Setup and Notation

Let $f : \mathbb{R}^{3\times224\times224} \to \mathbb{R}^K$ be the primary ResNet-50 classifier with $K = 10$ ID classes. For an input image $\boldsymbol{x}$, denote the logit vector $\boldsymbol{z} = f(\boldsymbol{x}) \in \mathbb{R}^K$, the softmax output $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z})$, the penultimate-layer embedding $\boldsymbol{h} = g(\boldsymbol{x}) \in \mathbb{R}^{2048}$ (output of the global average pool), and the predicted class $\hat{y} = \arg\max_k z_k$.

Integrated Gradients

For a single image $\boldsymbol{x}$, a baseline $\boldsymbol{x}'$ (we use $\boldsymbol{x}' = \mathbf{0}$), and a target class c, the Integrated Gradients attribution map is:

$$A_c(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}') \odot \int_{\alpha=0}^{1} \frac{\partial f_c(\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}'))}{\partial \boldsymbol{x}}\, d\alpha \tag{1}$$

where f_c is the c-th logit and the integral is approximated with $n = 50$ Riemann steps. Each $A_c(\boldsymbol{x}) \in \mathbb{R}^{3\times224\times224}$.

Single-class baseline. The original pipeline computed $A_{\hat{y}}(\boldsymbol{x})$ only—attributions for the predicted class. For OOD images, $\hat{y}$ is arbitrary (whichever ID class the model is most confused by), making the resulting map inconsistent across semantically similar OOD inputs.

Multi-class profiles. We instead compute attributions for *all* K classes:

$$\mathcal{A}(\boldsymbol{x}) = \big[A_0(\boldsymbol{x}),\ A_1(\boldsymbol{x}),\ \ldots,\ A_{K-1}(\boldsymbol{x})\big] \in \mathbb{R}^{K\times3\times224\times224} \tag{2}$$

This stack captures how the model responds across its entire output space for a single input.

Per-Class Feature Extraction

For each class c, let $M_c = A_c(\boldsymbol{x}) \in \mathbb{R}^{3\times H\times W}$. We extract 55 features $\boldsymbol{\phi}_c = \phi(M_c)$ organized as follows.

Channel statistics (8 per channel $\times$ 3 = 24). For channel $j \in \{R, G, B\}$ with values $\{m_{j,u,v}\}$:

- Mean: $\mu_j = \frac{1}{HW}\sum_{u,v} m_{j,u,v}$, Std: $\sigma_j = \sqrt{\frac{1}{HW}\sum_{u,v}(m_{j,u,v} - \mu_j)^2}$
- Max, min, mean of $|m_{j,u,v}|$, fraction positive: $\frac{1}{HW}|\{(u,v) : m_{j,u,v} > 0\}|$
- Skewness: $\frac{1}{HW}\sum_{u,v}\big(\frac{m_{j,u,v}-\mu_j}{\sigma_j}\big)^3$, Excess kurtosis: $\frac{1}{HW}\sum_{u,v}\big(\frac{m_{j,u,v}-\mu_j}{\sigma_j}\big)^4 - 3$

Spatial energy distribution (13). Define the energy map $E(u,v) = \sum_j |m_{j,u,v}|$ and total energy $\mathcal{E} = \sum_{u,v} E(u,v)$.

- Quadrant energies: $q_{TL}, q_{TR}, q_{BL}, q_{BR}$ (fraction of $\mathcal{E}$ in each image quadrant)
- Radial rings: normalized distance $d(u,v) = \sqrt{(u - H/2)^2 + (v - W/2)^2}\,/\,d_{\max}$. Ring energies r_{center} ($d < 0.25$), r_{mid} ($0.25 \le d < 0.5$), r_{edge} ($d \ge 0.5$), and ratio $r_{\text{center}}/r_{\text{edge}}$
- Row/column energy spread: CV$(\sum_v E(u,v))$ and CV$(\sum_u E(u,v))$ where CV $= \sigma/\mu$
- Diagonal ratio, horizontal/vertical balance

Frequency domain (6). Apply the 2D Type-II DCT to E, partition coefficients into low ($\le H/4$), mid, and high frequency bands. Report absolute and fractional energy per band.

