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)

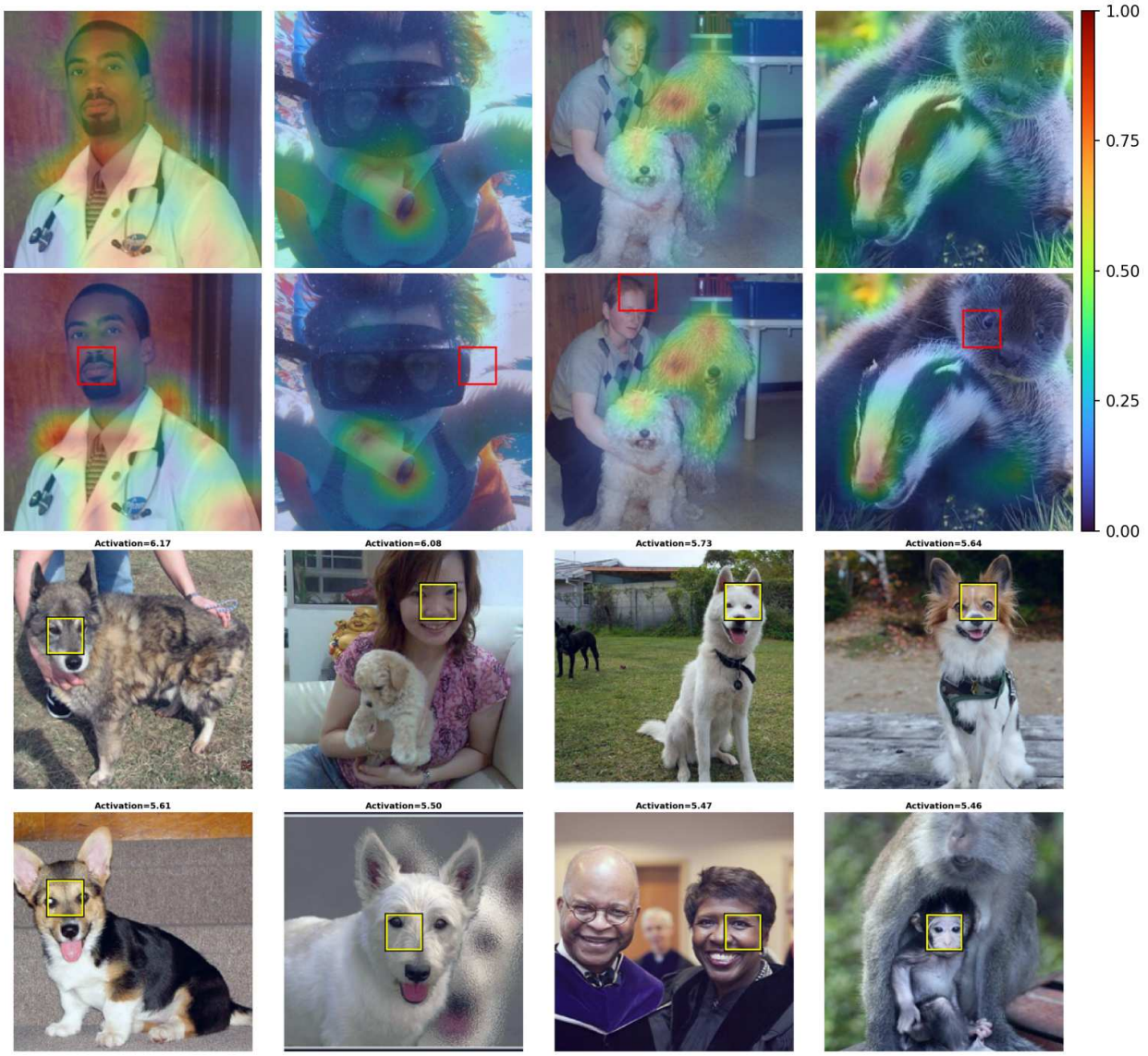

Figure 3. **Face/eye region detector (Feature 10968, Layer 6).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *lab coat*, *snorkel*, *komondor*, *badger*. Red boxes highlight patches where this feature drives the largest suppression. *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.

