A Benchmark Study on the Reliability of Explainability Methods

Ibna Kowsar; Shahbaz Rezaei, PhD; Fanyu Meng; Xin Liu, PhD
University of California, Davis

UC DAVIS UNIVERSITY OF CALIFORNIA

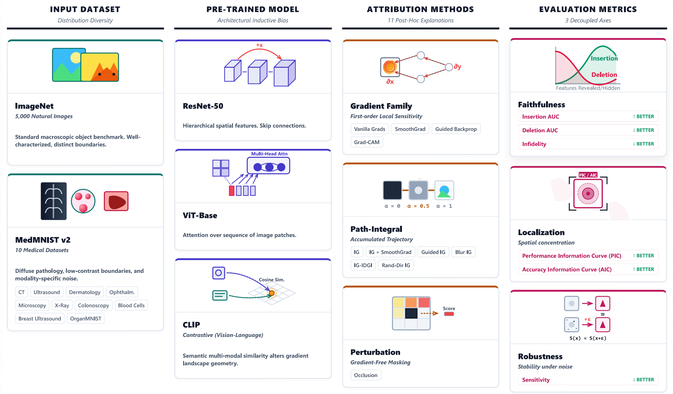
CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

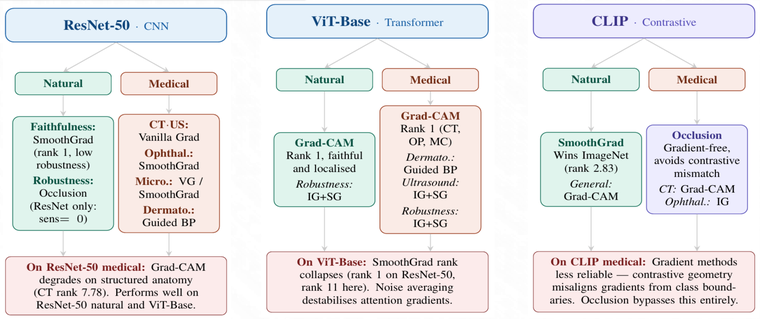
Post-hoc attribution methods are widely used to explain deep vision models, yet their reliability across architectures and imaging domains is poorly understood

- Existing benchmarks evaluate within a single model family or data domain
- Architecture-agnostic XAI selection is common but empirically unvalidated
- We present a rigorous, multi-metric benchmark evaluating 11 XAI methods across 3 architectures (ResNet-50, ViT-Base, CLIP) on ImageNet and 10 medical imaging datasets (MedMNIST v2)

Benchmark Pipeline

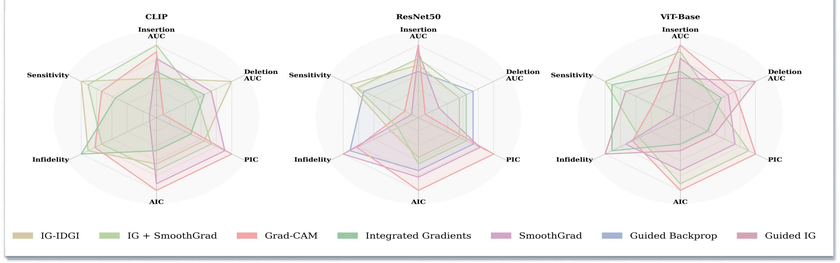


Architecture-Aware Selection Guide



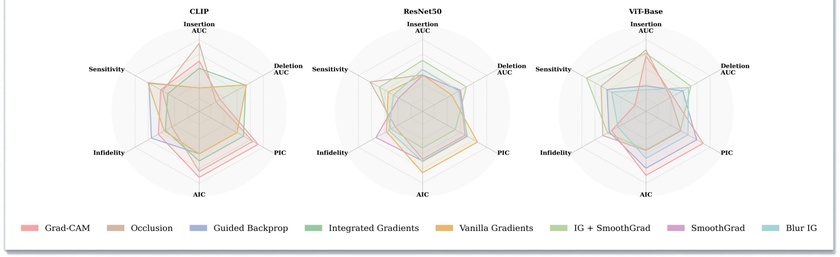
Performance — ImageNet

Normalized rank across 6 metrics (outer edge = best) for ResNet-50 (CNN), ViT-Base (Transformer), and CLIP (Contrastive). Larger area = stronger overall performance.