Gradient features (4). Compute finite-difference gradients $\nabla_u E, \nabla_v E$, then gradient magnitude $G = \sqrt{(\nabla_u E)^2 + (\nabla_v E)^2}$. Extract: mean G, entropy of normalized G, fraction of high-gradient pixels ($>$ 90th percentile), and anisotropy $\overline{|\nabla_u E|}\,/\,\overline{|\nabla_v E|}$.

Global statistics (8). ℓ_1 and ℓ_2 norms, peak-to-mean ratio, spatial entropy $H(E) = -\sum_{u,v}\tilde{E}(u,v)\log\tilde{E}(u,v)$ where $\tilde{E}$ is normalized, top-1%/5% pixel concentration, sparsity, dynamic range.

Cross-Class Features

Given the per-class energies $\mathcal{E}_c = \sum_{j,u,v}|A_c(\boldsymbol{x})_{j,u,v}|$ for $c = 0, \ldots, K-1$:

Energy distribution (8). Let $\boldsymbol{e} = [\mathcal{E}_0, \ldots, \mathcal{E}_{K-1}]$ and $\hat{\boldsymbol{e}} = \boldsymbol{e}/\|\boldsymbol{e}\|_1$. We compute:

- Dominance ratio: $\mathcal{E}_{(1)}/\mathcal{E}_{(2)}$ where $\mathcal{E}_{(1)} \ge \mathcal{E}_{(2)} \ge \cdots$ are sorted
- Top-1 and top-2 fractions: $\hat{e}_{(1)}$ and $\hat{e}_{(1)} + \hat{e}_{(2)}$
- Energy entropy: $H(\hat{\boldsymbol{e}}) = -\sum_c \hat{e}_c \log \hat{e}_c$
- Max, min, std of $\boldsymbol{e}$, and max/min ratio

Spatial correlation (4). Let $\boldsymbol{v}_c = \text{vec}(E_c)/\|\text{vec}(E_c)\|_2$ where $E_c(u,v) = \sum_j |A_c(\boldsymbol{x})_{j,u,v}|$. The pairwise cosine similarity matrix is $S_{ij} = \boldsymbol{v}_i^\top \boldsymbol{v}_j$. From the upper triangle we extract: mean, std, max, and min correlation.

Intuition. An ID image produces one dominant class with high $\hat{e}_{(1)}$, low $H(\hat{\boldsymbol{e}})$, and low mean cross-class correlation (the dominant map attends to class-specific regions while others are diffuse). An OOD image produces similar energies across classes ($H(\hat{\boldsymbol{e}}) \to \log K$), high cross-class correlation (all maps attend to the same confused regions), and low dominance ratio.

Softmax features (13). The full vector $\boldsymbol{p} \in \mathbb{R}^K$, entropy $H(\boldsymbol{p})$, margin $p_{(1)} - p_{(2)}$, and $\max_k p_k$.

Total: $55K + 19 + 13 = 582$ features per image.

Detector and Normalization

Features are standardized using mean and standard deviation computed from ID training data only (since OOD statistics are unavailable at deployment). A 3-layer MLP (256→128→64→2) with BatchNorm, GELU activations, and dropout (0.4/0.3/0.2) is trained with cross-entropy loss, balanced class sampling via `WeightedRandomSampler`, OneCycleLR scheduling (10% warmup, peak LR 2×10^{-3}), and early stopping on validation loss.

