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

Supplementary Material

Appendix

In the Appendix, we provide the following:

- metric design details, including Spearman rescaling and computational cost breakdown (Appendix A).
- per-perturbation stability scores, retention rates, and heatmaps for CIFAR-10 (Appendix B), ImageNet (Appendix C), and COCO (Appendix D).

All values are computed over prediction-invariant image pairs. Scores in conditions with near-zero retention (<0.1%) rest on very few evaluation pairs and should be interpreted with caution; such cells are marked with † in the tables below.

A Metric Design Details

A.1 Spearman Rescaling

Raw Spearman rank correlation [1] ρ_s ranges over $[-1, +1]$, where $+1$ indicates identical importance ordering, 0 indicates no linear rank association, and -1 indicates perfectly reversed ordering. SSIM [2] and top-k Jaccard [3] both range over $[0, 1]$. To ensure equal contribution to the composite FASS score, we rescale Spearman via $R = (\rho_s + 1)/2$, mapping perfect agreement to 1.0, random ordering to 0.5, and complete reversal to 0.0. Without this rescaling, negative Spearman values would pull the unweighted mean disproportionately downward relative to the other two components, which cannot take negative values.

A.2 Computational Cost

LIME was the most expensive method, accounting for approximately 70% of total runtime ($\sim$ 100–120 GPU hours) due to its sampling-based surrogate modeling. GradientSHAP required approximately 48 GPU hours owing to background sampling and gradient accumulation. Gradient-based methods (Integrated Gradients and Grad-CAM) were substantially cheaper, collectively requiring fewer than 30 GPU hours. Peak GPU memory usage occurred during GradientSHAP evaluations ($\sim$38.2 GB), approaching the 40 GB hardware limit. Given Colab session limits and runtime interruptions, effective wall-clock duration extended to several months.

B CIFAR-10

Table 1 reports prediction-invariant retention rates for all architecture–perturbation combinations on CIFAR-10. These rates are identical across attribution methods (retention depends only on the model and perturbation, not the explanation technique). Row-wise averages of the stability tables below correspond to the per-model scores in Table 4 of the main paper.

Table 1: Prediction-invariant retention (%) on CIFAR-10.

Architecture	Rot.	Trans.	Bright.	Noise	JPEG
ResNet-50	37.2	0.6	9.0	26.9	<0.1
DenseNet-121	20.8	<0.1	<0.1	12.1	<0.1
ConvNeXt-T	45.3	<0.1	<0.1	11.6	<0.1
ViT-B/16	36.1	<0.1	<0.1	29.2	<0.1

B.1 Integrated Gradients

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG†
ResNet-50	0.422	0.560	0.470	0.381	0.528
DenseNet-121	0.446	0.532	0.511	0.380	0.599
ConvNeXt-T	0.456	0.581	0.607	0.439	0.652
ViT-B/16	0.440	0.542	0.569	0.470	0.650

Table 2: IG stability on CIFAR-10. †Near-zero retention; interpret with caution.

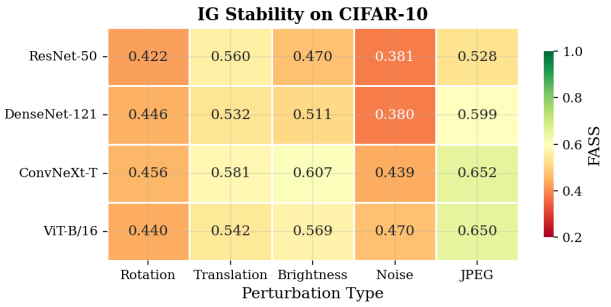

Figure 1: IG stability on CIFAR-10.

B.2 GradientSHAP

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG†
ResNet-50	0.423	0.501	0.425	0.393	0.468
DenseNet-121	0.445	0.491	0.479	0.384	0.520
ConvNeXt-T	0.457	0.536	0.563	0.441	0.583
ViT-B/16	0.442	0.517	0.523	0.463	0.588

Table 3: GradientSHAP stability on CIFAR-10. †Near-zero retention.

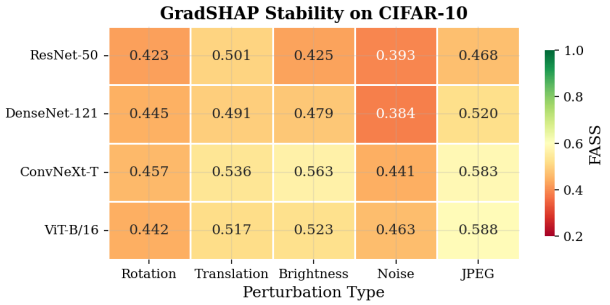

Figure 2: GradientSHAP stability on CIFAR-10.

