Faithful Attribution in Vision Transformers via Feature-Gradient Gating

Julius Šula[1*] Thomas Lukasiewicz[1,2] Bayar Menzat[1]
[1]Vienna University of Technology, Austria
[2]University of Oxford, UK
[*]julius.sula@gmail.com

Abstract

Attention-based attribution methods like TransMM identify where *a Vision Transformer attends but not* which *internal features drive the prediction. Sparse Autoencoders (SAEs) can decompose ViT activations into interpretable feature dictionaries, yet their signals have not been integrated directly into attribution mechanisms. We propose* feature-gradient gating*: residual-stream gradients are projected onto SAE decoder directions, combined with feature activations to score patches, and used as multiplicative gates on TransMM's gradient-weighted attention before relevance propagation. The resulting per-patch scores decompose linearly into per-feature contributions, enabling inspection of which learned features drive each region's relevance. Across chest X-ray, endoscopy, and natural-image benchmarks, feature-gradient gating consistently improves Faithfulness Correlation and Salience-guided Faithfulness Coefficient (SaCo) over vanilla TransMM, with smaller or mixed gains on Pixel Flipping.*

1. Introduction

Vision Transformers (ViTs) [5] have become a standard architecture in computer vision, achieving strong performance across tasks ranging from natural image classification to medical imaging [13, 16, 19]. Their self-attention mechanism enables global information flow, but, like other deep models, ViTs remain difficult to interpret. This lack of transparency is especially problematic in high-stakes settings, where users must verify that predictions rely on meaningful visual evidence rather than spurious correlations or dataset artifacts. Attention-based attribution methods aim to address this by producing spatial explanations for ViTs. Among these, TransMM [4] has emerged as a strong baseline by combining attention weights with gradients to propagate class-specific relevance through transformer layers. Yet it identifies *where* the model attends, while ignoring the rich semantic information encoded in residual-stream activations, making it difficult to determine *what* specific internal features drive those attention patterns.

In parallel, recent work in mechanistic interpretability has shown that Sparse Autoencoders (SAEs) can decompose dense activations into sparse, interpretable features that often correspond to semantically meaningful concepts, and that manipulating these features can causally affect model behavior [3, 10–12, 14]. Prior work uses gradient-based attribution to identify causally relevant SAE latents in language models [7, 8, 15] and to explain SAE features in vision [6, 9, 21], but not to modulate spatial ViT attribution directly.

We connect these two lines of research by proposing *feature-gradient gating*, an extension of TransMM that uses feature-level activation–gradient scores derived from SAEs trained on the residual stream to modulate gradient-weighted attention maps before relevance propagation. The SAE is not inserted into the forward pass; gradients are projected onto SAE decoder directions as a side channel, yielding attribution maps that are both spatially faithful and decomposable into contributions from individual SAE features, so users can trace relevance changes back to specific learned features rather than only to token-level scores. **Our contributions are threefold:**

- We introduce feature-gradient gating, a simple extension of TransMM that incorporates sparse SAE feature signals into attention attribution without modifying the model forward pass.
- The resulting per-patch score decomposes linearly into SAE feature contributions, enabling feature-level auditing of relevance at each spatial location.
- Across three datasets and three faithfulness metrics, feature-gradient gating improves SaCo and Faithfulness Correlation over vanilla TransMM in all settings, while remaining competitive on Pixel Flipping and being more consistent than direct activation-gradient gating across datasets.

Code available at: https://github.com/piragi/featuregradientgate

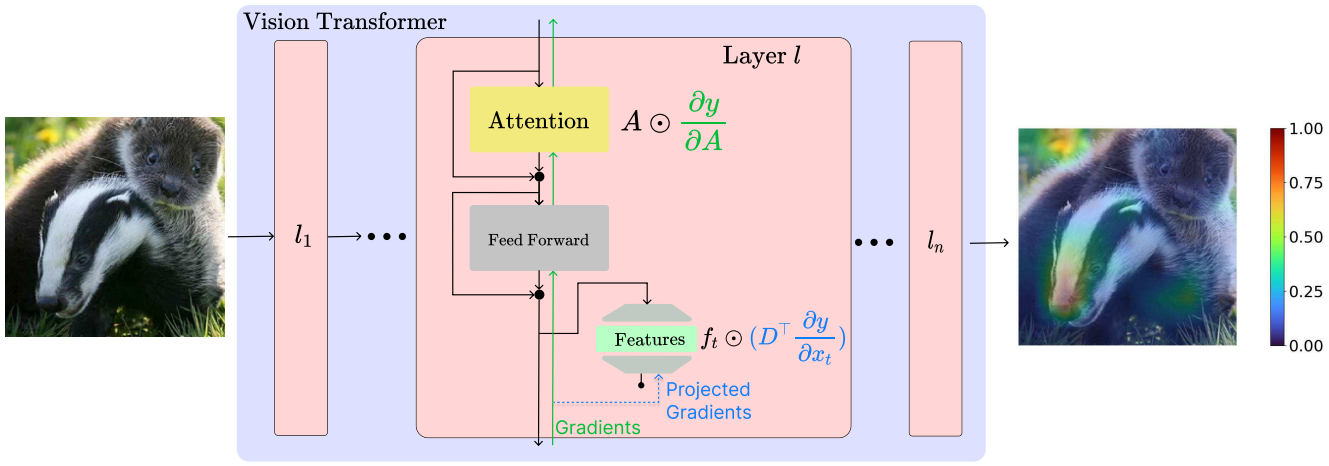

Figure 1. **Feature-gradient gating.** Residual-stream gradients are projected onto SAE decoder directions ($\tilde{g}_{f,t} = D^\top \frac{\partial y}{\partial x_t}$), combined with feature activations, normalized, and mapped to multiplicative gates. Gates modulate TransMM gradient-weighted attention before relevance propagation, yielding spatial attributions decomposable into per-feature contributions.

2. Method

2.1. Background and Notation

Consider a transformer with L layers and N spatial patches ($N+1$ tokens including CLS).

