Spurious-Aware Prototype Refinement for Reliable Out-of-Distribution Detection

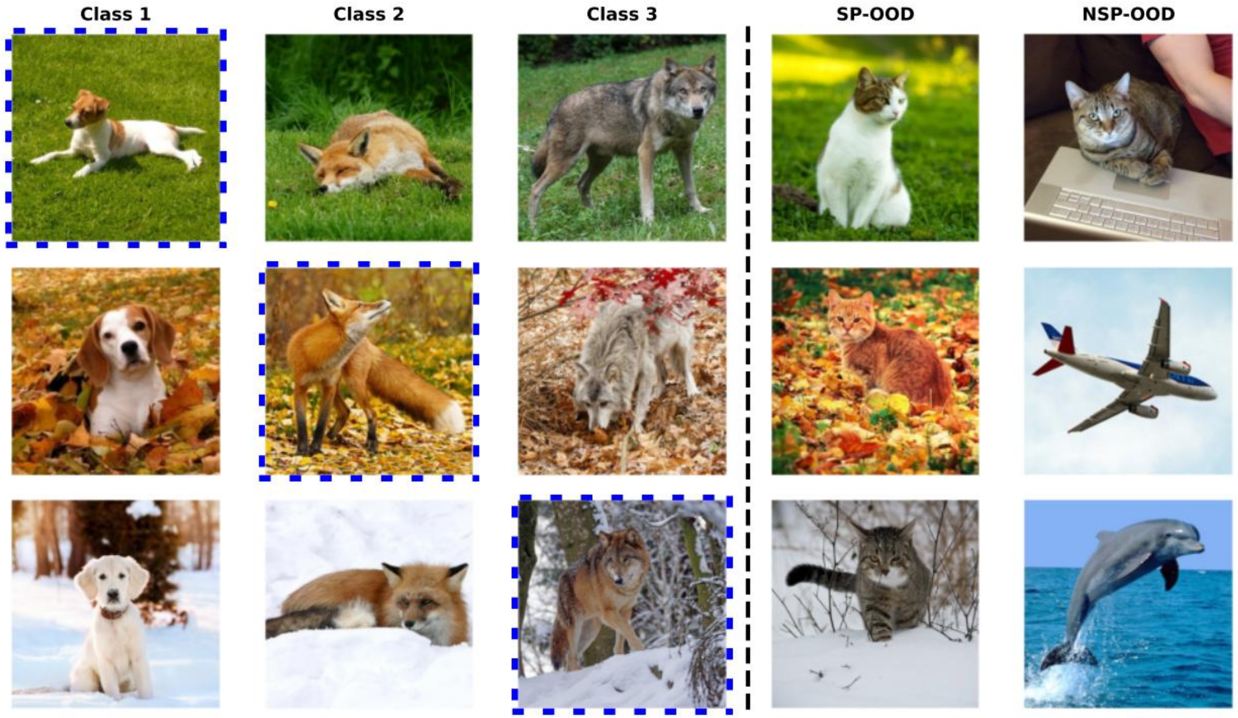
Reihaneh Zohrabi*, Hosein Hasani*, Mahdieh Soleymani, Anna Rohrbach, Marcus Rohrbach, Mohammad Hossein Rohban

