Sparsity as a Key: Unlocking New Insights from Latent Structures for Out-of-Distribution Detection

Supplementary Material

A. Activation Affinity of OOD Samples to Specific ID Core Features

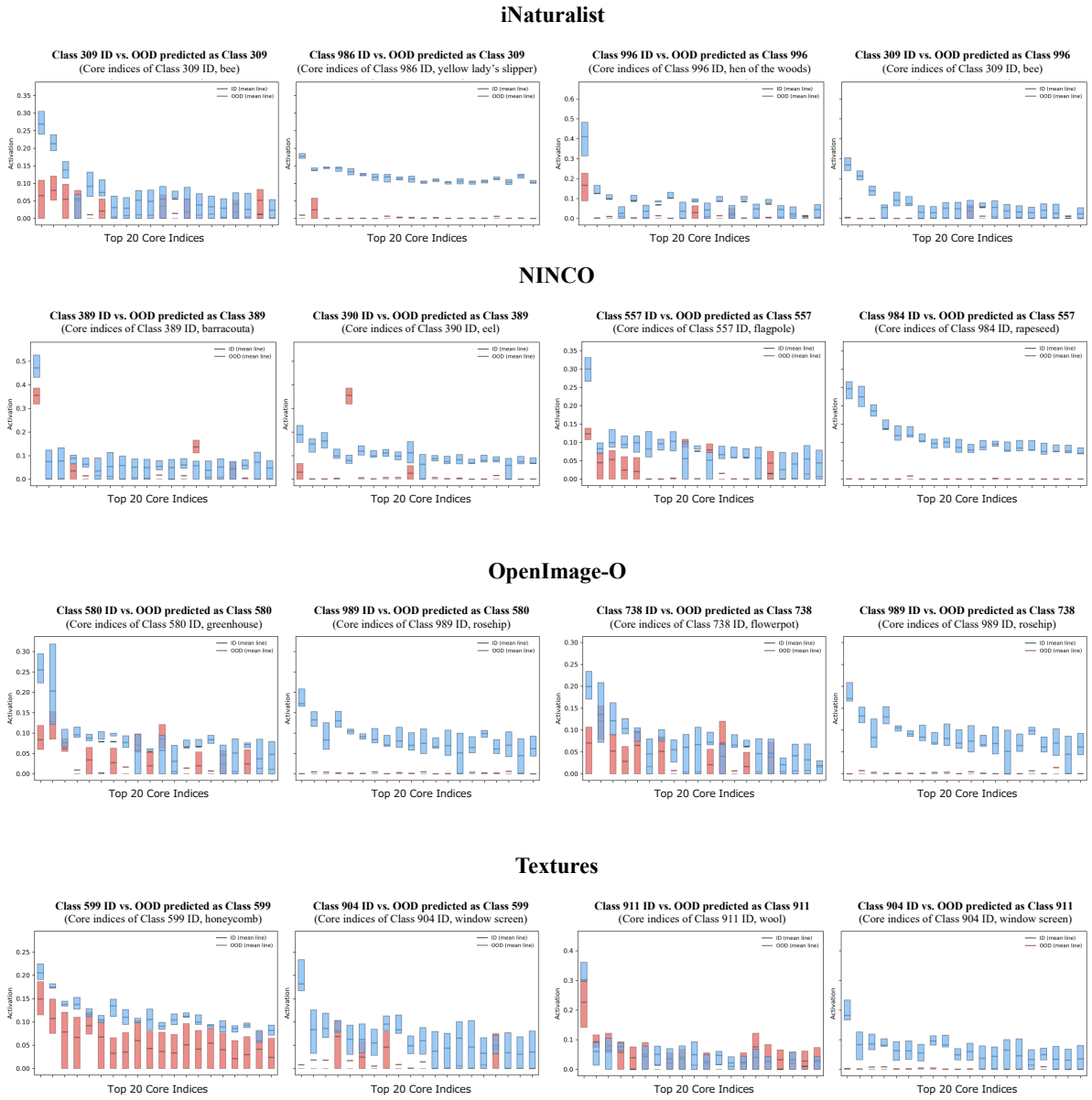


Figure 7. **Core feature activation by different OOD datasets.** This figure presents two pairs of graphs for each dataset. Each pair shows an OOD sample (Red) compared against ID samples (Blue) using two class feature sets. **Left panel (in each):** The OOD sample consistently exhibits high activation on the core features of its predicted ID class. **Right panel (in each):** The same OOD sample demonstrates negligible activation on the core features of an unrelated ID class. This pattern confirms the robust semantic alignment driving classifier predictions, irrespective of the OOD source dataset.

To demonstrate that the semantic alignment of OOD samples is a general phenomenon—a key finding discussed in Section 3.2—Figure 7 presents additional evidence drawn from diverse OOD datasets, including iNaturalist, NINCO, OpenImage-O, and Textures. As evident in the figure, the pattern remains remarkably consistent. In every illustrated case, the OOD sample (Red) exhibits significant, non-zero activation levels when evaluated against the core features of its misclassified ID class (Left panels in each). Conversely, the same OOD sample consistently demonstrates negligible activation when processed against the core features belonging to an entirely unrelated ID class (Right panels in each). This observation strongly reinforces our hypothesis that the classifier's decisions are driven by structural and semantic alignments in the sparse latent space, rather than resulting from random classification failures.