TransMM Attribution. TransMM [4] defines gradient-weighted attention at layer ℓ as

$$\bar{A}^{(\ell)} = \mathbb{E}_h \left[\left(\frac{\partial y}{\partial A_h^{(\ell)}} \odot A_h^{(\ell)} \right)_+ \right], \tag{1}$$

and propagates relevance via $\mathcal{R}^{(\ell)} = \mathcal{R}^{(\ell-1)} + \bar{A}^{(\ell)} \cdot \mathcal{R}^{(\ell-1)}$, with $\mathcal{R}^{(0)} = I$. Spatial attributions are read from CLS-to-patch entries of $\mathcal{R}^{(L)}$.

Notation. For layer ℓ and token t: residual-stream activation $x_t^{(\ell)} \in \mathbb{R}^d$; gradient $g_t^{(\ell)} = \partial y / \partial x_t^{(\ell)} \in \mathbb{R}^d$; SAE decoder $D^{(\ell)} \in \mathbb{R}^{d \times K}$ with columns as learned feature directions; sparse feature activations $f_t^{(\ell)} \in \mathbb{R}^K_{\geq 0}$ satisfying $D^{(\ell)} f_t^{(\ell)} + b^{(\ell)} \approx x_t^{(\ell)}$. The SAE is a *side channel*: not in the model's forward pass.

2.2. Feature-Gradient Gating

Feature-Direction Sensitivity. We estimate per-feature directional sensitivity (Fig. 1) by taking inner products between the residual-stream gradient and SAE decoder directions:

$$\tilde{g}_{f,t}^{(\ell)} := \left(D^{(\ell)}\right)^\top g_t^{(\ell)} \in \mathbb{R}^K. \tag{2}$$

The k-th component estimates first-order sensitivity of y along decoder direction $D_k^{(\ell)}$.

Feature-Gradient Scoring. Combining sensitivity with feature presence gives a per-patch importance score:

$$s_t^{(\ell)} := \left(\tilde{g}_{f,t}^{(\ell)}\right)^\top f_t^{(\ell)} = (g_t^{(\ell)})^\top \hat{x}_t^{(\ell)}, \tag{3}$$

where $\hat{x}_t^{(\ell)} := D^{(\ell)} f_t^{(\ell)}$ is the bias-free SAE reconstruction (derivation in Appendix A). This differs from the Direct baseline $(g_t^{(\ell)})^\top x_t^{(\ell)}$ by the reconstruction error.

Gate Computation. Scores are normalized using median and MAD per image per layer (details in Appendix B):

$$\hat{s}_t^{(\ell)} = \frac{s_t^{(\ell)} - \mu^{(\ell)}}{\sigma^{(\ell)} + \epsilon}, \qquad w_t^{(\ell)} = \alpha^{\tanh(\hat{s}_t^{(\ell)})}, \tag{4}$$

with base $\alpha > 1$, yielding $w_t^{(\ell)} \in [1/\alpha, \alpha]$ (we select α on validation; Appendix B). The CLS gate is fixed to 1 (it is not a spatial source).

2.3. Integration with TransMM

Gates modulate TransMM's gradient-weighted attention via right-multiplication:

$$\bar{A}_{\text{gated}}^{(\ell)} = \bar{A}^{(\ell)} \cdot \text{diag}\left(w_0^{(\ell)}, \ldots, w_N^{(\ell)}\right). \tag{5}$$

Right-multiplication scales source-token contributions: tokens with negative feature-gradient scores are downweighted as attribution sources. Gating is applied at selected layers $\mathcal{L}_{\text{SAE}}$ identified by validation performance (Appendix D).

Normalization. We do not renormalize $\bar{A}_{\text{gated}}^{(\ell)}$ after gating, instead bounding gates to $[1/\alpha, \alpha]$ to limit scale changes while preserving relative source-token reweighting.

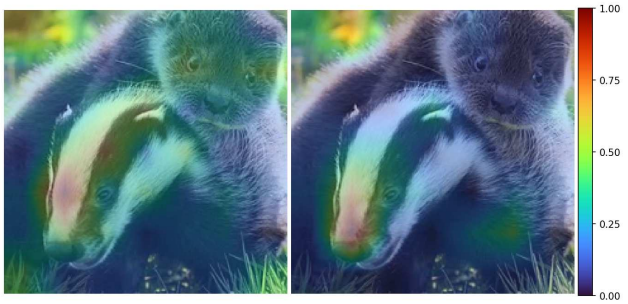

Figure 2. **Qualitative comparison (predicted class: *badger*).** Left: vanilla TransMM assigns relevance across both the badger and the otter. Right: feature-gradient gating concentrates on the badger itself. Much of the suppression is driven by the face/eye region feature (Feature 10968, Layer 6), which detects faces in animals but is not discriminative for the predicted class. See supplementary material for prototype analysis of this feature across additional examples.

3. Experiments

Setup. We evaluate on three datasets: **COVID-QU-Ex** (chest X-rays, 3 classes, ViT-B/16 fine-tuned by Mondal et al. [16]), **Hyperkvasir** (endoscopy, 23 classes, ViT-B/16 fine-tuned by Sanderson and Matuszewski [19]), and **ImageNet-1k** (natural images, 1000 classes, CLIP ViT-B/32 [17]). We use a unified ReLU SAE architecture (expansion factor 64×, L1 sparsity); training details are in Appendix C. For COVID-QU-Ex and Hyperkvasir, SAEs are trained on residual-stream patch activations at layers 2–10; for ImageNet, we use pre-trained SAEs from the Prisma multimodal release [12]. We select the gated-layer set $\mathcal{L}_{\text{SAE}}$ and gate base α on a validation subset of 500 images per dataset using joint validation performance across SaCo, Faithfulness Correlation, and Pixel Flipping (Appendix D, Appendix B), and keep these values fixed for all test-set results. **Direct** and **Random Dict.** use the same $(\mathcal{L}_{\text{SAE}}, \alpha)$ as feature-gradient gating, while other baselines use their standard published settings. Faithfulness is measured via: **SaCo** [23], **Faithfulness Correlation** [2], and **Pixel Flipping** [18]. Unless stated otherwise, perturbations replace masked patches with the per-image mean. We average each metric over n_{trials}=3 trials. Faithfulness Correlation uses 20 random perturbations per image with subset size 20 patches (10 for ViT-B/32). Pixel Flipping perturbs 1 patch per step. Statistical significance is assessed using two-tailed paired t-tests on per-image metric differences relative to Vanilla TransMM.

