Summary of Master's Thesis
Feature-Gradient Attribution: Enhancing Vision Transformer Explanations with Sparse Autoencoders

Julius Šula
Advisors: Prof. Thomas Lukasiewicz, Bayar Ilhan Menzat

Abstract

Vision Transformers achieve strong performance in medical and natural imaging, yet their decision processes remain opaque. TransMM, a leading attribution method, combines attention with gradients to highlight influential patches. We introduce *Feature-Gradient Attribution*, which extends TransMM's principle from attention space to semantic feature space using Sparse Autoencoders (SAEs). SAEs decompose activations into interpretable features; we project gradients onto this feature basis to compute feature-gradient scores capturing both feature presence and influence on predictions. These scores modulate TransMM's attention maps, forming a lightweight, semantically-informed correction. Across three datasets (chest X-rays, endoscopy, natural images), two architectures (fine-tuned ViT-B/16, CLIP ViT-B/32), and three complementary faithfulness metrics, our method achieves consistent improvements: 10.5–44.3% on SaCo, 14.0–43.0% on Faithfulness Correlation, and 1.3–10.8% on Pixel Flipping. Ablation studies confirm that both feature activations and gradients are necessary. To our knowledge, this is the first integration of sparse semantic features during attribution computation, demonstrating that mechanistic feature structure can materially enhance Transformer attributions.

1 Introduction

Vision Transformers (ViTs) have become prominent in computer vision, particularly in high-stakes medical imaging applications like chest X-ray diagnosis [16] and endoscopy [18]. However, their decision-making processes remain opaque, creating critical challenges for clinical deployment where understanding *why* a model makes predictions is as important as predictive accuracy.

TransMM (Transformer Attribution using Multimodal Mixing) [7] has emerged as a leading attribution method, generating saliency maps of input images to visualize importance of regions inside the input image by combining attention weights with gradients to identify class-specific important patches. Its core principle is simple: patch importance emerges from combining *attention maps* (where the model attends) with *gradients* (how that attention affects predictions). However, TransMM operates on dense gradient signals that aggregate influences across all learned representations, potentially conflating relevant semantic features with noise.

Recent advances in mechanistic interpretability have introduced Sparse Autoencoders (SAEs) [5, 11] as tools for understanding model internals. SAEs decompose polysemantic activations into interpretable features where each dimension corresponds to a distinct semantic concept. While initially developed for language models, recent work has demonstrated their effectiveness in vision domains [12, 15], with features exhibiting clear semantic structure and causal relevance through intervention studies [2].

Research Gap. While prior work has used attribution to explain SAE features [10] or SAE features to validate attribution post-hoc [19], neither integrates semantic feature information *directly into*

attribution computation. We investigate whether TransMM's principle of combining activations with gradients extends from attention space to feature space.

Contributions. We introduce Feature-Gradient Attribution, which: (1) projects residual gradients through SAE feature space to identify which semantic features drive patch importance, (2) integrates this decomposition into TransMM through attention map modulation, and (3) demonstrates substantial faithfulness improvements across diverse datasets and architectures over multiple faithfulness metrics. Ablation studies confirm that attribution improvement requires the interaction between feature activations and gradients and that neither signal alone suffices.

2 Related Work

Attribution Methods for Transformers. Layer-wise Relevance Propagation (LRP) [14] and Deep Taylor Decomposition [17] provide principled frameworks for attribution through relevance conservation. TransLRP [6] extended these to Transformers but requires complex relevance propagation through all layers.

TransMM [7] simplified this by directly weighting attention with gradients. For each transformer layer ℓ, it computes gradient-weighted attention maps:

$$\bar{A}^{(\ell)} = I + \mathbb{E}_h[(\nabla A^{(\ell)} \odot A^{(\ell)})_+] \tag{1}$$

where $A^{(\ell)}$ are the raw attention weights, $\nabla A^{(\ell)}$ denotes gradients of the target class logit with respect to these attention weights, $\odot$ is element-wise multiplication, $(\cdot)_+$ retains only positive contributions, $\mathbb{E}_h$ averages across attention heads, and I is the identity matrix accounting for residual connections. These weighted attention maps are then aggregated across layers through relevancy propagation to produce final attribution scores. This approach achieves comparable performance to TransLRP with dramatic computational savings.

