Andreas Mock, and Klaus-Robert Müller. Beyond attention heatmaps: How to get better explanations for multiple instance learning models in histopathology. *arXiv preprint arXiv:2603.08328*, 2026.

Joshua Brendan Melander, Zaki Alaoui, Shenghua Liu, Surya Ganguli, and Stephen A. Baccus. Causal interpretation of neural network computations with contribution decomposition. In *International Conference on Learning Representations*, 2026. arXiv:2603.06557.

Neel Nanda. Attribution patching: Activation patching at industrial scale, 2023. URL `https://www.neelnanda.io/mechanistic-interpretability/attribution-patching`.

Lars Nieradzik, Henrike Stephani, and Janis Keuper. Reliable evaluation of attribution maps in cnns: A perturbation-based approach. *arXiv preprint arXiv:2411.14946*, 2024.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018. arXiv:1806.07421.

Peachapong Poolpol, Henrik H. J. Detjen, and Eike Petersen. Faithful faithfulness evaluations: Challenges & pitfalls learned from a breast mri case study. *arXiv preprint arXiv:2609.25978*, 2026.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018. doi: 10.3390/data3030025.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *PMLR*, 2022. arXiv:2202.00449.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 9046–9057, 2020. arXiv:1912.09818.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 3319–3328, 2017. arXiv:1703.01365.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024. arXiv:2310.10348.

Abhishek Verma, Anila Krishna, Abhishek Gajanan Bankar, and Juan Miguel Lopez Alcaraz. Explainable diabetic retinopathy classification using vision foundation models. *arXiv preprint arXiv:2608.28207*, 2026.

A PER-CONDITION FIGURES FOR THE FIXED-WIDTH SWEEP

condition	accuracy	n_eff	rank	del gain	ins gain
full training	0.848 ± 0.005	10.17 ± 0.46	0.690 ± 0.096	0.145 ± 0.122	0.061 ± 0.197
short training	0.719 ± 0.007	9.75 ± 0.64	0.827 ± 0.022	0.260 ± 0.101	0.368 ± 0.059
very short	0.490 ± 0.005	9.38 ± 0.41	0.829 ± 0.027	0.272 ± 0.017	0.106 ± 0.189
interference penalty λ = 0.03	0.526 ± 0.110	8.83 ± 1.36	0.902 ± 0.022	1.680	1.262
interference penalty λ = 0.10	0.243 ± 0.099	6.96 ± 2.16	0.957 ± 0.023	0.628 ± 0.524	0.872 ± 0.444
high weight decay	0.685 ± 0.019	4.71 ± 0.16	0.914 ± 0.031	0.167 ± 0.071	0.506 ± 0.324
extreme weight decay	$\mathbf{0.181 \pm 0.001}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.467	0.467

B PER-CONDITION FIGURES FOR THE ACCURACY BAND

condition	accuracy	n_eff	rank
lr 0.20, 8 ep	0.8546	9.8	0.707 ± 0.072
wd 2×10^{-3}, 8 ep	0.8487	9.9	0.645 ± 0.014
wd 5×10^{-4}, 8 ep	0.8453	10.2	$\mathbf{0.674 \pm 0.114}$
lr 0.20, wd 2×10^{-3}	0.8439	9.4	0.700 ± 0.016
lr 0.05, wd 5×10^{-3}	0.8396	9.8	0.709 ± 0.089
lr 0.05, 8 ep	0.8386	10.2	$\mathbf{0.611 \pm 0.034}$
wd 5×10^{-3}, 8 ep	0.8340	9.6	0.761 ± 0.022
wd 5×10^{-4}, 6 ep	0.8263	10.1	0.641 ± 0.022
wd 5×10^{-3}, 6 ep	0.8218	9.1	0.722 ± 0.027
wd 10^{-2}, 8 ep	0.8013	7.8	$\mathbf{0.803 \pm 0.032}$

C SCALAR CORRECTIONS IN FULL

Gain over random does not correct the effect. del_gain and ins_gain already subtract a random ordering, yet their correlations with n_eff remain -0.26 and -0.45. Degeneracy makes both the attribution-ordered and random-ordered problems easier, so subtracting the two curves does not produce a difficulty-independent quantity.