Baselines. We compare against: **Rollout** [1], which aggregates raw attention across layers recursively; **Grad-CAM** [20], adapted for ViTs using last-layer patch features weighted by class gradients; **TokenTM** [22], which extends TransMM with MLP token transformations for improved localization; **Random Dict.**, a permutation control that keeps the gating pipeline identical but breaks learned feature–direction alignment by permuting SAE decoder columns $D^{(\ell)}$; this tests whether improvements depend on the learned correspondence between feature activations and decoder directions rather than generic gating alone; and **Direct**, which scores patches as $s_t^{(\ell)} = (g_t^{(\ell)})^\top x_t^{(\ell)}$ with identical normalization and gate mapping, isolating the effect of SAE decomposition over raw activation-gradient scoring.

Results. Tab. 1 shows that feature-gradient gating consistently improves SaCo and Faithfulness Correlation across all three datasets, while Pixel Flipping gains are smaller and mixed (Fig. 2 shows a qualitative example). The strongest baseline is **Direct**, which performs especially well on ImageNet, including the best Pixel Flipping score, but is less consistent: on Hyperkvasir it underperforms vanilla TransMM on Faithfulness Correlation and Pixel Flipping, whereas feature-gradient gating maintains gains on both main faithfulness metrics across all datasets. Direct also only produces a scalar token-level score, whereas feature-gradient gating decomposes each patch attribution into per-feature contributions, making explanations auditable at the feature level. The **Random Dict.** control generally underperforms feature-gradient gating and often underperforms vanilla TransMM, suggesting that learned feature structure matters. However, on Hyperkvasir SaCo, Random Dict. captures most of the gain over vanilla TransMM (0.504 vs. 0.510), indicating that generic token reweighting contributes substantially in this setting; the advantage of learned features is more evident on Faithfulness Correlation and the other two datasets. Random Dict. SaCo on Hyperkvasir ranges from 0.466 to 0.508 over five permutation seeds, with the reported value near the upper end.

4. Discussion

4.1. Feature-Level Analysis

To probe the mechanism, we analyze 15,000 active (feature, patch) pairs from ImageNet case studies. Of these, 89.4% are suppressions (gate < 1) and 10.6% boosts (gate > 1), indicating that gains primarily come from suppressing non-discriminative regions rather than amplifying class-relevant ones. Manual inspection of the top 50 suppressed features reveals coherent categories including text/watermark detectors, co-occurring human faces, and generic background textures; other features respond to lower-level edge patterns rather than clear semantic concepts (Appendix E). Because per-patch scores decompose linearly into per-feature contributions (Eq. (3)), users can trace which SAE features drive each region's attribution and compare them against feature prototypes. We do not claim full interpretability by design, but the decomposition supports targeted auditing.

Table 1. **Faithfulness results on test sets.** Bold marks the best value per metric per dataset. Layer configs: COVID-QU-Ex [2,3,4], Hyperkvasir [3,4,6], ImageNet [3,4,9]. All differences vs. Vanilla TransMM are statistically significant ($p < 0.05$, two-tailed paired t-test on per-image metric differences) except where marked with $^\circ$.

Dataset	Model	Method	SaCo↑	F.C.↑	Pixel↓
COVID-QU-Ex	ViT-B/16	Rollout	-0.114	-0.018	120.83
		GradCAM	-0.106	0.002	124.53
		TokenTM	0.307	0.274	94.44
		Vanilla TransMM	0.352	0.310	93.17
		Random Dict.	0.328	0.302	96.34
		Direct	0.349°	0.359	90.09
		Feature-gradient gating (ours)	**0.391**	**0.420**	**84.95**
Hyperkvasir	ViT-B/16	Rollout	0.076	0.131	154.32
		GradCAM	0.335	0.167	146.29
		TokenTM	0.342	0.430	93.73
		Vanilla TransMM	0.441	0.481	94.27
		Random Dict.	0.504	0.468	102.05
		Direct	0.449°	0.454	97.15
		Feature-gradient gating (ours)	**0.510**	**0.525**	**92.53**
ImageNet	CLIP ViT-B/32	Rollout	0.121	0.061	12.54
		GradCAM	0.265	0.186	13.58
		TokenTM	0.325	0.280	8.78
		Vanilla TransMM	0.306	0.275	8.95
		Random Dict.	0.298	0.243	9.49
		Direct	0.374	0.328	**8.62**
		Feature-gradient gating (ours)	**0.399**	**0.351**	8.78

4.2. Impact of SAE Decomposition

Feature-gradient gating replaces $(g_t)^\top x_t$ with $(g_t)^\top \hat{x}_t$, discarding gradient signal in directions outside the SAE reconstruction. The **Direct** baseline shows that raw activation-gradient gating already captures a substantial portion of the benefit and is a strong comparator, especially on ImageNet. However, feature-gradient gating is more consistent across datasets and metrics, and uniquely provides a per-feature decomposition for inspecting which SAE features drive suppression or amplification. One possible explanation is that the SAE reconstruction acts as a denoising step by emphasizing directions captured by learned sparse features, though alternative explanations remain plausible, including the effects of sparsity, non-negativity, or better-conditioned projections. The Random Dictionary baseline indicates that learned feature structure contributes to the gains, but does not isolate the precise mechanism.

4.3. Limitations

SAE training requires an offline preprocessing step and additional memory per model. Results differ across metrics: improvements are strongest on SaCo and Faithfulness Correlation, while Pixel Flipping gains are smaller and mixed. The gate formulation involves design choices (α, median/MAD normalization) with moderate sensitivity; alternative mappings or normalization schemes may improve specific metric trade-offs. Because $\mathcal{L}_{\text{SAE}}$ and α are selected using joint validation performance across SaCo, Faithfulness Correlation, and Pixel Flipping, the evaluation is partially selection-aligned; a stronger evaluation would further test robustness under alternative selection criteria. Cross-dataset comparability is limited by different ViT architectures (fine-tuned ViT-B/16 vs. CLIP ViT-B/32); within-dataset comparisons are controlled.

5. Conclusion

We study the direct integration of SAE feature-gradient signals into ViT attention attribution and find that a simple gating mechanism consistently improves SaCo and Faithfulness Correlation over vanilla TransMM across three datasets. Direct activation-gradient gating remains competitive in some settings, but feature-gradient gating is more consistent overall and additionally provides a per-feature decomposition that supports feature-level auditing. These results suggest that SAE-derived feature signals are a useful ingredient for modulating ViT attribution and that mechanistic interpretability tools can be incrementally combined with existing attribution methods.