TU Darmstadt, Sharif University of Technology

NeurIPS 2025

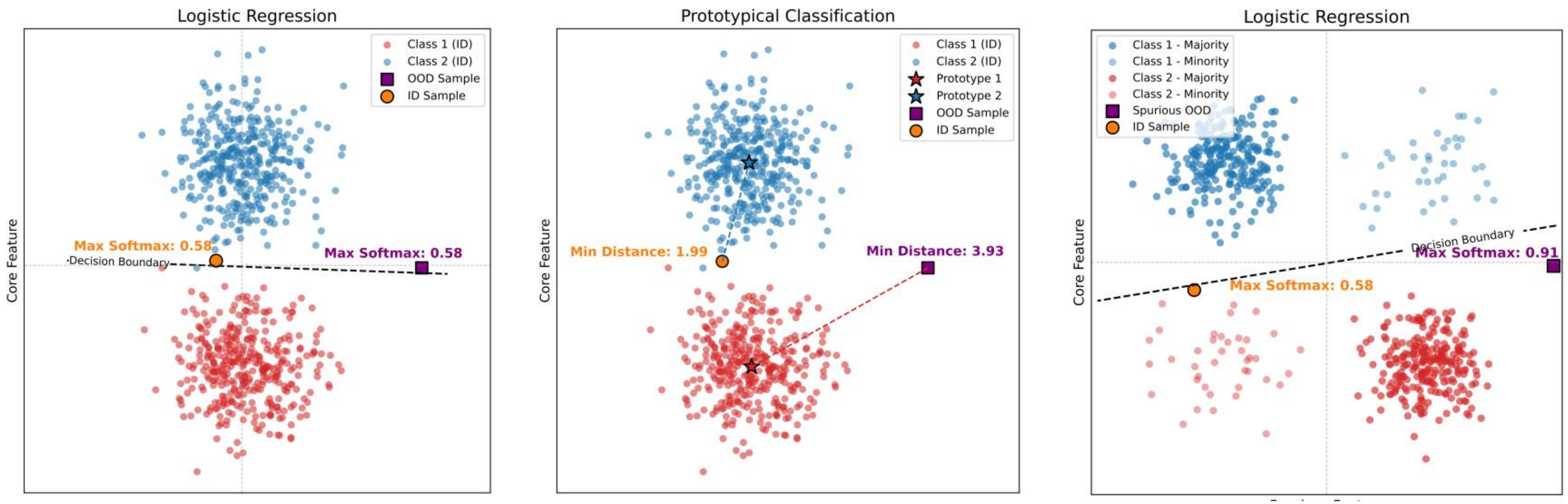
Problem Definition & Background

- **OOD Detection**: Identify inputs outside the training distribution.
- **Spurious Features**: Irrelevant cues learned by models.
- **SP-OOD**: OOD samples sharing spurious cues with in-distribution data.

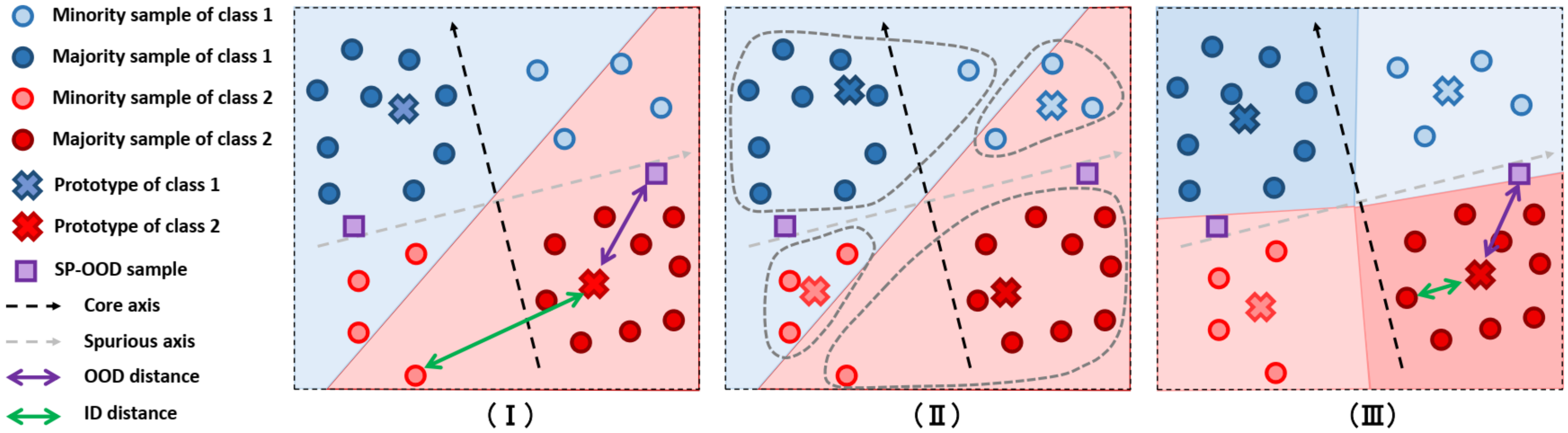


Motivation: Score Calculation

- **Score Calculation:**
  a scalar score $S(x)$ that separates ID and OOD samples.
- Two Perspectives:
  - **Discriminative**: defining score based on $p(y|x)$; e.g. softmax probabilities.
  - **Generative**: defining score based on $p(x|y)$; e.g. distance from prototypes.
- Discriminative approach has certain **failure cases**.