Normalising by the target's own reproducibility also leaves the effect intact. Let t_A and t_B be per-channel intervention effects measured on two disjoint image sets. The cross-image reproducibility $\text{ceiling} = \text{spearman}(t_A, t_B)$ directly measures how stable the target ranking is. A natural

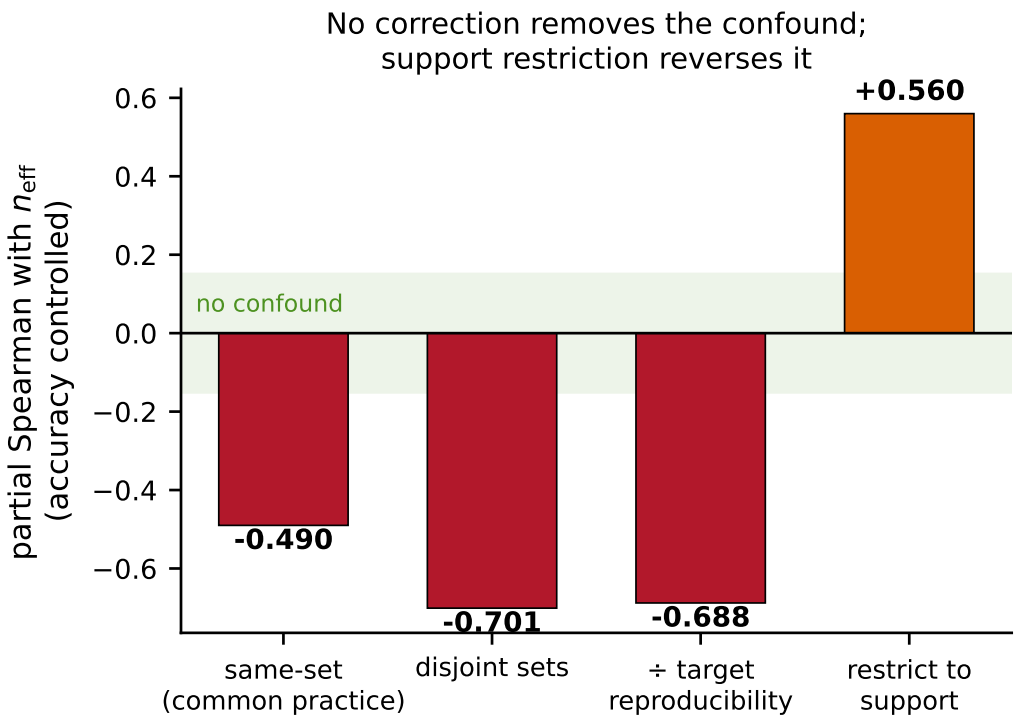


Figure 6: **The tested scalar corrections do not remove the concentration effect.** Partial Spearman correlation with `n_eff`, controlling for accuracy, under four evaluation protocols. The green band marks values near zero. Same-set evaluation is the common protocol; support restriction reverses the sign rather than eliminating the association.

correction is then $\text{achieved}/\text{ceiling}$, where achieved compares attributions from set A with intervention targets from set B.

quantity	vs `n_eff`	vs accuracy	partial, controlling accuracy
same-set rank Spearman	−0.842	−0.451	−0.490
cross-set achieved	−0.776	−0.490	−0.701
the ceiling itself	−0.719	−0.477	−0.657
normalised	−0.648	−0.284	−0.688

The ceiling itself correlates with concentration at −0.719. Degenerate networks have a ceiling of exactly 1.000, so division leaves them unchanged; well-performing networks have a ceiling near 0.66, so normalisation inflates their scores. The range narrows, but the ordering remains.

The same-set row has a weaker partial correlation (−0.490) than the disjoint-set row (−0.701), but this does not make it the preferable protocol. Shared sampling noise inflates every model's score and dilutes the measured association with `n_eff`. Same-set evaluation therefore masks part of the confound rather than removing it.

The same comparison reveals a separate sampling effect. Computing attributions and intervention targets on the *same* images can substantially inflate the correlation: in one condition the same-set score is 0.950, compared with 0.444 under disjoint sets. Because the two quantities share image-specific noise, we use disjoint sets whenever they enter the same statistic.