B. Global Statistics of Activation Intensity

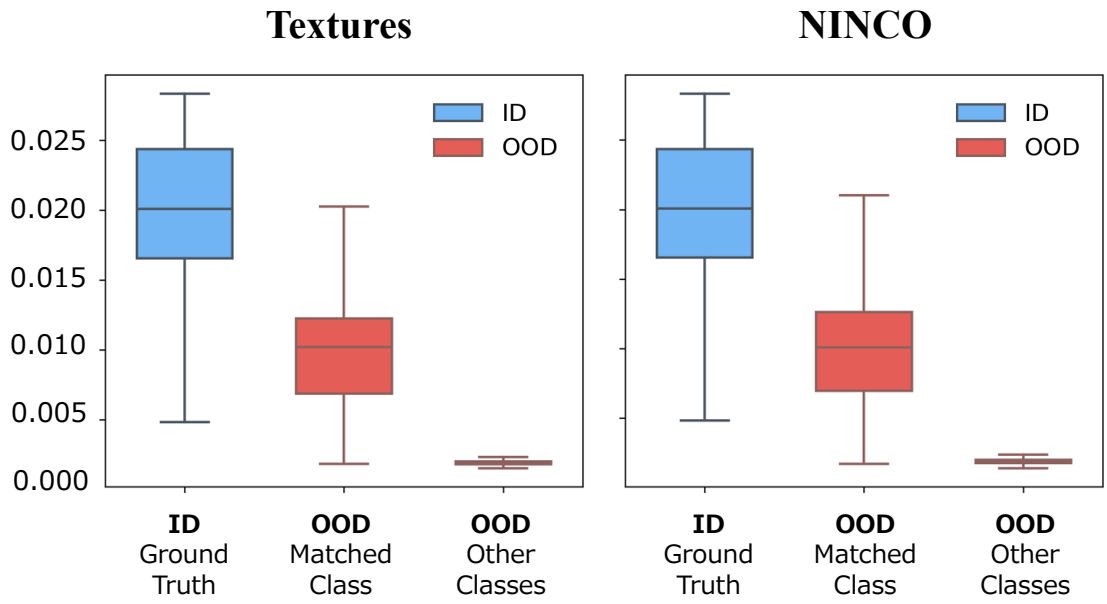


Figure 8. **Global statistics of core feature activation intensity.** We aggregate mean activation values on core indices across all classes for OOD datasets. **ID Ground Truth (Blue, Left):** Demonstrates strong activation on the core features corresponding to the ground-truth class. **OOD Matched Class (Red, Middle):** OOD samples activate the core features of their *predicted* ID class, though with consistently lower intensity than true ID samples. **OOD Other Classes (Red, Right):** OOD samples exhibit minimal activation on the core features of unrelated ID classes.

As a supplement to the analysis Figure 5 in Section 3.2, Figure 8 demonstrates that our findings are not limited to specific examples but represent a general phenomenon across diverse OOD datasets. The aggregated statistics consistently show that OOD samples systematically activate their *matched* class features (Red, Middle) but not *other* class features (Red, Right). This clear separation provides another strong evidence supporting our hypothesis that this specific activation is the reason why the sample is classified into that particular ID class. Furthermore, this OOD activation magnitude (Red, Middle) is consistently weaker than that of true ID samples (Blue, Left).