B.3 Grad-CAM

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG†
ResNet-50	0.450	0.529	0.643	0.449	0.703
DenseNet-121	0.535	0.523	0.735	0.553	0.760
ConvNeXt-T	0.548	0.834	0.796	0.645	0.774
ViT-B/16	0.539	0.667	0.744	0.598	0.722

Table 4: Grad-CAM stability on CIFAR-10. †Near-zero retention.

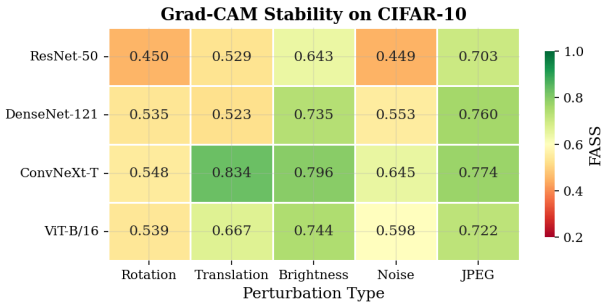

Figure 3: Grad-CAM stability on CIFAR-10.

B.4 LIME

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG†
ResNet-50	0.282	0.346	0.346	0.284	0.350
DenseNet-121	0.279	0.339	0.338	0.273	0.344
ConvNeXt-T	0.295	0.362	0.356	0.288	0.364
ViT-B/16	0.280	0.368	0.357	0.287	0.437

Table 5: LIME stability on CIFAR-10. †Near-zero retention.

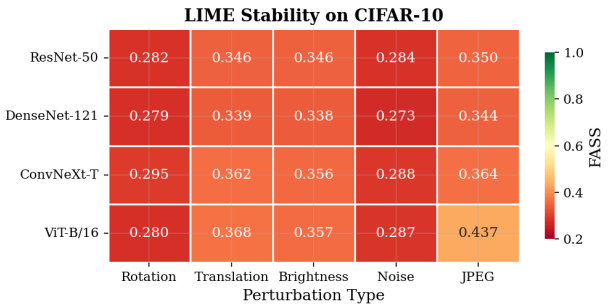

Figure 4: LIME stability on CIFAR-10.

Summary. CIFAR-10 yields the lowest stability across all datasets, consistent with distribution mismatch from upsampling 32×32 inputs to 224×224. Grad-CAM achieves the highest stability across all perturbations and architectures. IG and GradientSHAP track closely, reflecting their shared gradient-based formulation. LIME exhibits the lowest stability overall. ConvNeXt-T and ViT-B/16 show slightly improved stability compared to ResNet-50 and DenseNet-121, though differences remain moderate.

C ImageNet

Table 6 reports retention rates on ImageNet. Only rotation and noise retain substantial prediction-invariant pairs; translation, brightness, and JPEG yield near-zero retention across all architectures. Row-wise averages of the stability tables below correspond to the per-model scores in Table 4 of the main paper.

Table 6: Prediction-invariant retention (%) on ImageNet.

Architecture	Rot.	Trans.	Bright.	Noise	JPEG
ResNet-50	58.8	<0.1	<0.1	70.1	<0.1
DenseNet-121	58.4	<0.1	<0.1	70.5	<0.1
ConvNeXt-T	68.3	<0.1	<0.1	71.3	<0.1
ViT-B/16	35.7	<0.1	<0.1	73.0	<0.1

C.1 Integrated Gradients

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG†
ResNet-50	0.471	0.418	0.502	0.642	0.541
DenseNet-121	0.466	0.409	0.498	0.642	0.538
ConvNeXt-T	0.483	0.432	0.521	0.613	0.562
ViT-B/16	0.497	0.448	0.553	0.768	0.589

Table 7: IG stability on ImageNet. †Near-zero retention.

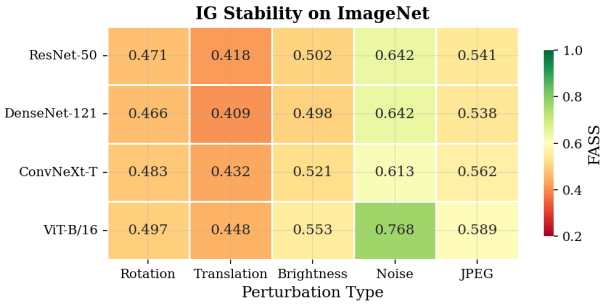

Figure 5: IG stability on ImageNet.