SPROD Overview



- **Stage 1 – Initial Prototypes**:
  Compute class prototypes from training features.
- **Stage 2 – Classification-Aware Prototypes**:
  Separate correctly and incorrectly classified samples to capture feature bias.
- **Stage 3 – Group Prototype Refinement**:
  Reassign samples and refine prototypes to reduce spurious influence.

Experimental Setup & Datasets

- **Baselines**:
  Compared with **19** post-hoc OOD detection methods.
- **Backbones**:
  Primarily ResNet-50; **10** different backbones in total.
- **Metrics**: AUROC and FPR@95, AUPR.
- **SP-OOD Benchmarks**:
  Waterbirds, CelebA, UrbanCars, Spurious ImageNet, and Animals MetaCoCo.

Results: SP-OOD benchmarks

- Comparative performance of **post-hoc OOD detection methods** on SP-OOD benchmarks using a **ResNet-50** backbone.

AUROC↑

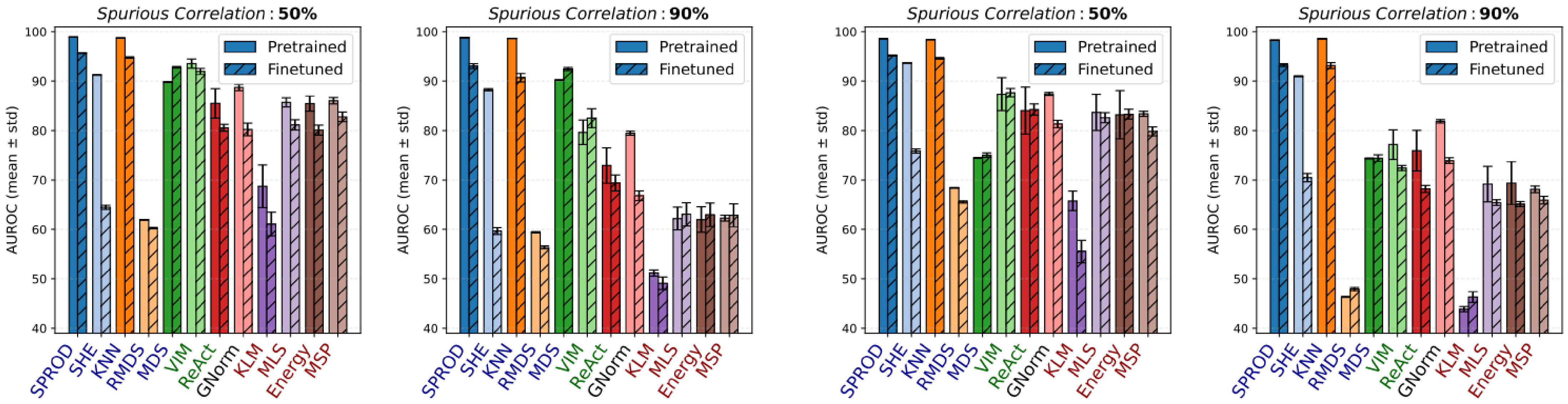
Method	WB	CA	UC	AMC	SpI	Avg.
MSP[7]	$62.3_{\pm0.6}$	$46.0_{\pm1.4}$	$38.5_{\pm0.3}$	$79.7_{\pm0.4}$	$83.1_{\pm0.3}$	61.9
Energy[13]	$62.0_{\pm2.6}$	$45.4_{\pm3.4}$	$38.4_{\pm2.1}$	$79.9_{\pm0.6}$	$80.6_{\pm0.4}$	61.3
MLS[44]	$62.2_{\pm2.3}$	$45.3_{\pm3.2}$	$38.4_{\pm1.4}$	$80.2_{\pm0.6}$	$81.9_{\pm0.3}$	61.6
KLM[44]	$51.2_{\pm0.7}$	$41.7_{\pm2.5}$	$57.0_{\pm0.2}$	$74.2_{\pm0.6}$	$79.6_{\pm0.8}$	60.7
GEN[45]	$62.3_{\pm0.6}$	$46.0_{\pm1.4}$	$38.5_{\pm0.3}$	$80.2_{\pm0.0}$	$80.8_{\pm0.4}$	61.6
GNorm[61]	$79.5_{\pm0.4}$	$38.0_{\pm1.3}$	$46.6_{\pm0.4}$	$74.2_{\pm0.5}$	$85.2_{\pm0.2}$	64.7