C. Structural Differences in ID and OOD Activation

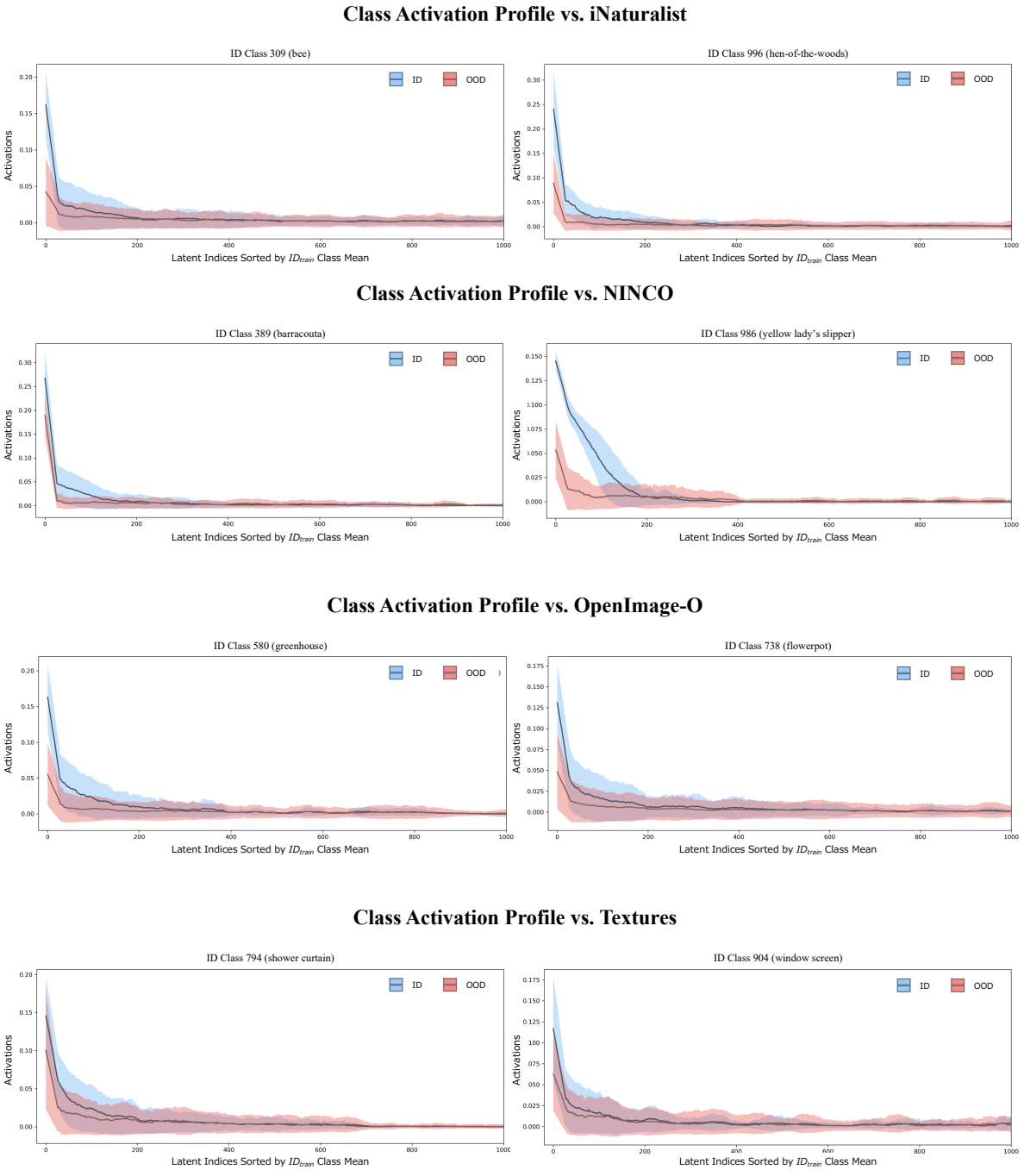


Figure 9. **Consistency of structural activation profiles across diverse OOD datasets.** This figure provides additional examples comparing the mean activation on CAP of ID samples (Blue) and OOD samples (Red) misclassified as the same class across four benchmark datasets: iNaturalist, NINCO, OpenImage-O, and Textures. The latent indices are sorted according to the mean ID activation profile of the respective class. In all cases, the distinctive activation head of ID samples is sharply attenuated and dispersed in OOD samples, confirming the generality of the structural disruption observed in the main paper.

To further substantiate the claims made in Section 3.3 regarding the fundamental structural disruption between ID and misclassified OOD samples, Figure 9 presents comprehensive activation profile comparisons. This supplementary evidence confirms that the phenomena illustrated in Figure 6 are systematic findings, not exceptions limited to specific datasets or classes. As shown across all tested OOD benchmarks (iNaturalist, NINCO, OpenImage-O, and Textures), the consistent trend is the inability of OOD samples (Red) to replicate the activation signature of true ID data (Blue). Specifically, OOD activation profiles display a diffused and lower-intensity distribution, particularly failing to match the high-energy, concentrated peak—the "head" of the profile—that is characteristic of the ID samples. This consistent structural failure across varied OOD sources underscores the robustness of our central observation, and provides the essential foundation for our EPD metric.