Results

Method	Setting	Acc	F1	AUROC	AUPRC
Single-class IG (24 feat)	Near	72.6	.706	.797	.811
Single-class IG (55 feat)	Near	73.8	.724	.821	.831
Single-class IG (58 feat)	Near	76.8	.751	.850	.862
Multi-class IG (582 feat)	Near	**81.3**	**.814**	**.898**	**.904**
Single-class IG (24 feat)	Far	78.5	.779	.870	.878
Single-class IG (55 feat)	Far	81.6	.815	.897	.900
Single-class IG (58 feat)	Far	86.4	.862	.937	.941
Multi-class IG (582 feat)	Far	**89.9**	**.901**	**.962**	**.963**
Single-class IG (24 feat)	Both	73.0	.779	.815	.899
Single-class IG (55 feat)	Both	76.4	.810	.850	.917
Single-class IG (58 feat)	Both	79.5	.834	.885	.941
Multi-class IG (582 feat)	Both	**82.3**	**.861**	**.914**	**.956**

Table 1: Full ablation results. Each row adds one component to the previous.

Ablation Analysis

The total improvement from the 24-feature single-class baseline to the full multi-class IG pipeline decomposes into three roughly equal steps (Table 2, Figure 1).

Step	Near	Far	Both	What changed
24 feat (baseline)	.797	.870	.815	6 channel + 6 global stats
Δ 24→55	+.024	+.027	+.035	Add skewness, kurtosis, DCT, gradients, spatial quadrants, radial rings
Δ 55→58	+.030	+.040	+.035	Add predicted confidence, softmax entropy approx., margin
Δ 58→Multi	+.047	+.025	+.029	10 IG targets + 19 cross-class features + 13 softmax features
Total	.797→.898	.870→.962	.815→.914	

Table 2: AUROC improvement decomposition. Each row shows the marginal gain from one change.

Richer features (24→55) contribute +0.024–0.035 AUROC. The additional spatial, frequency, and gradient features provide modest but consistent gains. DCT band energies and radial ring profiles capture structure that simple channel statistics miss.

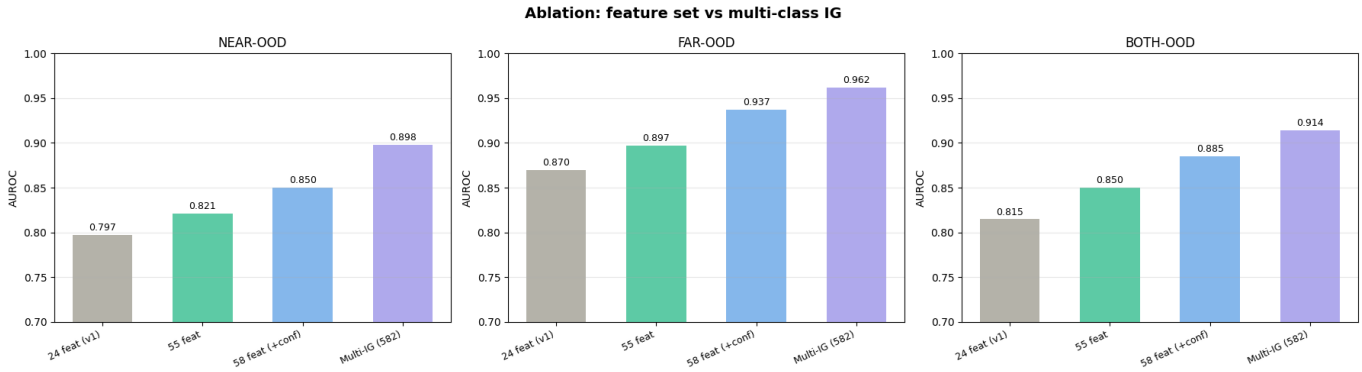

Figure 1: AUROC ablation across all three OOD settings. Each bar group adds one component: richer per-map features (24→55), model confidence (55→58), and multi-class IG profiles (58→582). All three steps contribute, with the multi-class profile providing the largest gain on the hardest setting (near-OOD).

Model confidence (55→58) contributes +0.030–0.040 despite adding only 3 features. This is the most efficient step per-feature. Even a simple confidence/entropy/margin triple carries substantial OOD signal, confirming that the primary model's uncertainty is informative.