ReAct[58]	$72.9_{\pm3.6}$	$45.6_{\pm5.3}$	$41.3_{\pm3.1}$	$80.1_{\pm0.6}$	$83.6_{\pm0.7}$	64.7
VIM[59]	$79.6_{\pm2.5}$	$50.4_{\pm3.1}$	$60.7_{\pm1.7}$	$78.6_{\pm0.6}$	$77.4_{\pm0.9}$	69.3
ASH[60]	$78.5_{\pm3.2}$	$47.3_{\pm2.8}$	$39.6_{\pm1.7}$	$78.0_{\pm0.2}$	**$86.6_{\pm0.7}$**	66.0
MDS[11]	$90.2_{\pm0.1}$	$57.8_{\pm0.5}$	$91.8_{\pm0.1}$	$62.9_{\pm0.8}$	$58.4_{\pm0.1}$	72.2
RMDS[46]	$59.4_{\pm0.1}$	$33.6_{\pm1.4}$	$47.4_{\pm0.2}$	<u>81.9</u>$_{\pm0.4}$	$68.8_{\pm0.1}$	58.2
KNN[47]	<u>98.6</u>$_{\pm0.0}$	$54.5_{\pm0.5}$	$91.1_{\pm0.1}$	$79.7_{\pm0.6}$	$77.4_{\pm0.0}$	<u>80.3</u>
SHE[48]	$88.3_{\pm0.2}$	$42.7_{\pm0.6}$	$73.2_{\pm0.1}$	$54.8_{\pm0.7}$	$83.0_{\pm0.1}$	68.4
NECO[51]	$53.5_{\pm1.6}$	$39.5_{\pm3.2}$	$35.1_{\pm1.5}$	$80.2_{\pm0.1}$	$67.2_{\pm0.3}$	55.1
NNGuide[49]	$70.6_{\pm2.9}$	$49.8_{\pm4.2}$	$43.6_{\pm2.1}$	$79.4_{\pm0.0}$	$85.1_{\pm0.8}$	65.7
Relation[50]	$80.7_{\pm0.2}$	<u>60.4</u>$_{\pm2.5}$	<u>96.0</u>$_{\pm0.5}$	$74.5_{\pm0.3}$	$81.8_{\pm0.7}$	78.7
SCALE[52]	$89.0_{\pm2.9}$	$44.9_{\pm3.2}$	$54.4_{\pm2.1}$	$78.4_{\pm0.4}$	<u>86.2</u>$_{\pm0.5}$	70.6
fDBD[53]	$71.1_{\pm0.5}$	$51.3_{\pm1.3}$	$47.4_{\pm0.2}$	$79.9_{\pm0.0}$	$84.2_{\pm0.3}$	66.8
NCI[54]	$84.0_{\pm0.1}$	$46.4_{\pm2.4}$	$54.8_{\pm0.8}$	$78.5_{\pm0.1}$	$84.9_{\pm0.2}$	69.7
SPROD	**$98.8_{\pm0.0}$**	**$61.6_{\pm0.9}$**	**$97.4_{\pm0.0}$**	**$82.4_{\pm0.5}$**	$85.3_{\pm0.0}$	**85.1**