Faithfulness Metrics. We employ three complementary perturbation-based metrics. Faithfulness Correlation [3] measures whether attribution magnitudes predict impact when features are perturbed. Pixel Flipping [14] tests whether removing features in attribution order causes appropriate performance degradation. SaCo [21] addresses cumulative perturbation limitations through independent perturbation, testing whether attribution magnitude differences between patches align with actual impact differences.

Sparse Autoencoders. SAEs address polysemanticity in neural networks [8] by learning overcomplete sparse representations. Recent work demonstrates that SAE features in vision models exhibit semantic structure [15], are causally relevant [2], and can validate attribution hypotheses [19]. However, no prior work integrates SAE features during attribution computation.

3 Feature-Gradient Attribution

3.1 Extending TransMM's Principle to Feature Space

TransMM's success stems from combining attention patterns with gradient information. We propose that this principle extends to sparse semantic features: feature activations indicate *which* semantic concepts are present, while feature-space gradients indicate *how* those concepts influence predictions.

Consider a residual layer ℓ with SAE decoder $D \in \mathbb{R}^{d \times K}$. For each spatial token t, let $x_t \in \mathbb{R}^d$ be the residual vector, $f_t \in \mathbb{R}^K_{\geq 0}$ the SAE feature activations, and $g_t = \nabla_{x_t} y$ the gradient with respect to target logit y.

Feature-Space Gradient Projection. We project gradients into feature space via the decoder transpose:

$$\nabla f_t = D^\top g_t \in \mathbb{R}^K \tag{2}$$

where $\nabla f_t^{(k)}$ represents the sensitivity of the target logit to feature k at token t.

Feature-Weighted Scoring. For each feature, we compute its contribution as the product of gradient sensitivity and activation strength:

$$s_t = \sum_{k=1}^{K} \nabla f_t^{(k)} \cdot f_t^{(k)} \tag{3}$$

This extends TransMM's attention-gradient product to feature space: just as $\nabla A \odot A$ combines attention presence with gradient influence, our $\nabla f \odot f$ combines feature presence with feature-space gradient influence.

3.2 Gate Construction and Integration

We convert per-patch scores s_t into multiplicative gates through robust normalization and exponential mapping:

$$\hat{s}_t = \frac{s_t - \text{median}(s)}{\text{MAD}(s) + \epsilon} \tag{4}$$

$$w_t = \exp(\log(c_{\max}) \times \tanh(\kappa \cdot \hat{s}_t)) \tag{5}$$

where MAD is median absolute deviation (scaled by 1.4826), κ controls sensitivity, and $c_{\max}$ determines the range $[1/c_{\max}, c_{\max}]$. This produces symmetric multiplicative corrections centered at 1.0.

We modulate TransMM's gradient-weighted attention through columnwise gating via diagonal matrix multiplication. This operation scales each column of $\bar{A}^{(\ell)}$ by the corresponding patch gate, modulating how much attention all patches can give to each spatial token:

$$\bar{A}^{(\ell)}_{\text{gated}} = \bar{A}^{(\ell)} \cdot \text{diag}(1, w_0, w_1, \ldots, w_{N-1}) \tag{6}$$

with the CLS token gate fixed at 1, since we don't want to influence the global representation but only the local. Patches with high feature-gradient scores ($w_t > 1$) attract more attention, while patches with low scores ($w_t < 1$) receive reduced attention from all other patches. The standard TransMM relevancy propagation then proceeds: $\mathcal{R}^{(\ell)} = \mathcal{R}^{(\ell-1)} + \bar{A}^{(\ell)}_{\text{gated}} \cdot \mathcal{R}^{(\ell-1)}$.