D. Hyperparameter Ablation: Latent Dimension and Sparsity Level

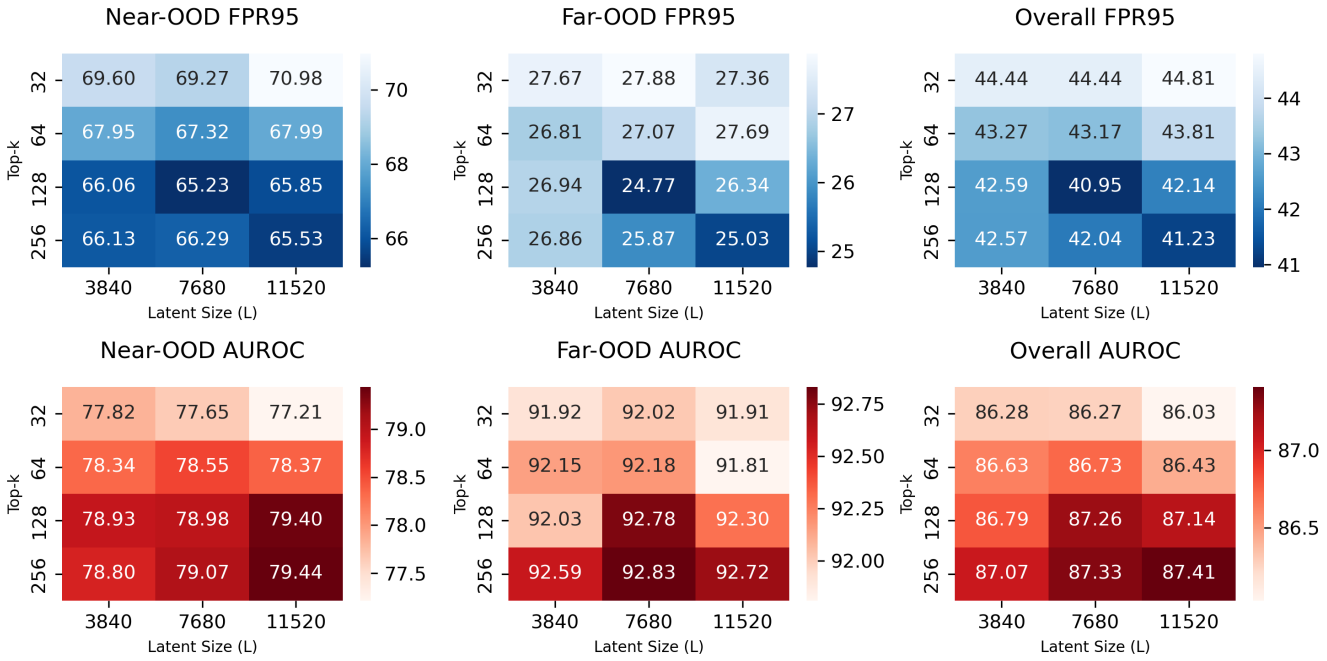


Figure 10. **Hyperparameter sensitivity analysis.** FPR95 (blue, lower is better) and AUROC (red, higher is better) across combinations of latent dimension (L) and sparsity level (k). Darker colors indicate better performance. The selected configuration ($L = 7680$, $k = 128$) minimizes overall FPR95 while maintaining strong AUROC, representing the optimal trade-off between robustness and separability.

Figure 10 presents a comprehensive sensitivity analysis over our two primary architectural hyperparameters: the latent dimension $L \in \{3840, 7680, 11520\}$ and the sparsity level $k \in \{32, 64, 128, 256\}$. The results reveal a consistent pattern: increasing L generally requires a proportional increase in k to maintain reconstruction quality, but this scaling is eventually constrained by model instability at larger k values. Our selected configuration ($L = 7680$, $k = 128$) yields the best overall FPR95 (40.96%), prioritizing robustness in safety-critical detection over marginal AUROC gains. The relative flatness of AUROC across configurations further confirms that our method is not overly sensitive to the precise hyperparameter choice within a reasonable range.

E. Activation Head Size Sensitivity

To support the parameter choices detailed in Section 4.3, Figure 11 provides a comprehensive sensitivity analysis on the Activation Head Ratio (p). This hyperparameter defines the percentage of top-activated latent indices utilized for calculating our EPD score.