Performance — Medical Imaging

Rankings on MedMNIST are more architecture-dependent and domain-specific than on ImageNet. Methods that excel on natural images can collapse on medical data.



Interesting Insights and Open Questions

Architecture Drives Rankings

No XAI method dominates across CNN, Transformer, and Contrastive backbones → selection should be architecture aware

Why Gradient Methods Fail on CLIP?

Could it be that contrastive objective optimizes cross-modal similarity between image and text embeddings rather than class-discriminative visual boundaries, and the resulting feature geometry is organized around semantic separation across modalities [2].

Robustness is Not a Proxy for Faithfulness

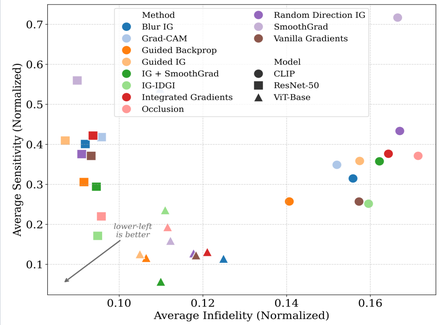
Infidelity values cluster tightly; Sensitivity varies substantially. “Methods that achieve low infidelity do not necessarily achieve low sensitivity”

Why Methods that scale on Natural Image, Fails on Medical Images?

Natural images has well-localized objects; medical images do not. Diffuse pathology, low-contrast boundaries, and texture-driven labels invalidate the coarse-localization assumption built into many attribution methods.

Faithfulness ≠ Robustness

Infidelity clusters tightly while Sensitivity varies by ~10×. These axes are empirically decoupled.



FAITHFULNESS ⊥ ROBUSTNESS

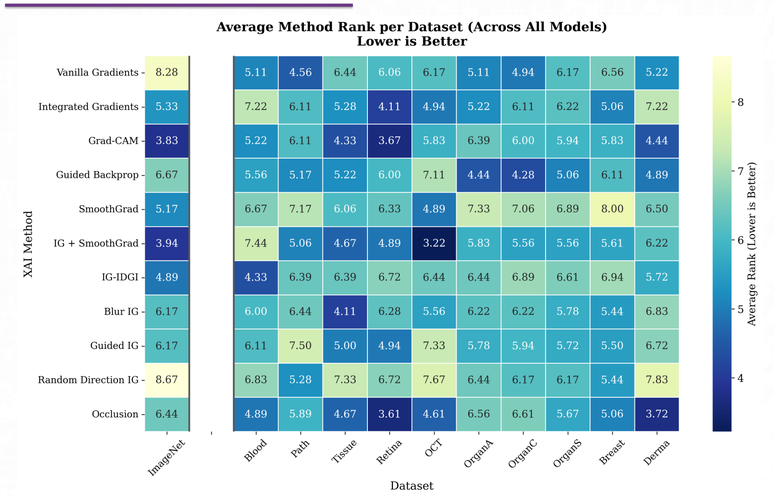
Infidelity *clusters tightly* — **0.08 — 0.10** across all methods

Sensitivity *varies ~10×* — **0.00 — 0.90** across all methods

→ Evaluating only faithfulness is insufficient.
Contradicts joint-decrease prediction of Yeh et al. (NeurIPS '19).

CLIP CONTRASTIVE GEOMETRY MISMATCH

Domain × Method Heatmap



Contact

Ibna Kowsar
University of California, Davis
Email: ikowsar@ucdavis.edu
Website: https://kawseribn-github-io.vercel.app/

References

1. Chih-Kuan Yeh, Been Kim, Changhua Hsieh, and PradeepRavikumar. On the (in)fidelity and sensitivity of explana-tions. In Advances in Neural Information Processing Sys-tems (NeurIPS), 2019. 1, 3, 6, 8
2. Alec Radford, Jong Wook Kim, Chris Hallacy, AdityaRamesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,Amanda Askell, Pamela Mishkin, Jack Clark, GretchenKrueger, and Ilya Sutskever. Learning transferable visualmodels from natural language supervision. In Proceedingsof the 38th International Conference on Machine Learning(ICML), 2021. 1, 2, 4, 8