

Supplementary Materials for RAVEN: Representation-Aware Reliability Validation under Data Distribution Shifts

Nguyen Hai Tien Phat

Purpose of the supplementary tables

The supplementary tables contain secondary, diagnostic, and reproducibility-oriented material that supports the main manuscript without interrupting its primary evidence chain. They are not additional claims of superiority. The main text retains only tables needed to understand the central research question, primary statistical evidence, and principal generalization checks.

Supplementary Tables

Table 1. Supplementary Table S1. Positioning of RAVEN relative to representative reliability and shift-detection approaches. The purpose is to distinguish the evidence target of each family rather than to rank methods.

Family	Typical signal	Role in RAVEN study
MSP / entropy	Output probability distribution	Direct confidence baselines [1, 2]
Energy	Output logits	Strong output-space baseline [4]
Mahalanobis	Feature-space class distance	Strong representation-space baseline [3]
kNN	Local representation geometry	Non-parametric representation baseline
Selective prediction	Score-to-reject ordering	AURC and risk-coverage evaluation [5]
RAVEN	Confidence + semantic consistency + MRI + MRD	Representation-aware reliability validation

Table 2. Supplementary Table S2. Core RAVEN quantities and their direction of interpretation. The table is intended to prevent ambiguity when reading later tables and figures.

Quantity	Meaning	Reliability interpretation
C	Output confidence	Higher is nominally safer, but can be overconfident.
SC	Semantic consistency	Higher indicates greater semantic agreement.
MRI	Multi-level representation instability	Higher indicates stronger clean-to-shift representation change.
MRD	Representation displacement	Higher indicates greater relative representation displacement.
RARS	Joint reliability score	Higher indicates greater estimated reliability.
F	Prediction failure indicator	$F = 1$ denotes an incorrect shifted prediction.

Table 3. Supplementary Table S3. Evaluation layers and the question each metric addresses. The table explains why no single metric is treated as sufficient evidence of reliability.

Layer	Purpose and metrics
Association	Monotonic relationships and measured confounding: Spearman ρ and partial ρ .
Regression	Incremental association conditional on confidence and severity: β , SE, OR, random effects.
Detection	Ranking quality for incorrect predictions: AUROC, paired bootstrap, DeLong.
Selectivity	Accept/reject utility across coverage: AURC and Risk@Coverage.
Transfer	Portability beyond the core setting: cross-dataset and cross-backbone tests.
Robustness	Component necessity and internal agreement: ablations, extensions, consistency checks.

Table 4. Supplementary Table S4. Complete corruption-by-severity results (cell-macro protocol). Each corruption–severity cell receives equal weight. The table’s purpose is to show the monotonic behavior of accuracy, failure, MRI, confidence, and RARS across the controlled shift grid.

Corruption	Sev	Acc	Flip	MRI	MRD	Conf	RARS
Gaussian noise	1	0.730	0.16	0.073	0.35	0.82	0.670
Gaussian noise	2	0.640	0.26	0.111	0.42	0.75	0.568
Gaussian noise	3	0.530	0.38	0.162	0.51	0.66	0.452
Gaussian noise	4	0.410	0.51	0.227	0.62	0.56	0.326
Gaussian noise	5	0.280	0.62	0.317	0.74	0.46	0.198
Brightness	1	0.800	0.11	0.046	0.26	0.87	0.728
Brightness	2	0.740	0.18	0.069	0.31	0.82	0.652
Brightness	3	0.660	0.26	0.102	0.38	0.75	0.564
Brightness	4	0.570	0.35	0.146	0.46	0.67	0.466
Brightness	5	0.470	0.44	0.204	0.56	0.58	0.354
Contrast	1	0.770	0.14	0.054	0.29	0.85	0.704
Contrast	2	0.690	0.22	0.083	0.35	0.78	0.616
Contrast	3	0.590	0.31	0.124	0.43	0.69	0.518
Contrast	4	0.480	0.42	0.177	0.53	0.59	0.408
Contrast	5	0.360	0.53	0.246	0.64	0.48	0.286
Blur	1	0.810	0.10	0.038	0.24	0.88	0.742
Blur	2	0.750	0.17	0.058	0.29	0.83	0.670
Blur	3	0.680	0.24	0.085	0.35	0.77	0.586
Blur	4	0.600	0.33	0.124	0.42	0.70	0.492
Blur	5	0.500	0.42	0.174	0.51	0.62	0.390
Pixel dropout	1	0.750	0.15	0.062	0.32	0.83	0.682
Pixel dropout	2	0.670	0.24	0.097	0.39	0.76	0.588
Pixel dropout	3	0.570	0.35	0.142	0.48	0.67	0.480
Pixel dropout	4	0.450	0.47	0.198	0.58	0.57	0.356
Pixel dropout	5	0.330	0.57	0.273	0.70	0.47	0.230
Mean	All	0.5932	0.3146	0.1357	0.4338	0.6972	0.5090

Table 5. Supplementary Table S5. Within-dataset MRI–failure associations. The table’s purpose is to demonstrate that the central association persists within each dataset rather than arising solely from between-dataset differences.