References

[1] Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meet-*

ing of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 4190–4197. Association for Computational Linguistics, 2020. 3

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

[3] 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. 1

[4] 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, 2021. 1, 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[6] Maximilian Dreyer, Lorenz Hufe, Jim Berend, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. From what to how: Attributing clip's latent components reveals unexpected semantic reliance. *arXiv preprint arXiv:2505.20229*, 2025. 1

[7] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024. 1

[8] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Representation Learning*, pages 26721–26754, 2025. 1

[9] Sangyu Han, Yearim Kim, and Nojun Kwak. Causal interpretation of sparse autoencoder features in vision, 2025. 1

[10] 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. 1

[11] Sonia Joseph, Praneet Suresh, Ethan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Robert Graham, Danilo Bzdok, Wojciech Samek, and Blake Aaron Richards. Steering CLIP's vision transformer with sparse autoencoders. In *Mechanistic Interpretability for Vision at CVPR 2025 (Non-proceedings Track)*, 2025.

[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. 1, 3, 2

[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, 2023. 1

[14] 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. 1

[15] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024. 1

[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. 1, 3

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 3

[18] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 3

[19] Edward Sanderson and Bogdan J Matuszewski. A study on self-supervised pretraining for vision problems in gastrointestinal endoscopy. *IEEE Access*, 12:46181–46201, 2024. 1, 3

[20] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 3

[21] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models, 2025. 1

[22] Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, and Yan Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10926–10935, 2024. 3

[23] 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. 3

Faithful Attribution in Vision Transformers via Feature-Gradient Gating

Supplementary Material

A. Feature-Gradient Score Derivation

Algebraic Expansion. Starting from Eq. (3):

$$s_t^{(\ell)} = \sum_{k=1}^{K} \left[(D^{(\ell)})^\top g_t^{(\ell)} \right]_k \cdot f_{t,k}^{(\ell)}$$
$$= (g_t^{(\ell)})^\top \sum_{k=1}^{K} D_k^{(\ell)} f_{t,k}^{(\ell)} = (g_t^{(\ell)})^\top \hat{x}_t^{(\ell)}, \quad (6)$$

where $\hat{x}_t^{(\ell)} := D^{(\ell)} f_t^{(\ell)}$ is the bias-free SAE reconstruction.

Relationship to Direct Baseline. Since $x_t^{(\ell)} = \hat{x}_t^{(\ell)} + e_t^{(\ell)}$, the Direct baseline differs by $(g_t^{(\ell)})^\top e_t^{(\ell)}$: gradients in directions orthogonal to learned SAE features. The per-feature decomposition $s_t = \sum_k \tilde{g}_{f,t,k} \cdot f_{t,k}$ exposes individual feature contributions unavailable from the Direct baseline's scalar score.

B. Normalization Details

For a fixed image and layer ℓ, let $\mathcal{S}^{(\ell)} = \{s_t^{(\ell)}\}_{t=1}^N$ denote scores over spatial patches. We compute:

$$\mu^{(\ell)} = \text{median}(\mathcal{S}^{(\ell)}), \quad (7)$$
$$\sigma^{(\ell)} = 1.4826 \cdot \text{median}\Big(|s_t^{(\ell)} - \mu^{(\ell)}| : t \in \{1, \ldots, N\} \Big). \quad (8)$$

The factor 1.4826 makes MAD consistent with standard deviation for Gaussian distributions. We choose median/MAD because they robustly identify the inactive-feature baseline in the heavy-tailed, sparse distribution of feature-gradient scores.

Alternative Normalizations. We evaluated z-score (mean/std) and min-max normalization on COVID-QU-Ex validation data. Median/MAD outperformed both, likely because:

- Z-score is sensitive to outliers from highly-active features
- Min-max compresses the distribution when extreme values are present
- Median/MAD robustly centers on the "typical" patch score

Gate Base Selection. For the exponential base α in Eq. (4), we sweep $\alpha \in \{2, 5, 10, 20\}$ on validation and observe a trade-off between stronger modulation (higher SaCo/F.C.) and more aggressive gating that can hurt Pixel Flipping.

- $\alpha = 2$: Gates too close to unity; minimal modulation effect
- $\alpha = 5$–10: Stronger modulation with stable gains across datasets
- $\alpha \geq 20$: Can increase SaCo/F.C. but may degrade Pixel Flipping

We use $\alpha = 10$ for COVID-QU-Ex and Hyperkvasir, and $\alpha = 5$ for ImageNet.

Table 2. Effect of gate base α on ImageNet validation for fixed layers [3,4,9].

α	SaCo↑	F.C.↑	Pixel↓
2	0.340	0.321	8.149
5	0.384	0.352	**8.139**
10	0.409	**0.360**	8.305
20	**0.426**	**0.360**	8.515

C. Sparse Autoencoder Training

Architecture. All SAEs use a unified ReLU architecture with expansion factor 64× (768 → 49,152 features). The encoder is initialized as the transpose of the decoder. Input activations are normalized with layer norm before encoding. Ghost gradients are enabled to prevent feature collapse during training. The learning rate follows a cosine annealing schedule with a 200-step linear warm-up. SAEs are trained on frozen residual-stream *patch token* activations (CLS token excluded, as our method gates spatial tokens only). Training uses a batch size of 4,096 tokens.

Per-dataset configuration. Hyperparameters are selected by sweeping L1 coefficient and learning rate, choosing the configuration with highest explained variance and no dead features. For ImageNet we use pre-trained SAEs from the Prisma multimodal release [12], trained on CLIP ViT-B/32 `hook_resid_post` activations. Tab. 3 summarises the selected configurations; per-layer quality metrics are in Tab. 4. Complete training scripts, sweep results, and random seeds will be released with the code upon acceptance.

Table 3. Selected SAE training configurations per dataset.

Dataset	L1	LR	Epochs	Source
COVID-QU-Ex	10^{-5}	4×10^{-4}	6	trained
Hyperkvasir	2×10^{-6}	9×10^{-4}	18	trained
ImageNet	10^{-5}	–	–	Prisma [12]

Table 4. SAE per-layer results. All use ReLU activation, expansion 64×, L1 sparsity. EV = explained variance (%), L0 = mean active features per token. †Pre-trained Prisma multimodal SAEs [12].