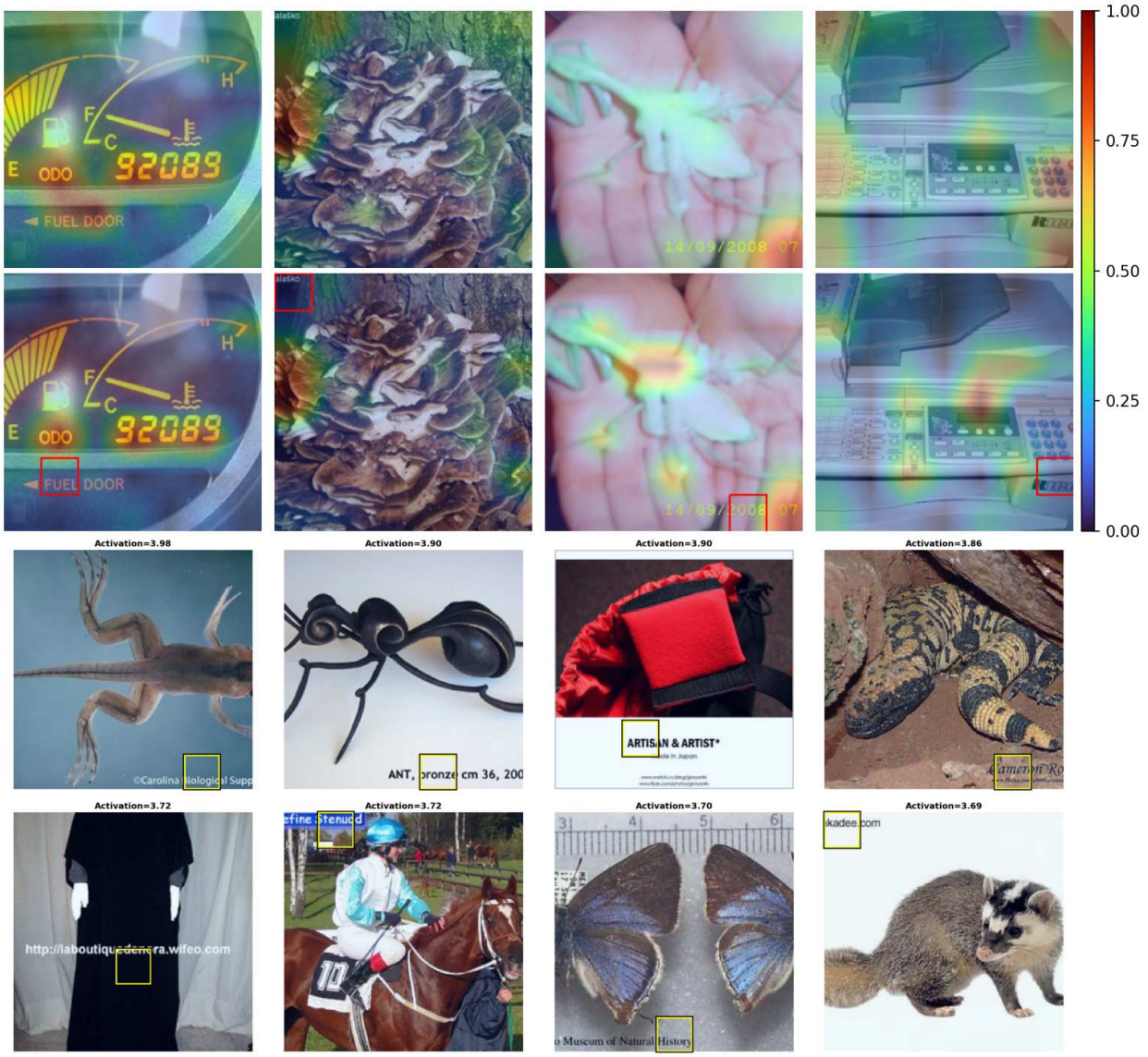

Figure 4. **Suppression of spurious text artifacts (Feature 538, Layer 6).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *odometer*, *hen-of-the-woods*, *mantis*, *photocopier*. Red boxes highlight patches where this feature 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.