Multi-Layer Configuration. Feature-gradient gating can be applied at selected layers $\mathcal{L} \subseteq \{1, \ldots, L\}$ to compound corrections, with validation experiments identifying optimal layer combinations.

4 Experiments

4.1 Experimental Setup

Datasets and Models. We evaluate on three diverse datasets:

- **COVID-QU-Ex** [20]: 33,920 chest X-rays across three classes, using fine-tuned ViT-B/16 (95.4% accuracy) [13]
- **Hyperkvasir** [4]: Gastrointestinal endoscopy with six anatomical landmarks, using fine-tuned ViT-B/16 [18]
- **ImageNet-1k**: 10,000-image subset, using CLIP ViT-B/32 trained on DataComp-1B [9] (72.7% zero-shot accuracy)

SAE Training. For medical datasets, we train dataset-specific SAEs using Prisma [12] with 64× expansion factor, Top-K activation (K=128), and >95% explained variance on layers 2-10. For ImageNet, we use publicly available Prisma SAEs trained on CLIP-ViT-B/32. All SAEs are trained on patch tokens only, as our attribution method gates spatial patches.

Faithfulness Evaluation. We adapt all metrics to patch granularity: Faithfulness Correlation samples 20-patch subsets, Pixel Flipping removes patches sequentially, and SaCo partitions patches into groups for independent perturbation. All metrics use per-sample mean patch replacement.

4.2 Validation Experiments

Ablation Study. To isolate the source of improvement, we test three variants:

1. *Combined* (primary): $s_t = \sum_k \nabla f_t^{(k)} \cdot f_t^{(k)}$
2. *Activation-only*: $s_t = \sum_k f_t^{(k)}$

Table 1: Test set results. Bold indicates best performance. Significance: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (Welch's t-test vs. vanilla baseline).

Dataset	Variant	SaCo↑	Faith.Corr.↑	Pixel↓
COVID-QU-Ex	Vanilla	0.459 ± 0.004	0.309 ± 0.002	93.17 ± 0.68
	Real	$\mathbf{0.507 \pm 0.004}^{***}$	$\mathbf{0.442 \pm 0.002}^{***}$	$\mathbf{83.06 \pm 0.67}^{***}$
	Shuffled	$0.434 \pm 0.004^{***}$	$0.319 \pm 0.002^{**}$	$96.66 \pm 0.70^{***}$
Hyperkvasir	Vanilla	0.504 ± 0.016	0.479 ± 0.009	94.50 ± 2.41
	Real	$\mathbf{0.727 \pm 0.015}^{***}$	$\mathbf{0.550 \pm 0.008}^{***}$	$\mathbf{93.31 \pm 2.38}$
	Shuffled	$0.616 \pm 0.019^{***}$	$0.446 \pm 0.009^{*}$	97.87 ± 2.36
ImageNet	Vanilla	0.250 ± 0.004	0.314 ± 0.003	8.83 ± 0.09
	Real	$\mathbf{0.337 \pm 0.004}^{***}$	$\mathbf{0.358 \pm 0.002}^{***}$	$\mathbf{8.52 \pm 0.09}^{*}$
	Shuffled	0.242 ± 0.004	$0.250 \pm 0.003^{***}$	$9.70 \pm 0.09^{***}$

3. *Gradient-only*: $s_t = \sum_k |\nabla f_t^{(k)}|$

Single-layer experiments (500-image validation subsets, $\kappa = 0.5$, $c_{\max} = 10.0$) reveal that the combined method consistently outperforms vanilla TransMM across all datasets and layers, with peak improvements at intermediate layers (3-5 for medical datasets, 5-10 for ImageNet). Full layer-by-layer results are provided in the supplementary material. Critically, gradient-only systematically underperforms baseline, often degrading faithfulness correlation by up to -18% in late layers. Activation-only shows modest improvements (2-6%), but the combined method's substantial superiority demonstrates that gradient information combined with activations significantly enhances attribution.