Dataset	Layer	EV (%)	L0	Dead (%)
COVID-QU-Ex	2	95.5	1036	0.0
	3	95.3	1147	0.0
	4	95.7	1401	0.0
	5	96.8	1719	0.0
	6	97.1	1913	0.0
	7	97.3	2156	0.0
	8	97.5	2281	0.0
	9	97.4	2179	0.0
	10	97.4	1976	0.0
Hyperkvasir	2	94.2	833	0.0
	3	93.8	825	0.0
	4	93.4	814	0.0
	5	94.1	797	0.0
	6	94.6	766	0.0
	7	94.5	736	0.0
	8	99.1	769	0.0
	9	98.5	774	0.0
	10	97.3	785	0.0
ImageNet†	0	98.6	40	0.0
	1	98.3	27	0.0
	2	90.6	10	0.0
	3	98.1	78	0.0
	4	98.0	157	0.0
	5	98.1	229	0.0
	6	98.2	1718	0.0
	7	98.2	1688	0.0
	8	98.2	1571	0.0
	9	98.2	1053	0.0
	10	98.4	1010	0.0
	11	98.4	1189	0.0

D. Single-Layer vs. Multi-Layer Attribution

As shown in Tab. 5, the selected multi-layer sets consistently outperform the best single layer on Faithfulness Correlation, but they need not contain the top single layers by Faithfulness Correlation alone because selection was based on joint validation performance across all three metrics.

E. Feature Taxonomy Visualizations

Not all suppressed features correspond to easily human-interpretable concepts. While many SAE features capture recognizable patterns, including faces (Fig. 3), text/watermarks (Fig. 4), background colors (Fig. 5), and domain-specific detectors (Fig. 6), others respond to low-level structures such as edges or color transitions without a clear semantic label (Fig. 7). These features still receive consistent gradient signals and contribute to gating, demonstrating the method operates beyond human-concept-level semantics.

Table 5. Single-layer validation sweep (Faithfulness Correlation). The selected multi-layer sets were chosen using joint validation performance across SaCo, Faithfulness Correlation, and Pixel Flipping.

Dataset	**Top-3 single layers (F.C.)**	**Selected multi-layer set (SaCo / F.C. / Pixel)**
COVID-QU-Ex	3 (0.357), 4 (0.357), 2 (0.348)	[2,3,4] (0.400 / 0.414 / 86.51)
Hyper-kvasir	6 (0.515), 4 (0.509), 3 (0.481)	[3,4,6] (0.508 / 0.527 / 90.56)
ImageNet	6 (0.336), 5 (0.327), 8 (0.325)	[3,4,9] (0.409 / 0.360 / 8.30)