C.2 GradientSHAP

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG†
ResNet-50	0.470	0.421	0.472	0.572	0.492
DenseNet-121	0.465	0.412	0.468	0.572	0.487
ConvNeXt-T	0.484	0.438	0.501	0.568	0.519
ViT-B/16	0.490	0.443	0.512	0.626	0.531

Table 8: GradientSHAP stability on ImageNet. †Near-zero retention.

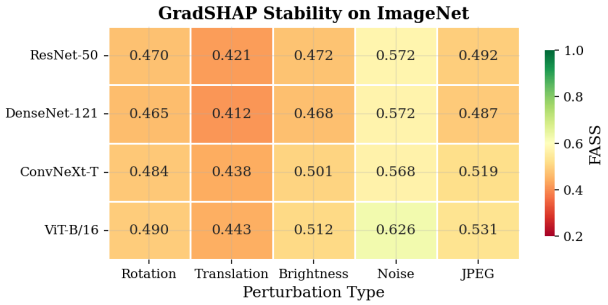

Figure 6: GradientSHAP stability on ImageNet.

C.3 Grad-CAM

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG†
ResNet-50	0.726	0.671	0.821	0.877	0.853
DenseNet-121	0.761	0.703	0.873	0.915	0.889
ConvNeXt-T	0.779	0.724	0.869	0.897	0.878
ViT-B/16	0.762	0.712	0.842	0.822	0.852

Table 9: Grad-CAM stability on ImageNet. †Near-zero retention.

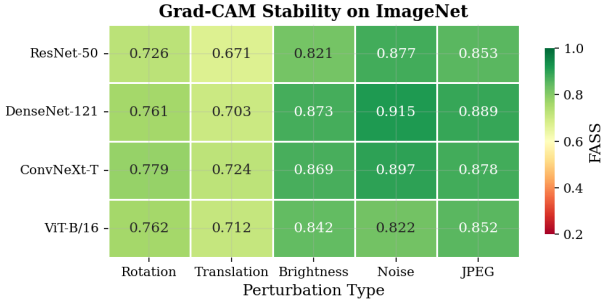

Figure 7: Grad-CAM stability on ImageNet.

C.4 LIME

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG†
ResNet-50	0.410	0.340	0.390	0.470	0.430
DenseNet-121	0.370	0.360	0.410	0.490	0.450
ConvNeXt-T	0.390	0.350	0.430	0.470	0.440
ViT-B/16	0.400	0.380	0.440	0.570	0.460

Table 10: LIME stability on ImageNet. †Near-zero retention.

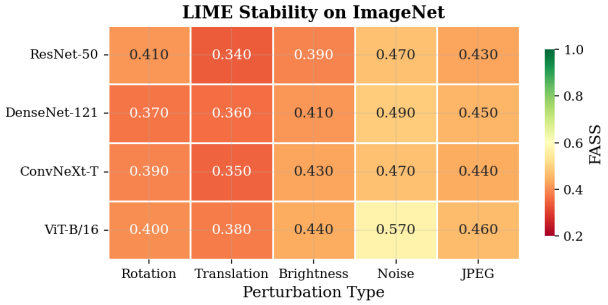

Figure 8: LIME stability on ImageNet.

Summary. ImageNet yields the highest stability among all three datasets for conditions with substantial retention (rotation and noise), consistent with pretrained feature alignment and native input resolution. Grad-CAM exceeds 0.85 under additive noise across all architectures. IG and GradientSHAP remain closely aligned. Translation, brightness, and JPEG produce near-zero retention, limiting the reliability of stability scores for those conditions.

D COCO

Table 2 in the main paper reports prediction-invariant retention rates by perturbation type. Rotation and noise retain substantial invariant pairs; JPEG retains a small fraction on ResNet-50 (11.7%). Row-wise averages of the stability tables below correspond to the per-model scores in Table 4 of the main paper.

Table 11: Prediction-invariant retention (%) on COCO.

Architecture	Rot.	Trans.	Bright.	Noise	JPEG
ResNet-50	88.1	<0.1	<0.1	74.3	11.7
DenseNet-121	51.7	<0.1	<0.1	90.6	<0.1
ConvNeXt-T	46.1	<0.1	<0.1	62.8	<0.1
ViT-B/16	45.4	<0.1	<0.1	76.7	0.3

D.1 Integrated Gradients

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG
ResNet-50	0.446	0.358	0.472	0.606	0.405
DenseNet-121	0.440	0.340	0.516	0.585	0.377†
ConvNeXt-T	0.474	0.390	0.548	0.597	0.416†
ViT-B/16	0.481	0.405	0.558	0.755	0.529

Table 12: IG stability on COCO. †Near-zero retention.