Multi-Layer Synergy. Multi-layer configurations consistently outperform best single layers: COVID-QU-Ex [2,3,4] achieves 10.0% SaCo vs. 4.1% single-layer; Hyperkvasir [4,6,7] achieves 30.9% vs. 16.5%; ImageNet [6,9,10] achieves 43.2% vs. 23.6%. This suggests features from adjacent layers provide complementary semantic information.

Hyperparameter Validation. Systematic sweeps over $\kappa \in \{0.1, 0.5, 1.0\}$ and $c_{\max} \in \{2.0, 10.0, 50.0\}$ on best multi-layer configurations reveal robust improvement regions. We select dataset-specific hyperparameters optimizing validation performance for the following test set evaluations: COVID-QU-Ex ($\kappa = 1.0$, $c_{\max} = 50.0$), Hyperkvasir ($\kappa = 0.1$, $c_{\max} = 50.0$), ImageNet ($\kappa = 0.5$, $c_{\max} = 10.0$).

4.3 Test Set Results

Table 1 presents test set performance comparing vanilla TransMM baseline against our method with real SAE features and a shuffled decoder control that preserves statistical properties while destroying semantic structure.

Key Findings. Real SAE features achieve best performance on all nine metric-dataset combinations. SaCo and Faithfulness Correlation show statistically significant improvements ($p < 0.001$) across all three datasets: 10.5% and 43.0% (COVID-QU-Ex), 44.3% and 14.8% (Hyperkvasir), 34.8% and 14.0% (ImageNet). Pixel Flipping shows consistent improvements with reductions of 10.8% (COVID-QU-Ex, $p < 0.001$), 1.3% (Hyperkvasir), and 3.5% (ImageNet, $p < 0.05$).

Semantic Structure Validation. The shuffled decoder control provides strong evidence that semantic structure is essential: real features consistently outperform shuffled variants by 6.6% (COVID-QU-Ex), 12.8% (Hyperkvasir), and 39.3% (ImageNet) on SaCo. Interestingly, shuffled features sometimes exceed vanilla baseline, suggesting sparsity provides incidental denoising benefits. However, full performance gains require semantic alignment—on ImageNet, shuffled features actually underperform vanilla (SaCo: 0.242 vs. 0.250), likely because random feature projections become harmful without semantic grounding.

Qualitative Analysis. While quantitative metrics demonstrate substantial faithfulness improvements, examining individual attribution maps reveals how Feature-Gradient Attribution refines TransMM's behavior. Representative examples from both medical imaging and natural images (Appendix A) show consistent patterns: regions semantically central to classification receive increased emphasis,

while artifacts and background elements show reduced attribution. Importantly, these qualitative observations align with quantitative improvements across all faithfulness metrics.

5 Discussion and Conclusion

We investigated whether TransMM's principle of combining activation patterns with gradient information naturally extends from attention space to learned sparse features. Our comprehensive evaluation across three datasets, two architectures, and three faithfulness metrics demonstrates consistent improvements across all employed metrics.

Why Multi-Layer Synergy? Multi-layer configurations substantially outperform single layers (60–87% better), suggesting features from adjacent layers provide synergistic information. This likely stems from both reinforcement (similar concepts activating across layers provide independent evidence) and complementarity (different layers capture different semantic aspects).

Limitations. Our method requires training SAEs before attribution computation, adding preprocessing overhead. The modest Pixel Flipping improvements (1.3–10.8%) compared to substantial SaCo gains suggest our method primarily refines moderately important patches rather than changing the most obvious regions, which is consistent with Pixel Flipping's documented sensitivity to top-ranked patches [21].