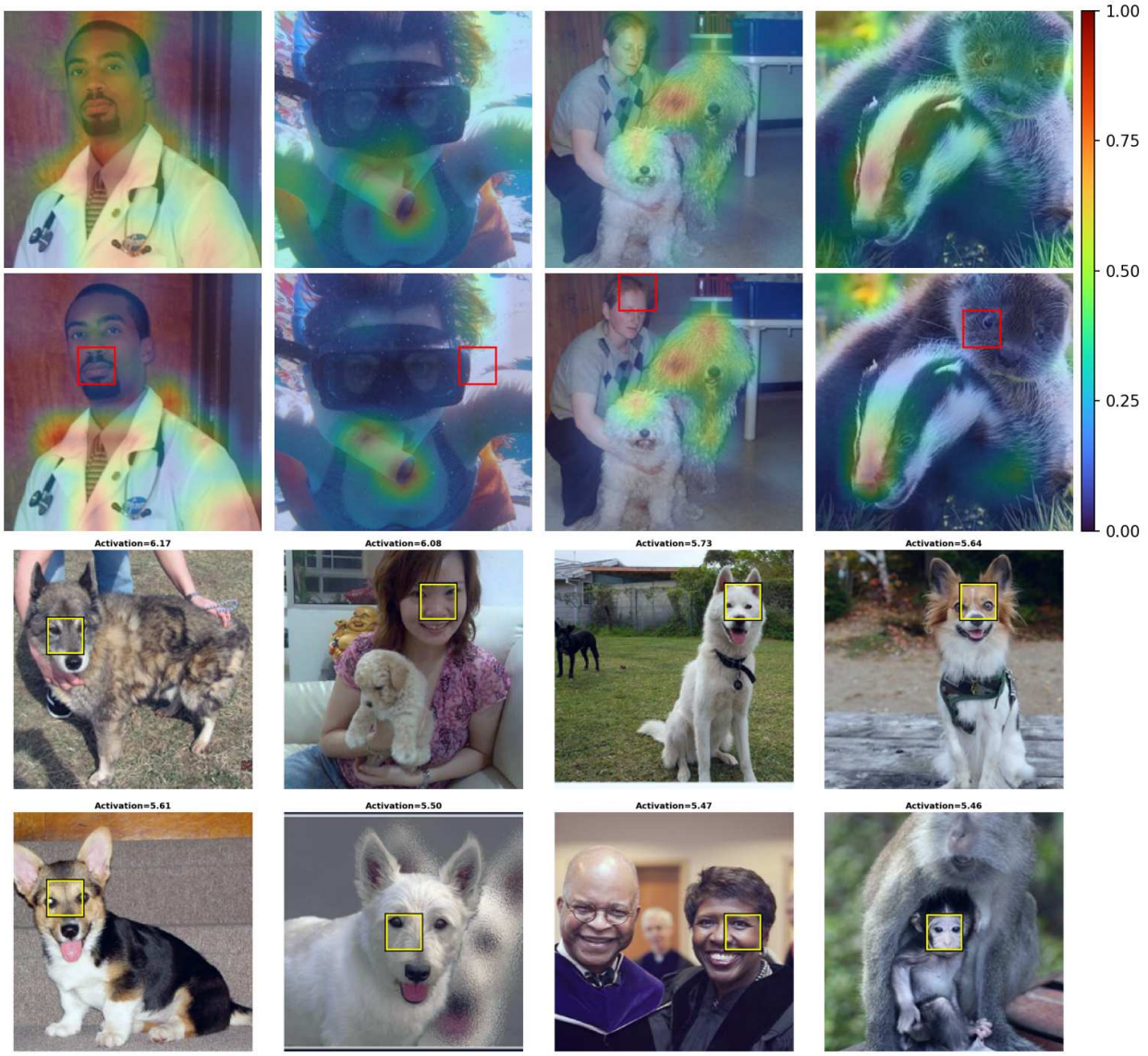

Figure 3. **Face/eye region detector (Feature 10968, Layer 6).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *lab coat*, *snorkel*, *komondor*, *badger*. Red boxes highlight patches where this feature drives the largest suppression. *Bottom rows:* Prototype analysis showing validation images where Feature 10968 activates most strongly; yellow boxes indicate the specific patches triggering the feature response. This feature detects facial regions in both humans and animals. Faces are semantically meaningful but often not class-discriminative: a generic face detector does not distinguish between dog breeds, and human faces are irrelevant for *lab coat* or *snorkel* classification.

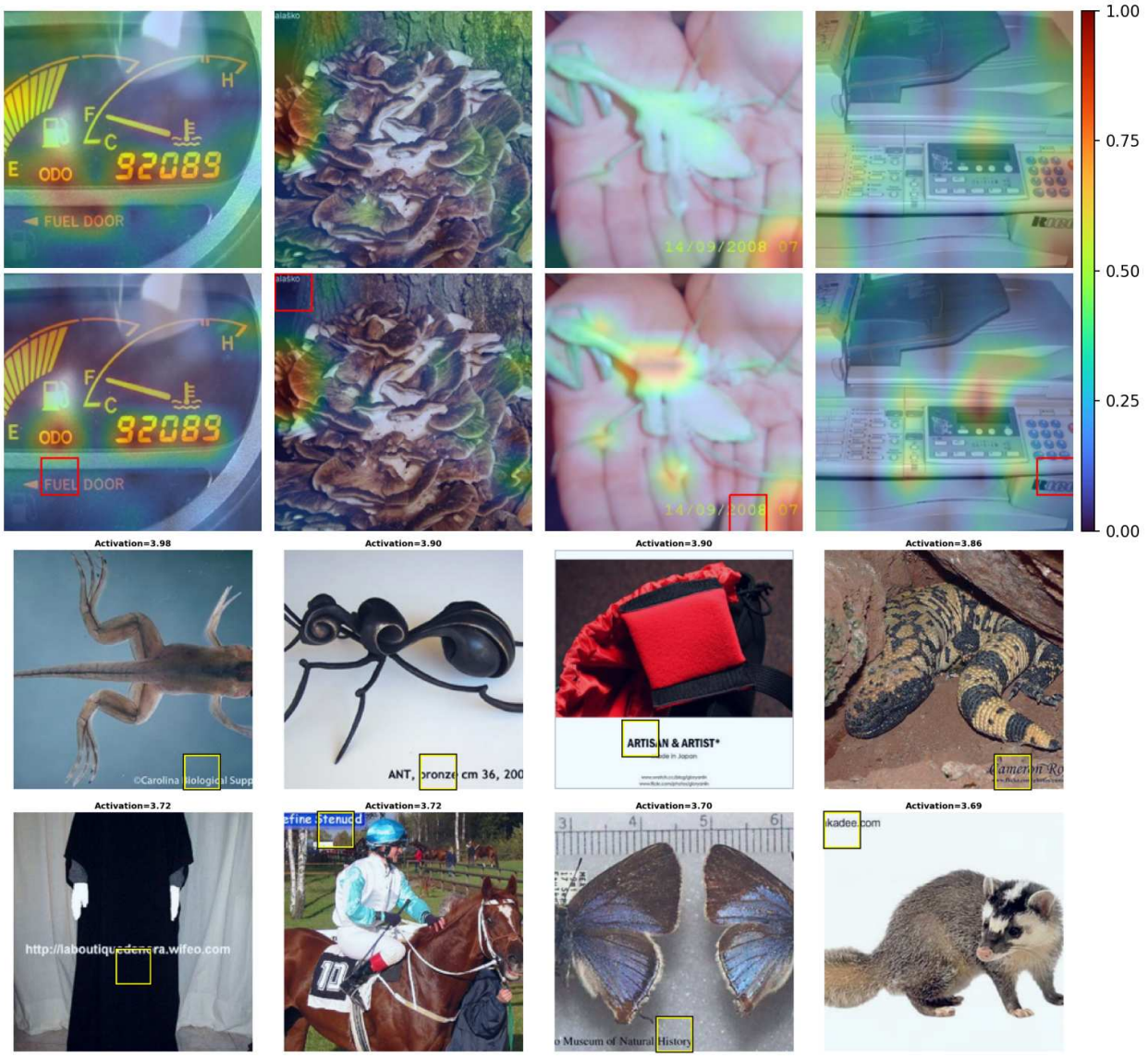

Figure 4. **Suppression of spurious text artifacts (Feature 538, Layer 6).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *odometer*, *hen-of-the-woods*, *mantis*, *photocopier*. Red boxes highlight patches where this feature drives the largest suppression; note how text regions ("FUEL DOOR", "alasko" watermark, date stamp, brand logo) are suppressed despite being visually salient. *Bottom rows:* Prototype analysis showing validation images where Feature 538 activates most strongly; yellow boxes indicate the specific patches triggering the feature response, revealing it detects text overlays and watermarks. This feature receives the strongest negative gradients, often suppressing text regions irrelevant to classification.