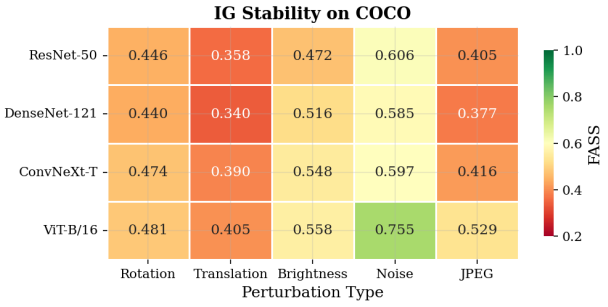

Figure 9: IG stability on COCO.

D.2 GradientSHAP

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG
ResNet-50	0.445	0.364	0.425	0.533	0.391
DenseNet-121	0.437	0.339	0.442	0.519	0.367†
ConvNeXt-T	0.475	0.393	0.506	0.568	0.411†
ViT-B/16	0.481	0.404	0.515	0.647	0.494

Table 13: GradientSHAP stability on COCO. †Near-zero retention.

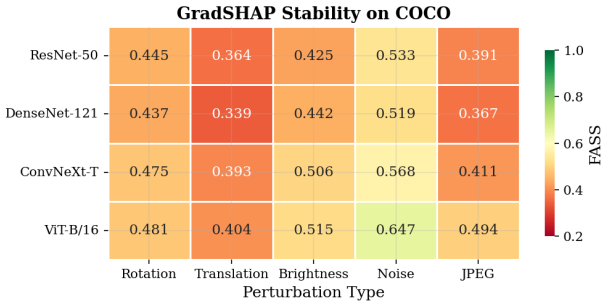

Figure 10: GradientSHAP stability on COCO.

D.3 Grad-CAM

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG
ResNet-50	0.645	0.609	0.676	0.825	0.662
DenseNet-121	0.661	0.595	0.824	0.829	0.707†
ConvNeXt-T	0.692	0.545	0.806	0.838	0.582†
ViT-B/16	0.752	0.716	0.637	0.849	0.658

Table 14: Grad-CAM stability on COCO. †Near-zero retention.

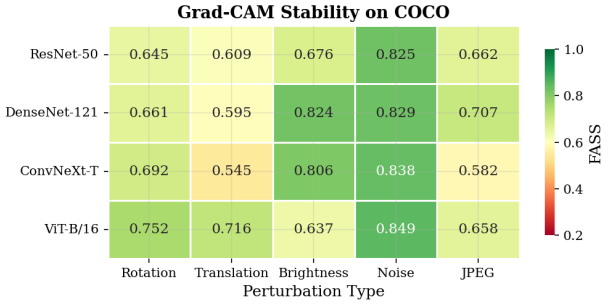

Figure 11: Grad-CAM stability on COCO.

D.4 LIME

Architecture	Rot.	Trans.†	Bright.†	Noise	JPEG
ResNet-50	0.427	0.489	0.529	0.415	0.512
DenseNet-121	0.427	0.466	0.526	0.432	0.495†
ConvNeXt-T	0.448	0.492	0.556	0.422	0.476†
ViT-B/16	0.429	0.445	0.517	0.413	0.489

Table 15: LIME stability on COCO. †Near-zero retention.

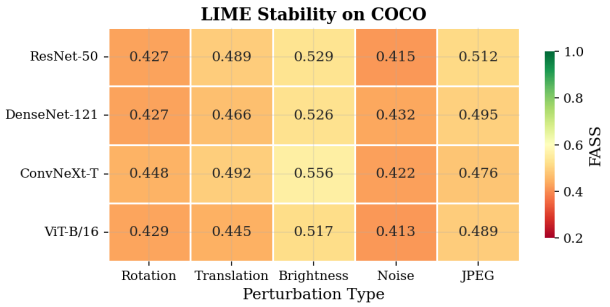

Figure 12: LIME stability on COCO.

Summary. COCO displays intermediate stability between CIFAR-10 and ImageNet. Grad-CAM remains the most stable method, though the gap with IG narrows relative to ImageNet. LIME exhibits stronger performance on COCO than on CIFAR-10, consistent with the hypothesis that multi-object scenes constrain perturbation-based sampling variability. Architectural differences are more pronounced under noise, where ViT-B/16 demonstrates elevated stability relative to convolutional architectures.

Appendix References

[1] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[3] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.