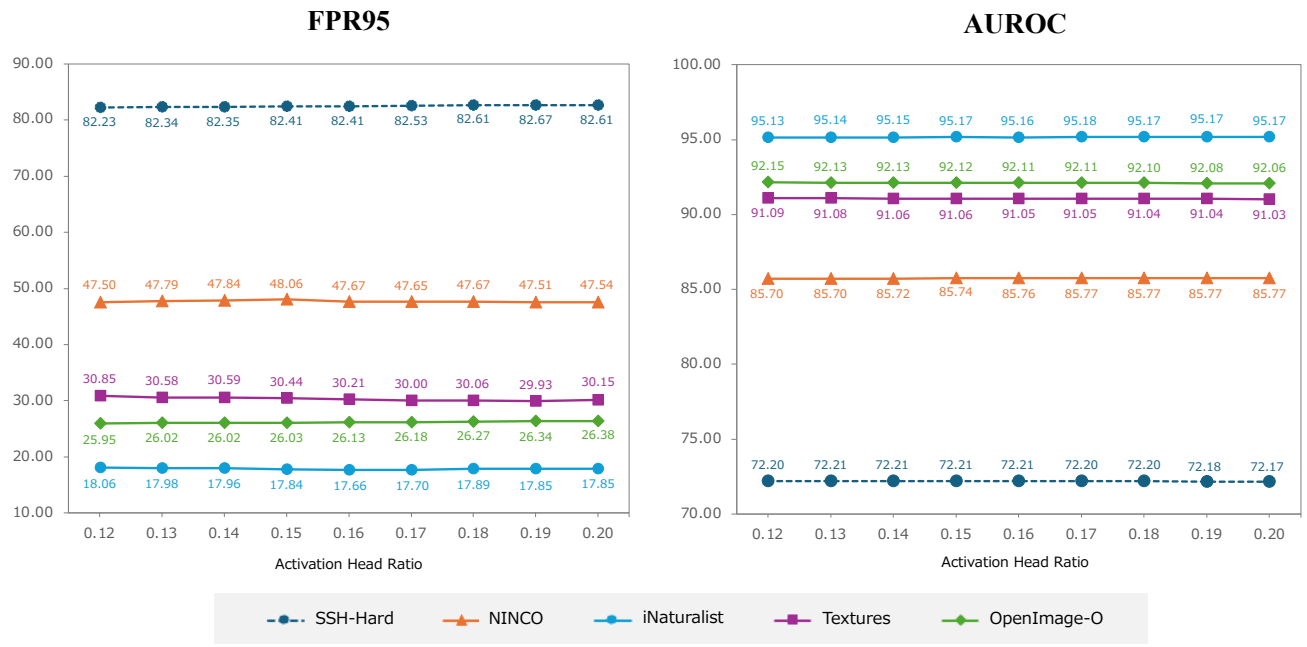


Figure 11. **Sensitivity analysis of the activation head ratio (p).** Performance metrics, FPR95 (Left) and AUROC (Right), are evaluated across various OOD benchmarks as a function of the activation head ratio (p). The ratio determines the size of the sorted latent feature set used for divergence calculation. The results demonstrate the stability of our proposed method EPD across the empirically derived meaningful range ($p = 0.12$ to $p = 0.20$), validating the robust selection of the chosen value ($p = 0.15$) used in the main paper.

The analysis systematically sweeps p across the empirically derived range. The resulting metrics, FPR95 and AUROC, show remarkable consistency across all tested benchmark OOD datasets (SSH-Hard, NINCO, iNaturalist, Textures, and OpenImage-O). The observed flatness of the performance curves explicitly confirms that the effectiveness of our EPD method is highly robust and not overly dependent on the precise selection of p. This validates our choice of p=0.15 as an optimal, stable setting that effectively balances detection sensitivity and generalizability across diverse OOD sources.

F. Ablation: Scoring Metric

Metric	Near-OOD		Far-OOD		Overall	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Euclidean	67.56	76.83	29.68	90.87	44.83	85.25
Cosine	71.74	77.09	47.93	84.21	57.45	81.36
EPD (Ours)	**65.23**	**78.98**	**24.77**	**92.78**	**40.96**	**87.26**

Table 4. **Metric Ablation on ViT-B/16.** EPD outperforms Euclidean and cosine distance across all evaluation splits, validating the use of KL divergence over normalized energy profiles.

EPD's advantage stems from its use of KL divergence over L_1-normalized profiles, which captures the *shape* of the energy distribution rather than its magnitude or direction alone. Euclidean distance performs competitively on Far-OOD but degrades on Near-OOD, while cosine similarity is notably weaker on Far-OOD—confirming that structural shape within the sparse simplex is the most discriminative signal for OOD detection.

G. Full Experimental Results

G.1. Full Results on ViT

Method	Near-OOD						Far-OOD								Overall	
	SSB-Hard		NINCO		Average		iNaturalist		Textures		OpenImage-O		Average		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
ASH	93.50	53.89	95.40	52.52	94.45	53.20	97.02	50.63	98.49	48.53	94.80	55.52	96.77	51.56	95.84	52.22
DICE	89.77	59.05	81.09	71.67	85.43	65.36	47.92	82.51	54.79	82.21	52.57	82.23	51.76	82.32	65.23	75.53
EBO	92.25	58.80	94.16	66.02	93.21	62.41	83.58	79.30	83.65	81.17	88.79	76.48	85.34	78.98	88.49	72.35
GEN	**82.24**	70.09	59.31	82.51	70.78	76.30	22.94	93.54	38.31	90.23	35.43	90.27	32.23	91.35	47.65	85.33
GradNorm	93.62	42.96	95.81	64.40	94.72	53.68	91.16	42.42	92.25	44.99	94.52	37.82	92.64	41.74	93.47	46.52
KNN	86.22	65.97	54.73	82.25	70.48	74.11	27.74	91.46	33.23	91.12	34.82	89.86	31.93	90.81	47.35	84.13
MDS	83.47	71.57	48.76	86.52	66.11	79.04	20.66	96.01	38.90	89.41	30.35	**92.38**	29.97	92.60	44.43	87.18
MLS	91.52	64.20	92.98	72.40	92.25	68.30	72.98	85.29	78.93	83.74	85.78	81.60	79.23	83.54	84.44	77.45
MSP	86.41	68.94	77.35	78.11	81.88	73.53	42.42	88.19	56.44	85.06	56.11	84.87	51.66	86.04	63.75	81.03
OpenMax	89.20	68.60	88.36	78.68	88.78	73.64	19.56	94.93	73.17	85.52	73.74	87.36	55.49	89.27	68.81	83.02
ReAct	90.46	63.10	78.50	75.43	84.48	69.27	48.22	86.11	55.87	86.66	57.68	84.29	53.92	85.69	66.15	79.12
RMDS	84.53	**72.87**	**46.22**	**87.31**	65.38	**80.09**	19.46	**96.10**	37.23	89.38	29.57	92.32	28.75	92.60	43.40	**87.60**
SHE	85.74	68.04	56.01	84.18	70.88	76.11	22.17	93.57	**25.65**	**92.65**	33.59	91.04	27.14	92.42	44.63	85.89
TempScale	87.36	68.55	81.90	77.80	84.63	73.18	43.08	88.54	58.22	85.39	60.00	85.04	53.77	86.32	66.11	81.06
ViM	90.06	69.42	57.45	84.64	73.76	77.03	**17.59**	95.72	40.41	90.61	29.59	92.18	29.20	**92.84**	47.02	86.51
Ours	82.41	72.21	48.06	85.74	**65.23**	78.98	17.84	95.17	30.44	91.06	**26.03**	92.12	**24.77**	92.78	**40.96**	87.26

Table 5. OOD detection performance on ImageNet-1K benchmarks, evaluated on a ViT-B/16 (ID Acc: 81.14%). For each metric, the best result is in **bold** and the second and third results are underlined.