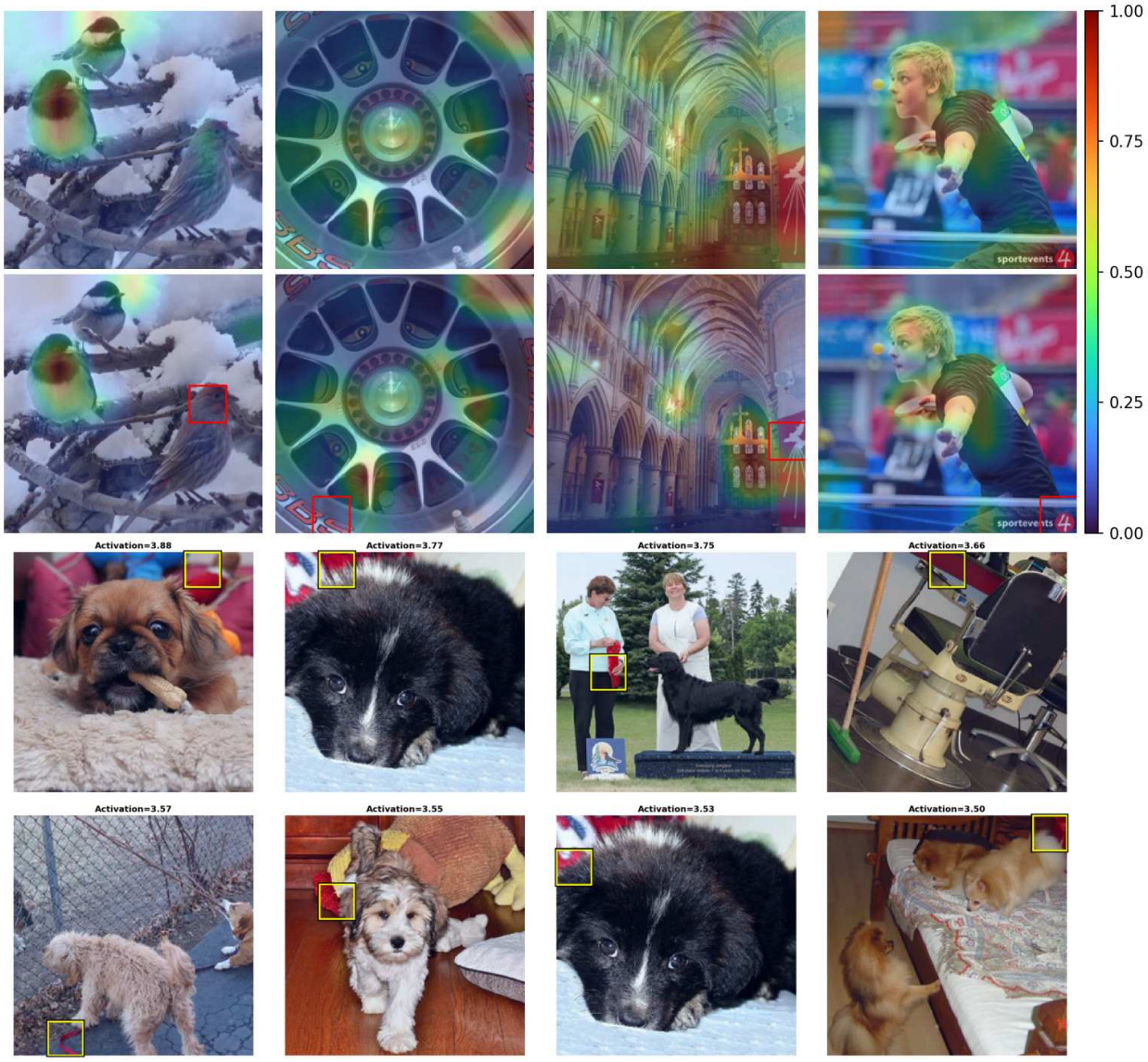

Figure 5. **Red background objects (Feature 21415, Layer 6).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *junco*, *car wheel*, *church*, *ballplayer*. Red boxes highlight patches where this feature drives the largest suppression. *Bottom rows:* Prototype analysis showing validation images where Feature 21415 activates most strongly; yellow boxes indicate the specific patches triggering the feature response. This feature detects red/orange-colored regions. Background colors are contextual but not causally related to class identity.