Restricting the ranking to the support of the measured effect reverses the association. Because much of the free credit comes from near-zero channels, one natural response is to rank only channels that carry substantial measured effect. We order channels by their effect on a held-out set, take the

smallest set accounting for 90% of the total, compute the correlation within that set, and return *undefined* when fewer than five channels remain.

With 20 images per set and three seeds, this leaves 24 models:

metric	n	vs n_eff	vs accuracy	partial, controlling accuracy
all channels, same set	24	−0.625	−0.603	−0.309
all channels, disjoint sets	24	−0.594	−0.535	−0.312
support-restricted, disjoint sets	21	**+0.499**	+0.330	**+0.560**

The association changes sign. A concentrated network contributes only about five channels to the correlation, producing a noisier estimate, whereas a broader network contributes about thirteen and tends to score higher. With six images per set, the same construction gives a partial correlation of +0.055, close to the point at which the two effects balance. The unrestricted metric is itself sensitive to evaluation-set size (partial correlation −0.512 with six images and −0.312 with twenty), so faithfulness numbers obtained with different evaluation budgets are not directly comparable.

The support criterion nevertheless captures one useful case: all degenerate networks with n_eff=1 are declared **undefined**. When too few components carry measurable effect, channel ranking does not provide a meaningful evaluation target.

D ALL THREE RANKING METRICS AGAINST CONCENTRATION

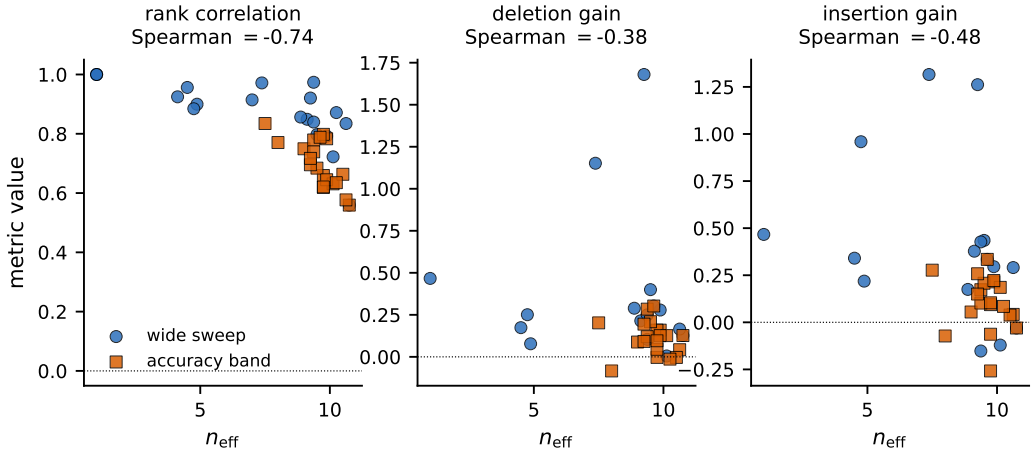


Figure 7: **All three ranking metrics vary with concentration.** Each metric is plotted against n_eff across both sweeps. The figure reports pooled Spearman correlations; the main-text tables report correlations computed within seed and then averaged. Deletion and insertion gains already subtract a random ordering but remain associated with concentration.

E FIVE ATTRIBUTIONS ACROSS THE RESNET-18 STAGES

Rank-based faithfulness on the same 100 ImageNet validation images and with the same intervention protocol as §6, changing only the attribution. grad×act is the attribution used in the main text and reproduces those values exactly. IG is Integrated Gradients in activation space with eight steps. |act| is summed channel activation magnitude and does not depend on the output; |grad| is summed output-gradient magnitude; random is a uniformly random ranking. Rows correspond to the four

ResNet-18 stages, labelled by channel count. The 512-channel stage feeds a linear readout, where grad×act equals the zero-ablation effect.

stage width	grad×act	IG	|act|	|grad|	random
64	0.595	0.467	0.465	0.405	0.002
128	0.728	0.548	0.536	0.348	−0.023
256	0.880	0.702	0.620	0.333	0.012
512	1.000	1.000	0.887	0.400	0.009