Feature Attribution Stability Suite: How Stable Are Post-Hoc Attributions?

Kamalasankari Subramaniakuppusamy[1]; Jugal Gajjar[1]
[1]The George Washington University, Washington D.C

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

PROBLEM STATEMENT

1) Predictions change, but benchmarks don't filter – stability is measured even when the model's output has already flipped, making scores meaningless.

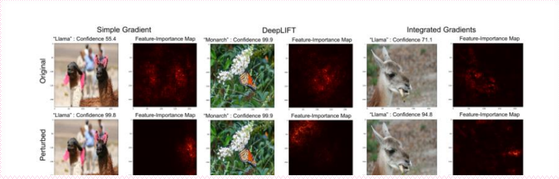
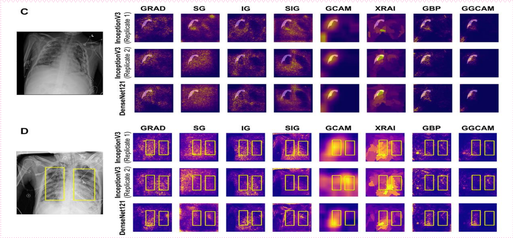


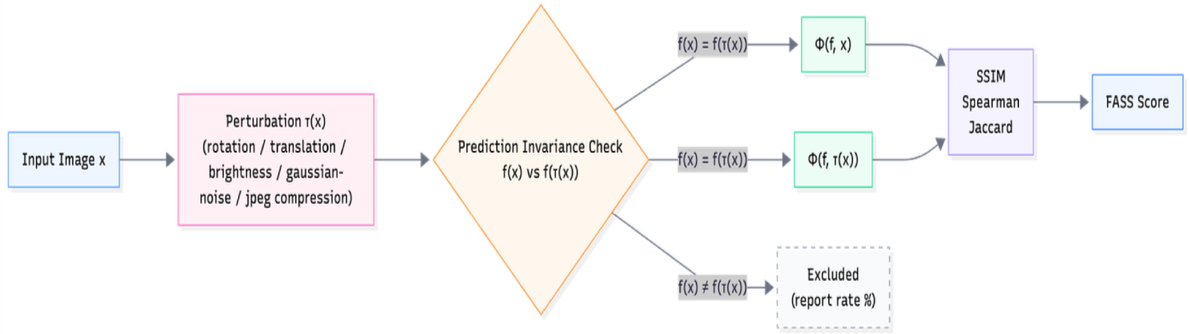
Fig.1 - from Ghorbani et al., AAAI 2019 [1] - Top row: original images with saliency maps. Bottom row: perturbed versions, same label, but attribution maps shift dramatically to irrelevant regions.

2) One number can't capture how an explanation failed



3) Only additive noise is tested – geometric, photometric, and compression perturbations ,the ones that actually occur in deployment are never evaluated.

Methods and Materials



PIPELINE

① **Perturbations**
a) Geometric: 15° rotation and 20px translation
b) Photometric: brightness ×1.5, Gaussian noise σ=0.15
Compression: JPEG quality 40

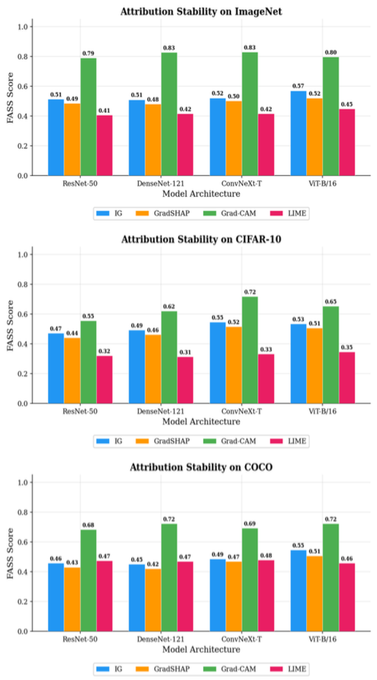
② **Prediction-Invariance Filter:** Keep only pairs(input, perturbed image pairs) where top-1 prediction is preserved. Report retention rate as a diagnostic.

③ **Attribute** *(on retained pairs only)*
4 XAI methods: Integrated Gradients, SHAP, Grad-CAM, LIME
evaluated each on **4 models**
ResNet-50, DenseNet-121, ConvNeXt-Tiny, ViT-B/16

④ **Measure**
SSIM: spatial coherence
Spearman ρ: feature importance ordering
Jaccard@100: top salient region overlap
FASS = (SSIM + Spearman + Jaccard) / 3

Images	*~ 70,000*
Explanation Computations	*6.4 million*
Evaluation Conditions	*(3 datasets × 4 arch. × 4 methods × 5 perturb.)*
Datasets	*ImageNet-1K MS-COCO CIFAR-10*
Resources Utilized	*NVIDIA A100 40GB PyTorch + Captum*

Category	SSIM	Spn.	Jac.	FASS
Geometric	.725	.666	.099	.497
Photometric	.770	.724	.178	.557
Compression (JPEG)	.791	.739	.196	.576



What do the results say?

① **Grad-CAM is the most stable** across all datasets and architectures

② **Method choice dominates architecture choice:** attribution method matters more than backbone

③ **Geometric perturbations are more destabilizing** than photometric or compression

④ **Without filtering, up to 99% of pairs involve changed predictions:** retention as low as 0.1% under translation

⑤ **ImageNet retains most pairs (63.3%):** CIFAR-10 lowest (11.5%) due to upsampling mismatch

SCAN THIS FOR GIT REPO

Perturbation	Mean	Std	Range
Rotation	30.9	27.5	0.0–88.1
Translation	0.1	0.2	0.0–0.6
Brightness	0.8	2.6	0.0–9.0
Noise	34.5	38.7	0.0–94.4
JPEG	1.0	3.4	0.0–11.7

What's NOVEL about FASS?

① **First to enforce prediction invariance:** stability scored only on prediction-preserving pairs; retention rate reported as diagnostic

② **Three-axis decomposition:** SSIM · Spearman · Jaccard reveal *how* explanations fail, not just *that* they fail.

③ **Beyond additive noise:** geometric, photometric & compression perturbations evaluated for the first time.

Limitations and Future Work

LIMITATIONS:

① **Stability ≠ faithfulness:** stable but incorrect attributions score highly; joint evaluation needed

② **Fixed perturbation magnitudes:** magnitude sweeps may reveal non-linear stability degradation

FUTURE WORK:

Extend to SmoothGrad, LRP, and confidence-based filtering beyond argmax-only retention.

References

[1] Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 3681–3688.

[2] Arun, N., Gidwani, M., Faber, J., Hajek, C., Vaickus, L., Salas, O., & Torresani, L. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence, 3(6), e200267.