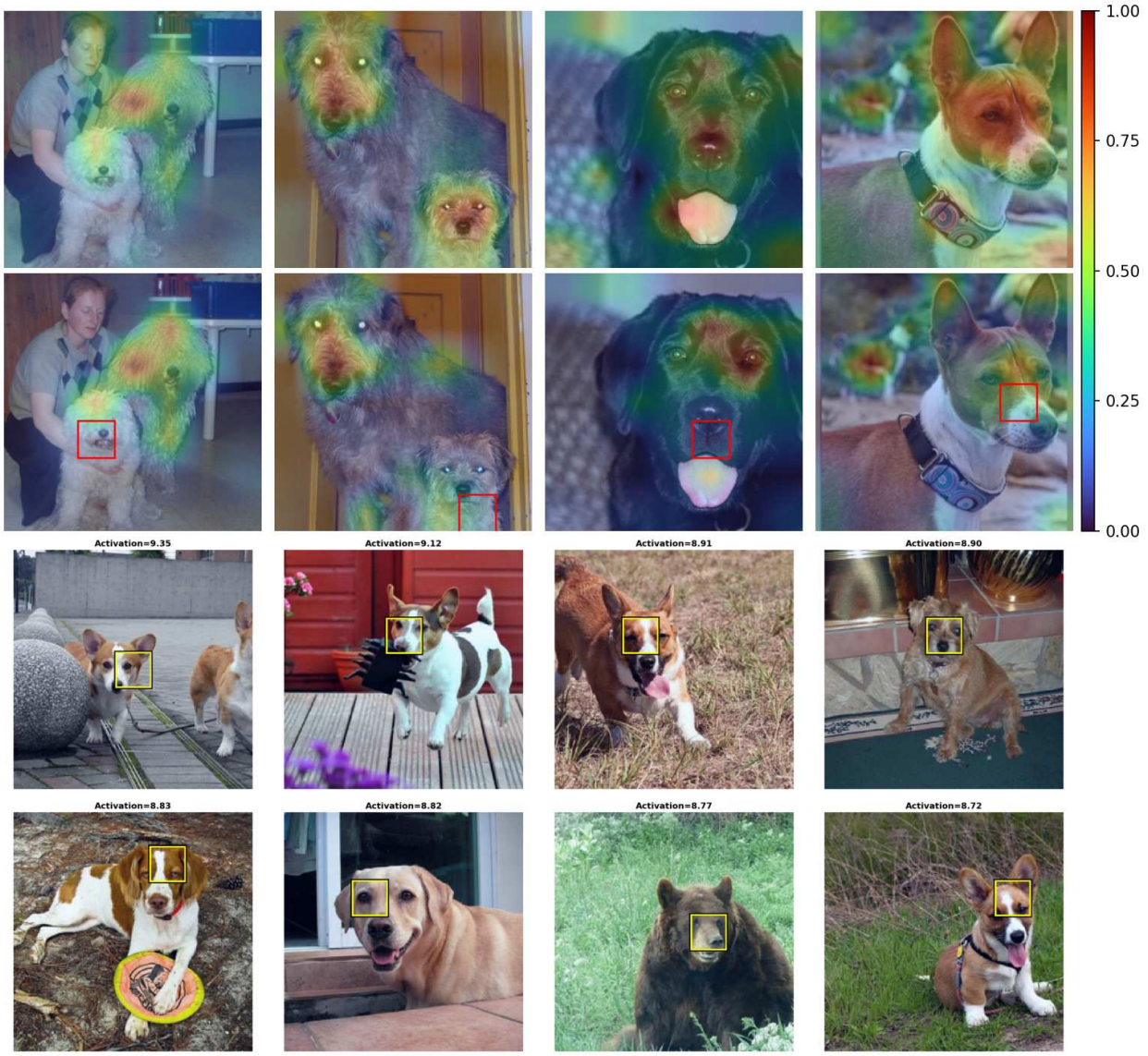

Figure 6. **Dog face detector (Feature 28865, Layer 7).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *komondor*, *Irish wolfhound*, *flat-coated retriever*, *basenji*. Red boxes highlight patches where this feature drives the largest suppression. *Bottom rows:* Prototype analysis showing validation images where Feature 28865 activates most strongly; yellow boxes indicate the specific patches triggering the feature response. This feature detects dog/wolf faces specifically. While relevant for canine classification, a generic "dog face" feature does not discriminate between breeds, so it receives moderate suppression to focus attribution on breed-specific markings (ear shape, coat pattern, muzzle structure).

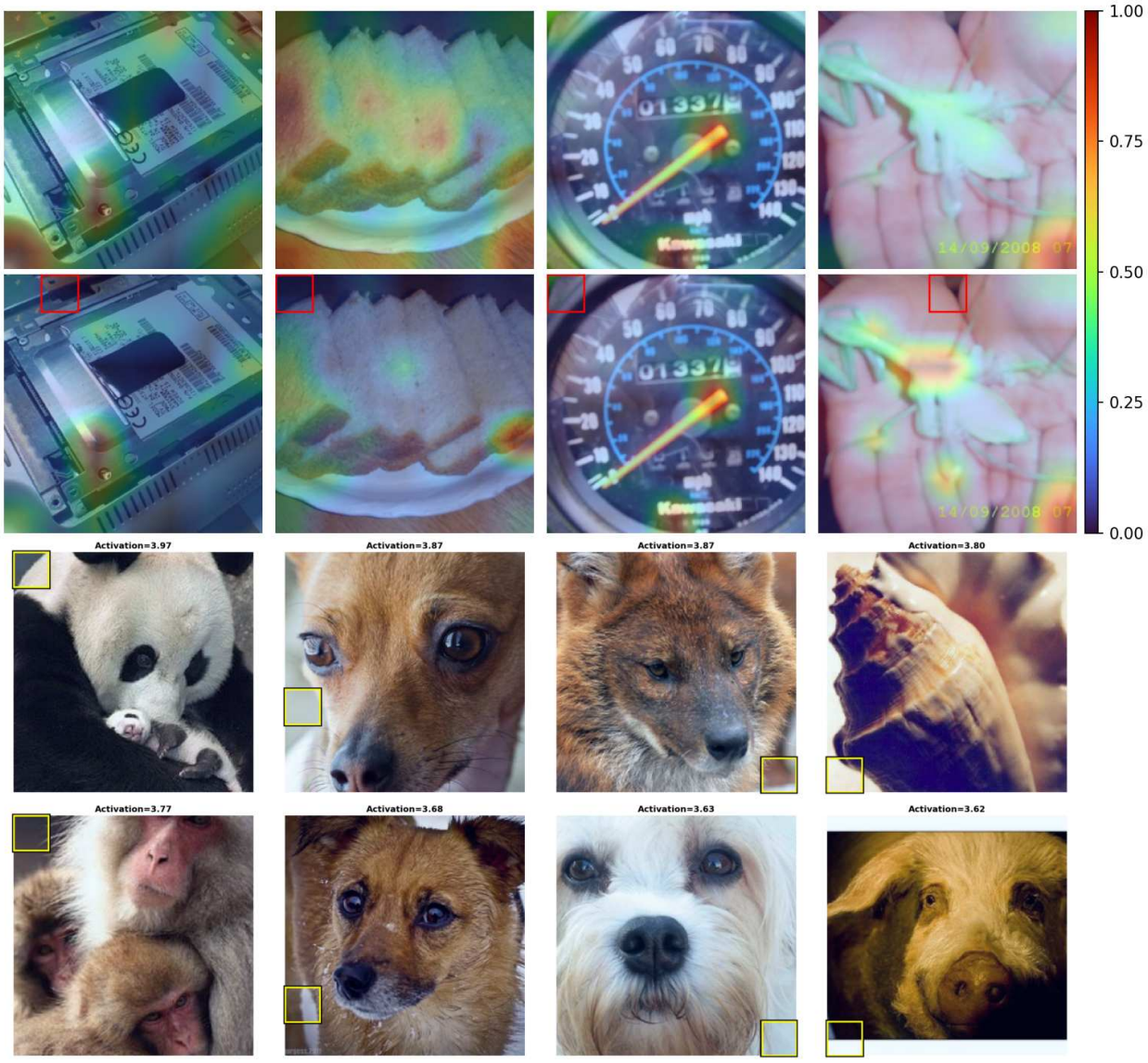

Figure 7. **Edge/ridge detector (Feature 29914, Layer 7).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *hard disc*, *French loaf*, *odometer*, *mantis*. Red boxes highlight patches where this feature drives the largest suppression. *Bottom rows:* Prototype analysis showing validation images where Feature 29914 activates most strongly. Unlike the semantically interpretable features above, this feature responds to low-level edge structures and color transitions rather than recognizable objects. Such features demonstrate that not all SAE directions correspond to human-interpretable concepts, yet they still receive consistent gradient signals that modulate attribution. The suppression of edge-heavy regions suggests the model treats these low-level patterns as non-discriminative for the predicted classes.