This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.



# Supplementary Table S1: Lesion-Disjoint Split Distribution Audit

To verify that the deterministic lesion-disjoint split is not a class-composition outlier, we sampled 10,000 random lesion-disjoint validation sets of the same size (700 lesions, 1,400 images). The fixed sorted-ID split fell within the random distribution for all six classes (maximum absolute $z$-score 0.71, all $p > 0.47$), indicating that the fixed split does not introduce class-composition confounds.

| Class | Fixed count | Random mean | Random 95% CI | $z$-score |
|---|---|---|---|---|
| MEL | 268 | 270.4 | [242, 299] | −0.30 |
| BCC | 204 | 200.2 | [174, 227] | +0.41 |
| SCC | 208 | 204.8 | [178, 232] | +0.39 |
| ACK | 256 | 258.6 | [230, 287] | −0.27 |
| NEV | 252 | 247.9 | [220, 276] | +0.44 |
| SEK | 212 | 218.1 | [191, 246] | −0.71 |

# Implementation and Reproducibility Details

**Datasets.** MILK10k (9,344 images, 4,672 lesions, two images per lesion) is used for in-domain training. The deterministic clean split sorts lesion IDs and reserves the final 15% (700 lesions, 1,400 images) for validation, yielding zero lesion overlap. PAD-UFES-20 ($N$=2298) is the OOD test set; clinical text uses only age, sex, and anatomical site to avoid biopsy-field leakage. DDI ($N$=343) provides Fitzpatrick-stratified fairness evaluation. Text-Only accuracy across five random stratified lesion-disjoint splits averaged 46.28% (95% CI 33.23–59.32%), below the 53.92% majority baseline ($\Delta$=−7.65 pp, $p$=0.179).

**Training Hyperparameters.**

| Setting | Value |
|---|---|
| Optimizer | AdamW |
| Loss | Cross-entropy |
| Learning rate | $2 \times 10^{-5}$ |
| Weight decay | 0.01 |
| Batch size | 16 |
| Maximum epochs | 5 |
| Early stopping | Validation-loss patience of 2 epochs |
| Warmup | Linear schedule, 10% warmup ratio |
| Gradient clipping | Maximum norm 1.0 |
| Mixed precision | Enabled on CUDA |

**Architecture Specifications.** All discriminative models use frozen ViT-Base-patch16-224 (768-d) and frozen Bio_ClinicalBERT (768-d) encoders.

| Model | Specification |
|---|---|
| Image-Only | Frozen ViT-B/16 $\rightarrow$ linear 6-class head. |
| Text-Only | Frozen BioClinicalBERT CLS $\rightarrow$ linear 6-class head. |
| Late Fusion | $[h_v; h_t]$ (1536-d) $\rightarrow$ 512-d hidden, ReLU, dropout 0.1, 6-class head. |
| GMU | Visual and text projected to 512-d; sigmoid gate from 1536-d concat; 6-class head. |
| CA $V \rightarrow T$ | Single 8-head cross-attention, visual query attends to text keys/values; 6-class head. |
| CA $T \rightarrow V$ | Single 8-head cross-attention, text queries attend to ViT patch tokens (mean-pooled); 6-class head. |
| Top-K SAE | Trained on mean-pooled fused residual of CA $T \rightarrow V$; $d_{\text{SAE}}$=6144, $k$=32. |

# Mathematical Derivations

**Lesion Leakage Under Image-Level Random Splitting.** MILK10k contains two images per lesion. Under image-level random splitting with $p_{\text{val}}$=0.15:

$$P(\text{paired image in train} \mid \text{one image in validation}) = 1 - p_{\text{val}} = 0.85.$$

A lesion-disjoint split removes this by construction: $\mathcal{L}_{\text{train}} \cap \mathcal{L}_{\text{val}} = \varnothing$.

**Matched Grounding Margin.** For a MEL/NEV pair, the per-image margin under text $t$ is $m_i(t) = f_{\text{MEL}}(x_i, t) - f_{\text{NEV}}(x_i, t)$. The grounding shift is $\Delta_i = m_i(t_i^{\text{aligned}}) - m_i(t_i^{\text{contradictory}})$. Positive $\Delta_i$ indicates aligned text increases the true-class margin. The reported real-image mean shift was $-0.0102$, i.e. aligned text slightly *decreased* the correct-class margin.

**SAE Feature Knockout Estimand.** Let $z_i$ be the fused representation, $a_{ij}$ the activation of SAE feature $j$, and $d_j$ its decoder direction. The knockout constructs $z_i^{(-j)} = z_i - a_{ij} d_j$. The reported causal estimand for Feature 1449 is the mean change in malignant-minus-benign logit contrast:

$$\Delta_{\text{mal-ben}} = \frac{1}{N} \sum_{i=1}^{N} \Big[ \big( \overline{f}_{\text{mal}}(z_i^{(-1449)}) - \overline{f}_{\text{ben}}(z_i^{(-1449)}) \big) - \big( \overline{f}_{\text{mal}}(z_i) - \overline{f}_{\text{ben}}(z_i) \big) \Big].$$

The random active-feature control uses the same estimator with a non-target active feature selected per row. Decoded-direction subtraction can move representations off manifold; this estimand is treated as causal-intervention evidence for a candidate shortcut-related feature, not as proof of an isolated artifact mechanism.