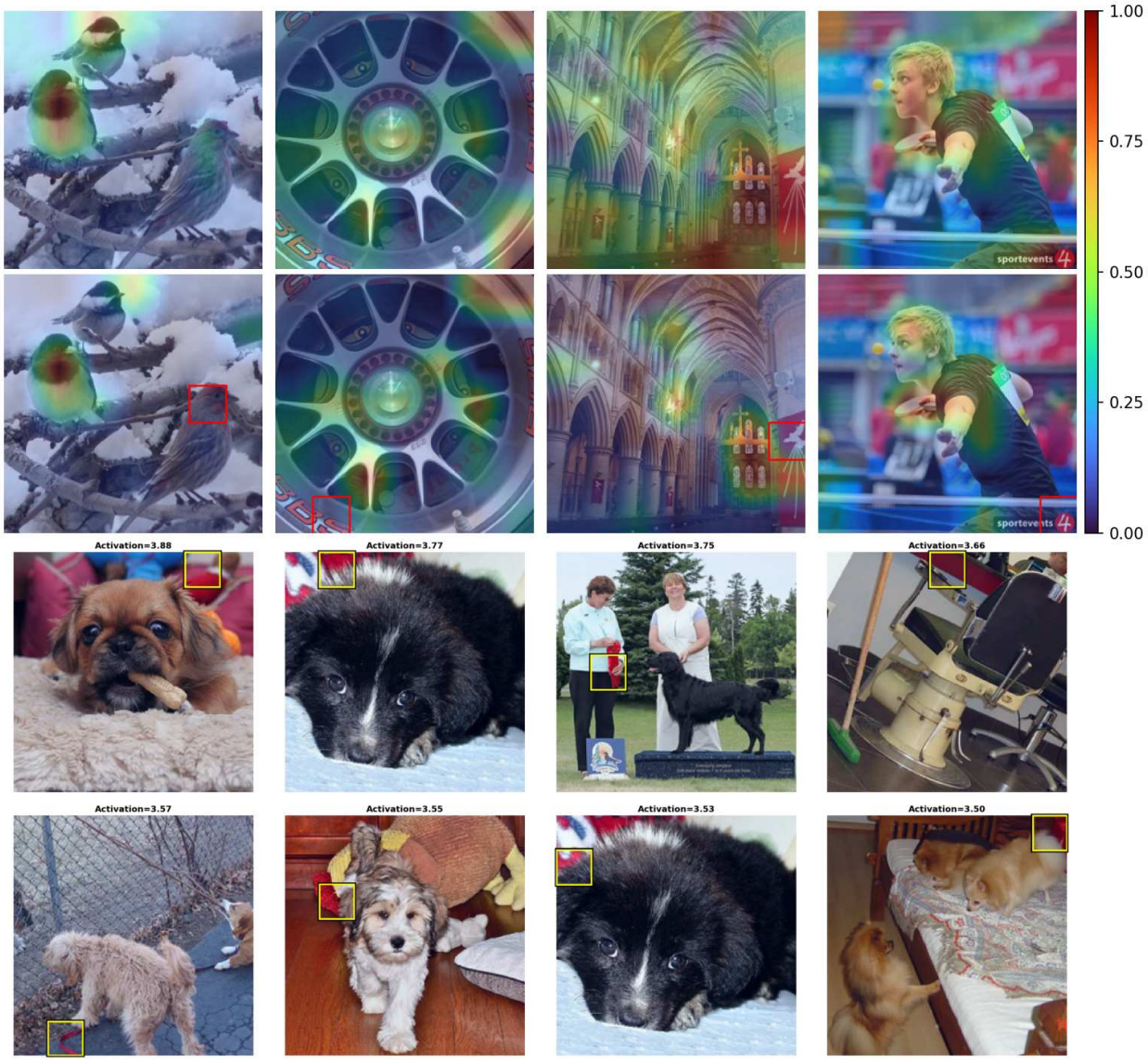

Figure 5. **Red background objects (Feature 21415, Layer 6).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *junco*, *car wheel*, *church*, *ballplayer*. Red boxes highlight patches where this feature drives the largest suppression. *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.