Implications. This work demonstrates that mechanistic interpretability tools can directly enhance practical explainability methods. The success across medical imaging and natural images, fine-tuned and contrastively pre-trained models, indicates broad applicability. Our results suggest attribution methods should leverage learned semantic representations that respect the structure of model computations, opening new directions for faithful visual explanations in Vision Transformers.

Acknowledgments and Disclosure of Funding

Project code available at `https://github.com/piragi/gradcamfaith/`.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[2] Anonymous. Steering CLIP's vision transformer with sparse autoencoders. In *Mechanistic Interpretability for Vision at CVPR 2025 (Non-proceedings Track)*, 2025.

[3] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3016–3022. ijcai.org, 2020.

[4] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K Stensland, Enrique Garcia-Ceja, Peter T Schmidt, Hugo L Hammer, Michael A Riegler, Pål Halvorsen, and Thomas de Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020.

[5] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, October 2021.

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, June 2021.

[8] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. `https://transformer-circuits.pub/2022/toy_model/index.html`.

[9] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

[10] Sangyu Han, Yearim Kim, and Nojun Kwak. Causal interpretation of sparse autoencoder features in vision, 2025.

[11] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[12] Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video, 2025.

[13] Piotr Komorowski, Hubert Baniecki, and Przemyslaw Biecek. Towards evaluating explanations of vision transformers for medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3726–3732, June 2023.

[14] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10:e0130140, 07 2015.

[15] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[16] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh. xViTCOS: Explainable vision transformer based COVID-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1100110, 2021. PMID: 34956741; PMCID: PMC8691725.

[17] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

[18] Edward Sanderson and Bogdan J Matuszewski. A study on self-supervised pretraining for vision problems in gastrointestinal endoscopy. *IEEE Access*, 12:46181–46201, 2024.

[19] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models, 2025.

[20] Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 2021.

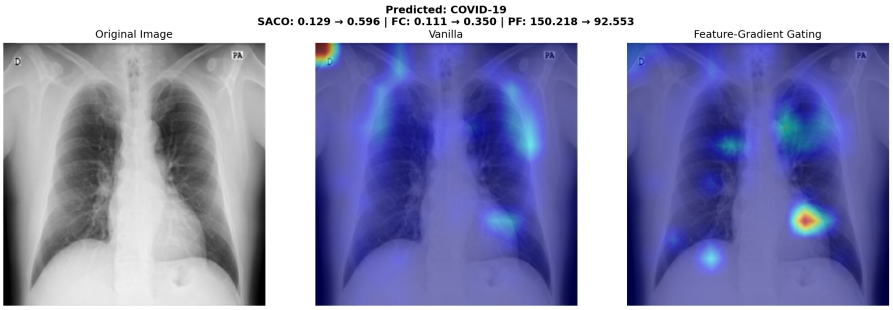
[21] Junyi Wu, Weitai Kang, Hao Tang, Yuan Hong, and Yan Yan. On the faithfulness of vision transformer explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10936–10945, 2024.

A Qualitative Examples

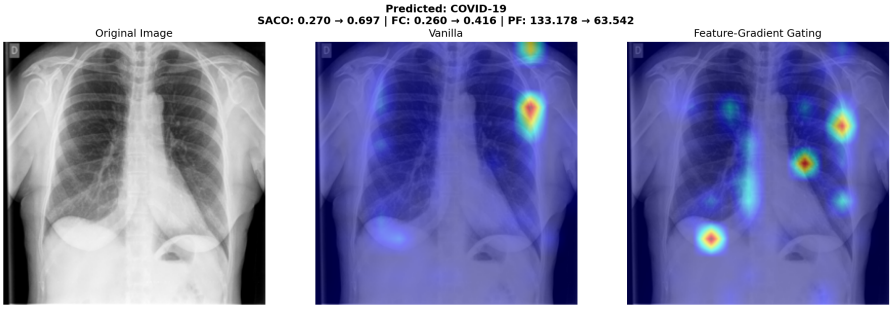
We present qualitative comparisons showing how Feature-Gradient Attribution refines TransMM's behavior. All examples demonstrate improved faithfulness metrics alongside visually refined attribution patterns.