Multi-class IG profile (58→Multi) contributes +0.025–0.047, with the largest gain on near-OOD (+0.047) where the task is hardest. This is the only step that requires additional IG computation (10× cost) and cannot be replicated from single-class maps. The cross-class features—energy entropy, dominance ratio, and pairwise spatial correlation—provide discriminative power that single-target attributions fundamentally lack.

Comparison with Embedding-Based Methods

We evaluated post-hoc methods operating on the 2048-dim penultimate-layer features, requiring no attribution computation (Table 3).

Method	Near-OOD AUROC	Training required
Max Softmax Prob. (MSP)	.776	No
Softmax Entropy	.772	No
Energy Score	.759	No
ViM	.618	No
Mahalanobis Distance	.618	No
Single-class IG (24 feat)	.797	Yes
Multi-class IG (582 feat)	**.898**	Yes

Table 3: Embedding-based methods vs. IG-based detection on near-OOD. The best embedding method (MSP) falls below even the simplest IG baseline, confirming that attribution maps carry spatial information absent from global average-pooled features.

On near-OOD, the best embedding method (MSP) achieves only 0.776 AUROC—below even the 24-feature single-class IG baseline (0.797). This confirms that the attribution maps carry spatial information not present in the global average-pooled embeddings. For semantically similar near-OOD pairs (e.g., hammerhead vs. great white shark), the embeddings land in similar feature-space regions, but the *spatial attention patterns* differ enough for IG-based detection to discriminate.

Error Analysis

Failures concentrate on specific semantic pairs (Table 4). For near-OOD, the most-missed OOD classes and most false-alarmed ID classes are symmetric partners.

Animal pairs (tiger/lion at 9%/4%, house finch/goldfinch at 8%/3%) have much lower error rates than object pairs (racer/sports car at 33%/36%), suggesting the classifier learns more discriminative spatial features for animals than for manufactured objects. Two ID classes—banjo (35% FP even against far-OOD) and teapot (30%)—consistently produce diffuse attribution maps resembling OOD patterns, motivating class-conditional detection thresholds as a future improvement.

Near-OOD class	FN%		ID counterpart	FP%
racer	33%	$\longleftrightarrow$	sports car	36%
coffeepot	25%	$\longleftrightarrow$	teapot	26%
timber wolf	25%	$\longleftrightarrow$	German shepherd	6%
acoustic guitar	22%	$\longleftrightarrow$	banjo	25%
lynx	20%	$\longleftrightarrow$	tabby cat	9%
airship	19%	$\longleftrightarrow$	airliner	26%
hammerhead shark	13%	$\longleftrightarrow$	great white shark	22%
vulture	12%	$\longleftrightarrow$	bald eagle	29%
tiger	9%	$\longleftrightarrow$	lion	4%
house finch	8%	$\longleftrightarrow$	goldfinch	3%

Table 4: Near-OOD error rates by semantic pair. FN% = fraction of that OOD class missed by the detector. FP% = fraction of the ID counterpart incorrectly flagged as OOD. Object pairs (top) show much higher confusion than animal pairs (bottom).

For far-OOD, errors are distributed more uniformly across classes (8–11% FN rate for most OOD classes), with the same two ID classes (banjo, teapot) dominating false positives (Table 5).

Far-OOD class	FN%	ID class	FP%
broccoli	11%	banjo	35%
jellyfish	11%	teapot	30%
cannon	11%	sports car	15%
wall clock	11%	great white shark	9%
scorpion	11%	bald eagle	7%
computer keyboard	8%	airliner	6%
cellular telephone	8%	German shepherd	5%
canoe	8%	tabby cat	3%
school bus	5%	goldfinch	2%
orange	3%	lion	1%

Table 5: Far-OOD error rates. FN% for OOD classes (left) and FP% for ID classes (right), sorted by error rate. Errors are more uniform across OOD classes but concentrated in the same ID classes (banjo, teapot) as near-OOD.