FPR@95↓

Method	WB	CA	UC	AMC	SpI	Avg.
MSP[7]	$87.9_{\pm0.8}$	$98.7_{\pm0.5}$	$97.3_{\pm0.3}$	$83.8_{\pm0.7}$	$74.1_{\pm1.2}$	88.4
Energy[13]	$89.2_{\pm3.2}$	$98.6_{\pm0.7}$	$95.5_{\pm3.1}$	$84.8_{\pm0.8}$	$76.3_{\pm0.9}$	88.9
MLS[44]	$88.1_{\pm2.0}$	$98.8_{\pm0.6}$	$96.7_{\pm2.0}$	$84.4_{\pm0.8}$	$74.6_{\pm0.9}$	88.5
KLM[44]	$89.1_{\pm0.7}$	$98.7_{\pm0.5}$	$97.1_{\pm0.3}$	$80.5_{\pm0.8}$	$76.1_{\pm1.7}$	88.3
GEN[45]	$87.9_{\pm0.8}$	$98.7_{\pm0.5}$	$97.3_{\pm0.3}$	$84.8_{\pm0.1}$	$76.3_{\pm0.7}$	89.0
GNorm[61]	$84.2_{\pm0.7}$	$98.8_{\pm0.4}$	$97.1_{\pm0.1}$	$84.2_{\pm0.6}$	$54.7_{\pm0.6}$	83.8
ReAct[58]	$86.9_{\pm7.0}$	$96.3_{\pm2.4}$	$95.5_{\pm3.2}$	$83.9_{\pm0.8}$	$57.5_{\pm1.6}$	84.0
VIM[59]	$61.4_{\pm3.5}$	$96.2_{\pm0.4}$	$69.0_{\pm1.5}$	$86.6_{\pm0.7}$	$79.5_{\pm0.5}$	78.5
ASH[60]	$85.2_{\pm7.0}$	$96.9_{\pm1.4}$	$96.1_{\pm1.5}$	$87.9_{\pm0.4}$	$52.9_{\pm3.1}$	83.8
MDS[11]	$49.2_{\pm0.2}$	$96.0_{\pm0.5}$	$39.0_{\pm0.3}$	$93.0_{\pm0.3}$	$90.5_{\pm0.1}$	73.5
RMDS[46]	$91.7_{\pm0.2}$	$99.6_{\pm0.1}$	$95.3_{\pm0.1}$	$83.4_{\pm0.9}$	$88.1_{\pm0.1}$	91.6
KNN[47]	<u>4.8</u>$_{\pm0.1}$	<u>94.4</u>$_{\pm1.0}$	$42.5_{\pm0.2}$	$79.9_{\pm1.1}$	$70.4_{\pm0.2}$	<u>58.4</u>
SHE[48]	$33.2_{\pm0.5}$	$96.4_{\pm0.5}$	$76.5_{\pm0.2}$	$93.9_{\pm0.3}$	<u>52.6</u>$_{\pm0.8}$	70.5
NECO[51]	$90.5_{\pm2.2}$	$98.8_{\pm0.6}$	$96.7_{\pm1.9}$	<u>78.2</u>$_{\pm0.0}$	$89.9_{\pm0.8}$	90.8
NNGuide[49]	$77.7_{\pm6.0}$	$97.6_{\pm1.2}$	$91.6_{\pm3.6}$	$86.3_{\pm0.1}$	**$52.2_{\pm1.4}$**	81.1
Relation[50]	$73.8_{\pm0.5}$	$95.4_{\pm0.2}$	<u>24.2</u>$_{\pm1.7}$	$84.6_{\pm0.2}$	$78.0_{\pm1.1}$	71.2
SCALE[52]	$61.0_{\pm22.5}$	$98.7_{\pm0.6}$	$94.6_{\pm3.1}$	$87.7_{\pm0.3}$	$53.0_{\pm1.5}$	79.0
fDBD[53]	$85.5_{\pm0.8}$	$98.6_{\pm0.5}$	$96.1_{\pm0.4}$	$85.4_{\pm0.2}$	$70.8_{\pm1.0}$	87.3
NCI[54]	$41.1_{\pm0.1}$	$99.4_{\pm0.3}$	$92.2_{\pm0.6}$	$85.8_{\pm0.3}$	$63.8_{\pm0.9}$	76.5
SPROD	**$4.7_{\pm0.1}$**	**$93.7_{\pm0.9}$**	**$19.0_{\pm0.4}$**	**$69.5_{\pm1.2}$**	$58.0_{\pm0.1}$	**49.0**