G.2. Full Results on Swin-T

Method	Near-OOD						Far-OOD								Overall	
	SSB-Hard		NINCO		Average		iNaturalist		Textures		OpenImage-O		Average		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
ASH	95.22	46.28	93.98	47.00	94.60	46.64	94.19	46.49	96.17	41.32	93.99	45.22	94.78	44.34	94.71	45.26
DICE	95.76	49.97	97.28	50.00	96.52	49.99	98.11	47.09	87.37	77.61	96.36	58.67	93.95	61.12	94.98	56.67
EBO	87.51	68.28	79.23	78.42	83.37	73.35	61.18	85.17	84.29	79.00	80.20	80.24	75.22	81.47	78.48	78.22
GEN	**79.35**	**72.78**	**48.10**	**85.16**	**63.73**	**78.97**	20.47	94.23	45.04	88.16	32.77	90.60	32.76	91.00	45.15	86.19
GradNorm	93.37	50.43	94.06	45.01	93.72	47.72	95.18	38.28	97.80	34.75	96.83	34.00	96.60	35.68	95.45	40.49
KNN	85.08	64.21	58.14	79.16	71.61	71.69	31.15	88.91	35.79	90.56	36.30	88.62	34.41	89.36	49.29	82.29
MDS	83.72	68.69	53.55	81.78	68.64	75.24	21.89	93.60	37.47	89.82	30.89	90.98	30.08	91.47	45.50	84.97
MLS	86.60	70.47	75.78	80.95	81.19	75.71	49.65	89.01	79.94	81.70	73.96	83.93	67.85	84.88	73.19	81.21
MSP	81.02	71.75	60.36	81.69	70.69	76.72	37.50	89.84	61.49	83.27	48.97	85.81	49.32	86.31	57.87	82.47
OpenMax	85.54	71.52	72.85	81.69	79.20	76.61	19.56	95.06	76.41	82.81	60.88	87.75	52.28	88.54	63.05	83.77
ReAct	85.01	69.36	60.08	82.12	72.55	75.74	31.34	90.08	53.98	87.04	42.54	87.85	42.62	88.32	54.59	83.29
RMDS	82.66	71.81	49.89	84.91	66.28	78.36	18.55	**95.57**	39.09	89.26	29.34	92.10	29.00	92.31	43.91	**86.73**
SHE	85.03	70.75	67.53	82.74	76.28	76.75	33.25	92.78	51.32	88.86	52.88	88.04	45.82	89.89	58.00	84.63
TempScale	82.12	71.83	63.02	82.09	72.57	76.96	36.84	90.36	62.89	83.68	50.30	86.20	50.01	86.75	59.03	82.83
ViM	88.55	68.94	60.80	81.85	74.68	75.40	**17.98**	94.62	**29.46**	**92.69**	**26.62**	**92.29**	**24.69**	**93.20**	44.68	86.08
Ours	81.62	70.50	55.40	83.01	68.51	76.76	20.37	94.07	32.64	90.86	27.90	91.71	26.97	92.21	**43.59**	86.03

Table 6. OOD detection performance on ImageNet-1K benchmarks, evaluated on a Swin Transformer (ID Acc: 81.60%. For each metric, the best result is in **bold** and the second and third results are underlined.

G.3. Full Results on Multiple Architectures (Ours Only)

Method	Near-OOD						Far-OOD								Overall	
	SSB-Hard		NINCO		Average		iNaturalist		Textures		OpenImage-O		Average		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Ours (ViT-B)	82.41	72.21	48.06	85.74	65.23	78.98	17.84	95.17	30.44	91.06	26.03	92.12	24.77	92.78	40.96	87.26
Ours (Swin-T)	81.62	70.50	55.40	83.01	68.51	76.76	20.37	94.07	32.64	90.86	27.90	91.71	26.97	92.21	43.59	86.03
Ours (DINOv2)	80.05	76.82	44.74	89.63	62.40	83.23	7.21	97.87	32.35	92.11	16.22	95.51	18.59	95.16	40.50	89.20

Table 7. **OOD Detection performance of our method across different architectures.** Detailed OOD detection performance on ImageNet-1K benchmarks for our proposed method (EPD) applied to DINOv2 B/14 (ID Acc: 84.64%), ViT-B/16 (ID Acc: 81.14%), and Swin-T (ID Acc: 81.60%).