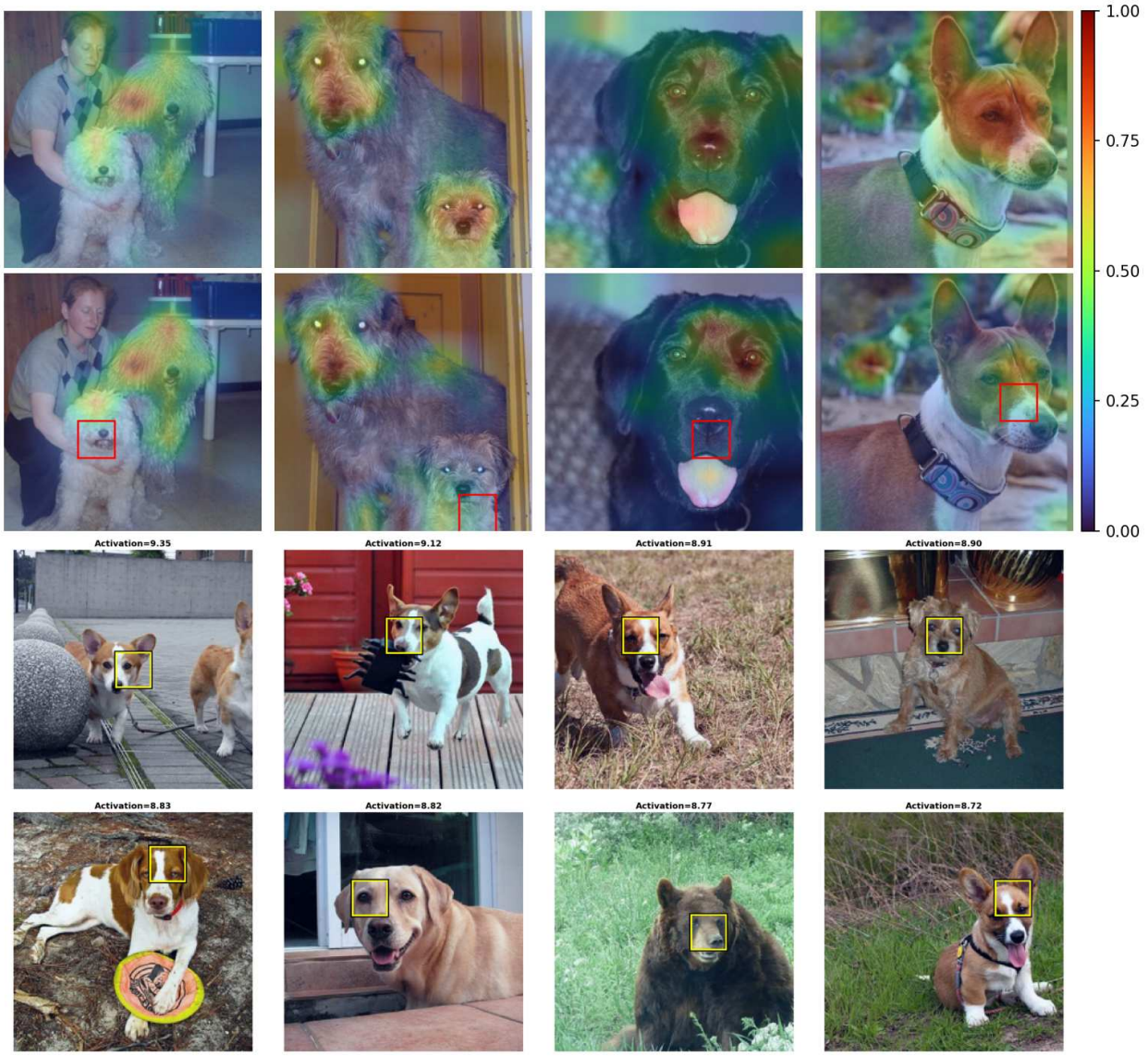

Figure 6. **Dog face detector (Feature 28865, Layer 7).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *komondor*, *Irish wolfhound*, *flat-coated retriever*, *basenji*. Red boxes highlight patches where this feature drives 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).