A.1 Medical Imaging

Figure 1 shows COVID-19 chest X-rays where our method produces more faithful attributions. The examples illustrate characteristic refinement: reduced attribution to positioning markers (technical artifacts) and increased focus on lung regions.



(a) Feature-Gradient Attribution (right) reduces attribution to the positioning marker (upper-right) while strengthening focus on lung regions. Vanilla TransMM (left) assigns significant attribution to this technical artifact.



(b) Feature-Gradient Attribution (right) provides more focused coverage of lung regions compared to vanilla TransMM (left).

Figure 1: COVID-QU-Ex qualitative comparison. Left: vanilla TransMM, Right: Feature-Gradient Attribution. Both examples show substantial faithfulness improvements (higher SaCo and Faithfulness Correlation, lower Pixel Flipping AUC).

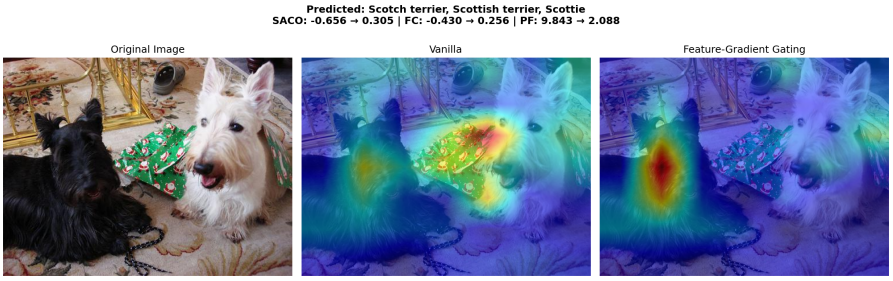
A.2 Natural Images

Figure 2 demonstrates generalization beyond medical imaging with ImageNet examples showing similar refinement patterns.

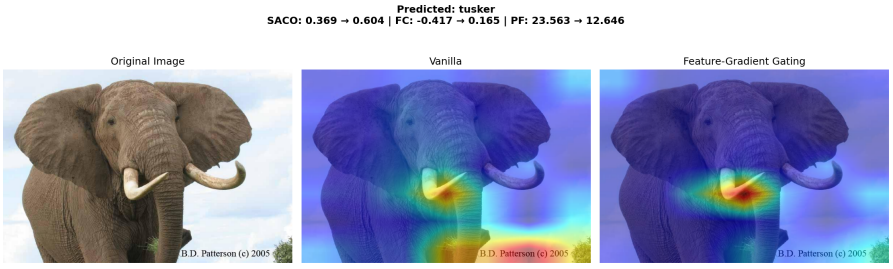
A.3 Discussion

These examples illustrate how Feature-Gradient Attribution refines TransMM's attention-gradient product: semantically relevant regions receive increased emphasis while artifacts (positioning markers, watermarks) and background elements show reduced attribution. Critically, qualitative observations align with quantitative improvements—all examples show higher SaCo and Faithfulness Correlation with lower Pixel Flipping AUC.

We emphasize that *visual plausibility does not guarantee faithfulness* [1]. Human intuitions about importance can diverge from actual model behavior. Our complementary quantitative metrics ensure



(a) Scotch Terrier. The gated method (right) strengthens attribution on the terrier while reducing attribution to background elements. Vanilla TransMM (left) shows diffuse attribution across background.



(b) Tusker. Feature-Gradient Attribution (right) de-emphasizes the watermark (lower-right) while strengthening focus on the tusks. Vanilla TransMM (left) distributes attribution more uniformly.

Figure 2: ImageNet qualitative comparison. Left: vanilla TransMM, Right: Feature-Gradient Attribution. Both examples show improved faithfulness metrics alongside refined attribution patterns.

observed improvements reflect genuine alignment with model behavior rather than merely matching human expectations.