The comprehensive OOD detection performance results for the ViT-B/16 (Table 5) and Swin-T (Table 6) backbones, including detailed metrics for all OpenOOD benchmark methods, are presented here. Moreover, to definitively demonstrate the generalizability and robust effectiveness of our core structural detection mechanism, Table 7 provides a side-by-side comparison of our method applied to three distinct Vision Transformer backbones: DINOv2, ViT-B/16, and Swin-T. These tables provide the detailed data that underpins the summary findings discussed in Section 4.5 and 4.6.

- **ViT-B/16**: The detailed results emphatically confirm that our method, EPD using CAP, achieves the best overall average FPR95 ($\mathbf{40.96}\%$) among all compared methods, outperforming the next best methods (RMDS at 43.40% and MDS at 44.43%). This performance gap validates EPD's exceptional capability in minimizing false positives across diverse OOD scenarios, which is paramount for safety-critical applications.
- **Swin-T**: Our method achieves the best overall average FPR95 ($\mathbf{43.59}\%$) among the compared methods on this architecture. The overall AUROC reflects the influence of the Swin-T architecture, where localized attention and hierarchical feature extraction result in latent representations with reduced global coherence. Despite this, our method secures the best overall average FPR95 and a competitive AUROC score.
- **DINOv2**: Our method achieves its highest overall performance on DINOv2, leading to the best overall FPR95 ($\mathbf{40.50}\%$) among all tested backbones. This inclusion of DINOv2 further validates the general applicability of EPD, proving that our method's efficacy of OOD detection stems from fundamental structural properties of the latent space. It is important to note that due to the lack of comprehensive OOD benchmark results for DINOv2 in existing literature, Table 7 presents the performance of only our method as an internal comparison.

This comparative analysis solidifies our hypothesis that the efficacy of OOD detection through structural deviation is not dependent on a specific backbone model.

H. Results on Different ID Dataset: CIFAR-100

Method	Near-OOD						Far-OOD								Overall	
	CIFAR10		TinyImageNet		Average		SVHN		Textures		Places365		Average		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
ASH	87.04	56.96	90.97	51.23	89.01	54.09	81.59	57.61	82.07	58.53	91.72	52.57	85.13	56.24	86.68	55.38
DICE	69.63	81.68	60.53	83.57	65.08	82.63	44.10	90.96	73.32	82.57	65.11	82.65	60.84	85.39	62.54	84.29
EBO	63.84	84.38	49.64	86.83	56.74	85.60	39.33	89.89	59.26	86.88	64.07	83.81	54.22	86.86	55.23	86.36
GEN	62.46	84.64	49.98	86.80	56.22	85.72	40.06	89.63	58.78	86.89	64.43	83.80	54.42	86.77	55.14	86.35
GradNorm	48.78	85.52	45.17	87.63	46.97	86.57	49.54	**91.47**	58.10	86.65	**44.07**	**88.41**	50.57	**88.84**	49.13	87.94
KNN	**43.61**	87.67	41.53	87.77	42.57	87.72	41.08	88.95	47.88	87.96	50.66	85.39	46.54	87.43	44.95	87.55
MDS	44.64	88.10	44.81	85.60	44.73	86.85	50.50	78.14	38.52	90.14	56.47	82.55	48.50	83.61	46.99	84.91
MLS	63.70	84.17	49.52	86.62	56.61	85.39	39.57	89.46	59.27	86.51	64.07	83.58	54.30	86.52	55.23	86.07
MSP	61.80	81.68	52.63	83.84	57.22	82.76	46.98	84.81	64.86	82.25	64.56	80.80	58.80	82.62	58.17	82.68
OpenMax	62.94	80.95	54.92	83.58	58.93	82.26	48.94	84.42	75.71	79.82	66.20	81.15	63.62	81.80	61.74	81.98
ReAct	51.21	86.02	41.68	88.11	46.44	87.06	41.10	91.04	57.11	87.31	47.42	87.14	48.54	88.50	47.70	87.92
RMDS	47.02	86.50	41.81	86.82	44.42	86.66	37.60	87.27	43.03	86.84	49.79	84.64	43.47	86.25	43.85	86.41
SHE	49.91	86.03	46.61	87.35	48.26	86.69	78.93	88.52	65.28	86.85	46.96	87.40	63.72	87.59	57.54	87.23
TempScale	62.41	83.02	50.92	85.18	56.67	84.10	44.23	86.85	61.70	84.23	64.19	82.13	56.71	84.40	56.69	84.28
ViM	43.89	**88.58**	**38.73**	**88.43**	**41.31**	**88.50**	**36.11**	86.88	**35.86**	**91.91**	46.51	86.53	**39.49**	88.44	**40.22**	**88.47**
Ours	44.76	87.25	42.76	87.48	43.76	87.37	41.76	87.92	45.21	87.93	52.44	85.21	46.47	87.02	45.39	87.16

Table 8. OOD detection performance on CIFAR-100 benchmarks, evaluated on a ViT-B/16 (ID Acc: 84.78%). For this experiment, we employed a fine-tuned version of `google/vit-base-patch16-224-in21k`. For each metric, the best result is in **bold** and the second and third results are underlined.