Results: Fine-Tuning and SP Correlation Rate

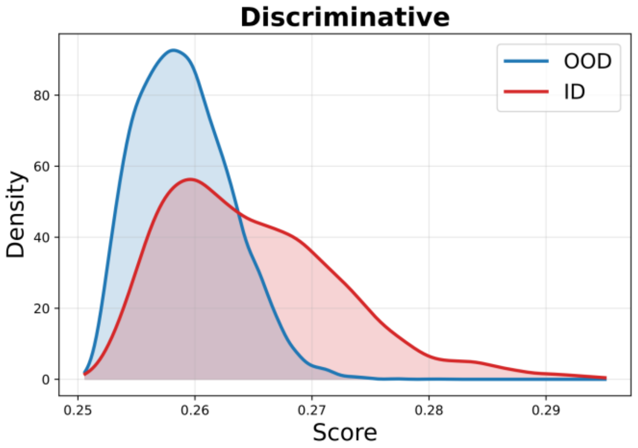
- **Fine-tuning** the backbone **reduces OOD performance**, especially for feature-based methods.
- **Higher spurious correlation (90% vs. 50%)** further degrades performance, mainly for output-based methods.
- **Smaller backbone (ResNet-18)** performs comparably to ResNet-50.

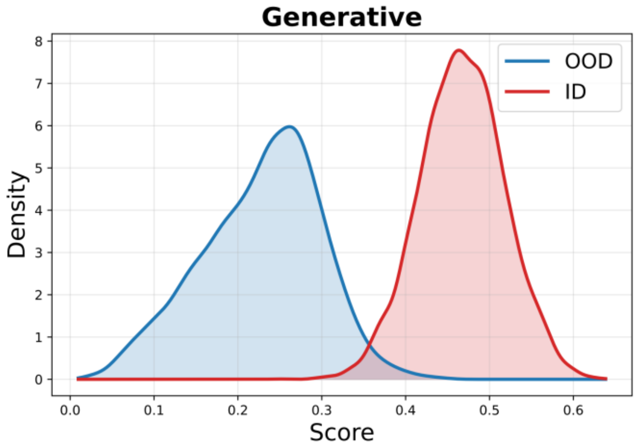


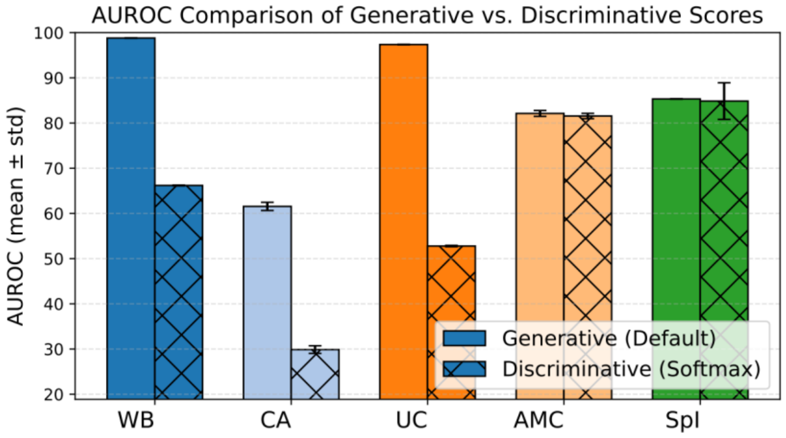
(a) ResNet-50 backbone. (b) ResNet-18 backbone.

Results: Generative & Discriminative (Ablation)

- **Distance-based** vs. **softmax-based** scoring, using identical SPROD features and prototypes.
- **Generative** scores yield **clearer ID/OOD separation.**
- **Generative** scoring **outperforms** on most SP-OOD datasets. Biggest gap on Waterbirds, CelebA, and UrbanCars.







Results: Low-Shot and Zero-Shot SP-OOD

- **Zero-Shot Comparison (CLIP-based):**
  - **MCM:** 98.36 **CMA:** 98.62 **SPROD: 99.01 (vision-only)**
- Text-free SPROD **outperforms CLIP-based zero-shot methods**.
- **Low-Shot Comparison:**
  - **SPROD maintains high AUROC** even with limited samples, unlike KNN.