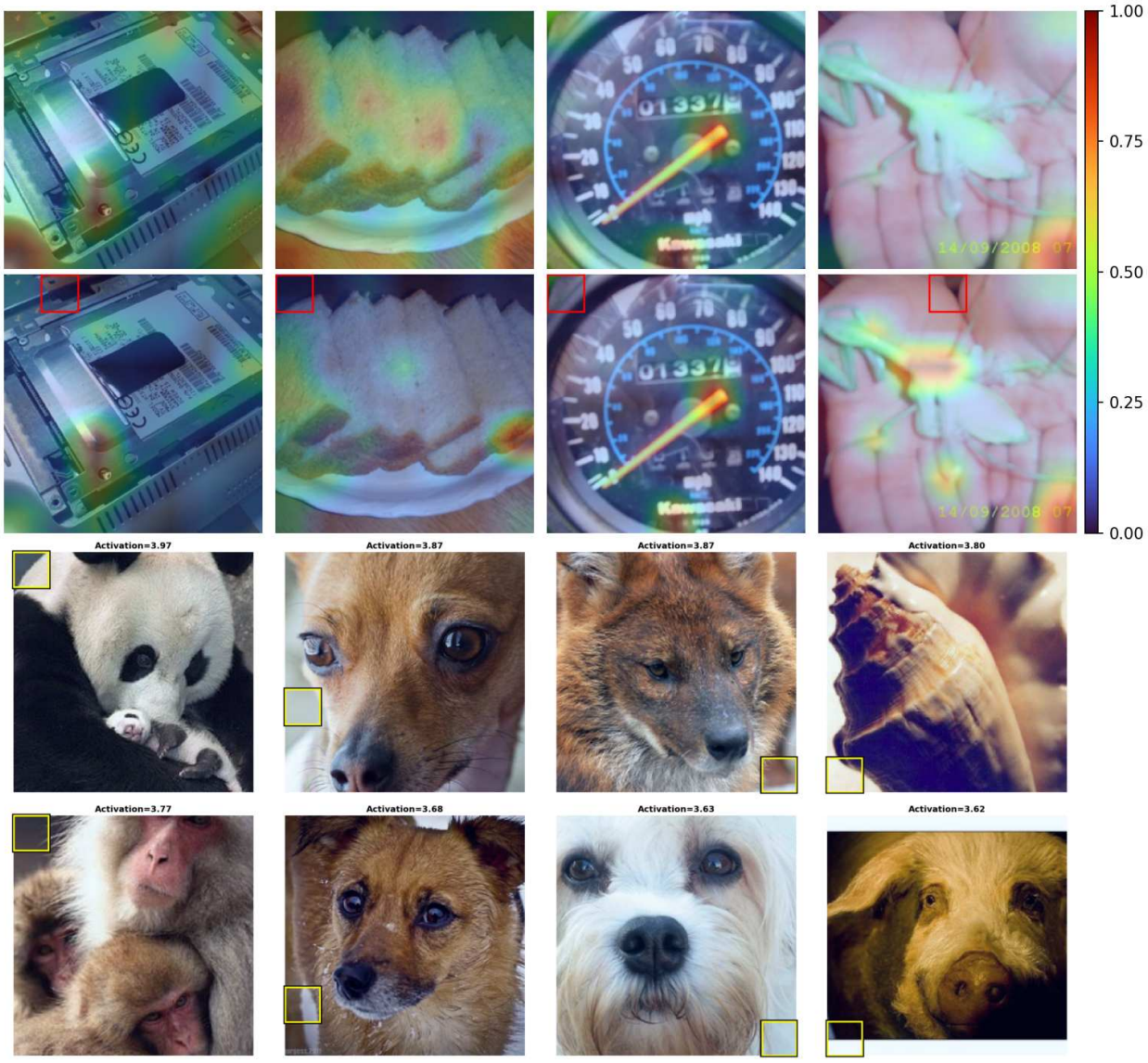

Figure 7. **Edge/ridge detector (Feature 29914, Layer 7).** *Top rows:* Attribution comparison showing vanilla TransMM vs. our feature-gated method. Predicted classes left-to-right: *hard disc*, *French loaf*, *odometer*, *mantis*. Red boxes highlight patches where this feature drives the largest suppression. *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.