To validate the general applicability and domain robustness of our EPD method, we conducted extensive experiments using CIFAR-100 as the ID dataset, a domain fundamentally different in scale and complexity from the ImageNet domain studied in the main experiment. The full set of results comparing EPD with established OpenOOD benchmarks is detailed in Table 8. The OOD datasets utilized for this analysis were selected from the suggested list for CIFAR-100 in the OpenOOD v 1.5 framework. The results demonstrate that EPD maintains its effectiveness despite the significant shift in the training domain. While some methods, such as KNN and GEN, showing a notable surge in performance on the CIFAR-100 domain compared to their ImageNet-1k score, our method demonstrates cross-domain stability. EPD maintains a robust performance where the score gap relative to the top-ranking method is small. Our method shows highly stable performance across both Near-OOD and Far-OOD categories, proving that the core mechanism is not architecture- or domain-specific.

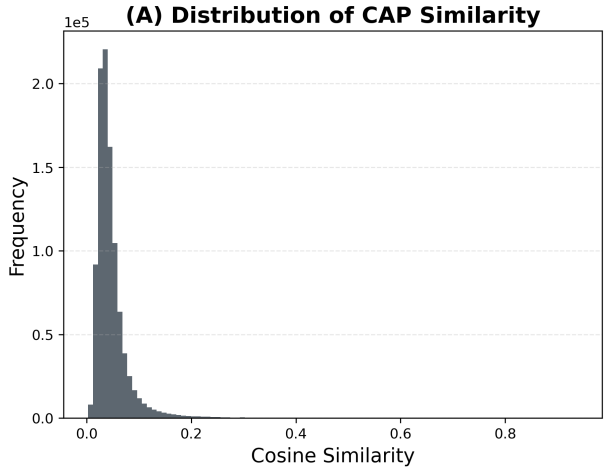
I. Comparison with Recent Baselines: MDS++ and RMDS++

Table 9 extends our comparison to include recently proposed methods MDS++ and RMDS++ [28], which are enhanced variants of MDS and RMDS. Note that unlike the other baselines evaluated via OpenOOD v1.5, these results were obtained through our own implementation. Our method remains highly competitive against these stronger baselines. In particular, our approach records the best FPR95 on SSB-Hard and Textures among all three methods, demonstrating that our structural detection mechanism provides complementary advantages. While MDS++ and RMDS++ outperform ours on iNaturalist and NINCO, our method shows superior or competitive performance on the remaining datasets.

Method	Near-OOD				Far-OOD					
	SSB-Hard		NINCO		iNaturalist		Textures		OpenImage-O	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
MDS++	83.97	74.04	44.90	88.55	**11.97**	**97.18**	31.37	90.57	**25.21**	**93.11**
RMDS++	83.29	**74.81**	**43.86**	**88.67**	13.34	96.95	34.74	89.28	26.45	92.66
Ours	**82.41**	72.21	48.06	85.74	17.84	95.17	**30.44**	**91.06**	26.03	92.12

Table 9. Comparison with recent methods on ImageNet-1K benchmarks using ViT-B/16. **Bold**: best result; underlined: second best.

J. CAP Cosine Similarity Analysis



(B) Examples of Semantic vs. Distinct Pairs

Type	Class A	Class B	Similarity
High (Semantic)	Eskimo_dog	Siberian_husky	0.939
High (Semantic)	missile	projectile	0.938
High (Semantic)	breastplate	cuirass	0.931
High (Semantic)	bathtub	tub	0.914
High (Semantic)	cassette_player	tape_player	0.912
Low (Distinct)	Shih-Tzu	passenger_car	0.032
Low (Distinct)	sulphur_butterfly	whistle	0.029
Low (Distinct)	beacon	moped	0.048
Low (Distinct)	Christmas_stocking	stage	0.035
Low (Distinct)	rhinoceros_beetle	pencil_box	0.025

Figure 12. **CAP cosine similarity analysis. (A)** Distribution of pairwise cosine similarity between all 1,000 ID class CAPs. The distribution peaks near zero, confirming that most ID classes form mutually orthogonal subspaces in the sparse latent space. **(B)** Examples of class pairs with high and low CAP similarity. High-similarity pairs (red) correspond to semantically related concepts (*e.g.*, Eskimo dog vs. Siberian husky), while low-similarity pairs (blue) represent semantically distinct classes. This validates that CAPs capture meaningful semantic identity rather than arbitrary activation patterns.

The pairwise cosine similarity distribution in Figure 12(A) provides further evidence for the disjointness of CAPs established in Section 3.1 of the main paper. While the Jaccard similarity in Figure 2 measures binary feature overlap, cosine similarity here captures the directional alignment of the full mean activation vectors. The zero-skewed distribution confirms that the vast majority of ID class pairs occupy nearly orthogonal subspaces in the sparse latent space. The qualitative examples in (B) further validate that the few high-similarity pairs correspond exclusively to semantically related classes (*e.g.*, near-synonym animal breeds or synonymous objects), demonstrating that the sparse latent space organizes classes according to genuine semantic relationships rather than arbitrary activation patterns.