Dataset	$\rho(\text{MRI}, F)$	95% CI	p	n
CIFAR10	0.452	[0.446, 0.458]	$< 10^{-300}$	125,000
CIFAR100	0.418	[0.412, 0.424]	$< 10^{-300}$	125,000
STL10	0.447	[0.441, 0.453]	$< 10^{-300}$	125,000
SVHN	0.502	[0.496, 0.508]	$< 10^{-300}$	125,000
FashionMNIST	0.463	[0.457, 0.469]	$< 10^{-300}$	125,000
DogsVsCats	0.491	[0.485, 0.497]	$< 10^{-300}$	125,000
Pooled	0.4618	[0.4603, 0.4633]	$< 10^{-300}$	750,000

Table 6. Supplementary Table S6. Within-corruption MRI–failure associations. The table’s purpose is to quantify heterogeneity across shift mechanisms while preserving the common positive direction.

Corruption	ρ	95% CI	n
Gaussian noise	0.538	[0.532, 0.544]	150,000
Brightness	0.424	[0.418, 0.430]	150,000
Contrast	0.468	[0.462, 0.474]	150,000
Blur	0.398	[0.392, 0.404]	150,000
Pixel dropout	0.485	[0.479, 0.491]	150,000
Macro mean	0.4626	[0.4566, 0.4686]	750,000

Table 7. Supplementary Table S7. Rank-based partial correlations for MRI and failure. The table’s purpose is to test whether the association survives measured sources of confounding. The final column reports the retained fraction of the raw association, computed as $100\rho_{\text{partial}}/0.4618$, so every row has a directly interpretable numeric quantity.

Conditioning set	Partial ρ	Retained vs raw (%)
None (raw)	0.4618	100.00
Severity	0.3184	68.95
Severity + dataset	0.2917	63.17
Severity + dataset + corruption	0.2643	57.23

Table 8. Supplementary Table S8. Paired AUROC differences for RARS. The table’s purpose is to distinguish statistically detectable gains from practically large gains.

Comparison	Δ AUROC	95% CI
RARS – MSP	+0.0551	[+0.0520, +0.0581]
RARS – Entropy	+0.0487	[+0.0456, +0.0518]
RARS – MaxLogit	+0.0400	[+0.0369, +0.0431]
RARS – Energy	+0.0338	[+0.0308, +0.0368]
RARS – kNN	+0.0209	[+0.0183, +0.0235]
RARS – Mahalanobis	+0.0053	[+0.0034, +0.0072]

Table 9. Supplementary Table S9. Leave-one-out component contribution. The table’s purpose is to quantify the marginal loss when each RARS component is removed. The final column reports retained AUROC as a percentage of the full score.

Component removed	AUROC	Retained vs full (%)
None (RARS full)	0.812	100.00
Semantic consistency	0.762	93.84
MRI	0.772	95.07
MRD	0.781	96.18

Table 10. Supplementary Table S10. Per-dataset component ablation. The table’s purpose is to show that the improvement of the full score is not confined to one dataset.

Dataset	Conf	Sem	C×Sem	C×MRI	C×MRD	Full
CIFAR10	0.745	0.698	0.768	0.774	0.766	0.802
CIFAR100	0.699	0.657	0.728	0.738	0.732	0.767
STL10	0.734	0.687	0.757	0.768	0.760	0.794
SVHN	0.784	0.747	0.808	0.814	0.810	0.838
FashionMNIST	0.762	0.722	0.784	0.792	0.786	0.816
DogsVsCats	0.818	0.784	0.838	0.842	0.840	0.860

Table 11. Supplementary Table S11. Calibration metrics for probability-like scores. RARS is intentionally excluded because it is not defined as a probability. The table’s purpose is to separate calibration from ranking-based reliability evaluation.

Method	ECE ↓	Brier ↓
MSP	0.147	0.198
Entropy (softmax-normalized)	0.152	0.204
MaxLogit (softmax)	0.149	0.201
Energy (softmax)	0.144	0.197
Temp-scaled MSP ($T = 1.4$)	0.061	0.171

Table 12. Supplementary Table S12. Cross-backbone validation on CIFAR10. The table’s purpose is to test whether the confidence-to-representation gain is specific to ResNet-18.

Backbone	Acc	Flip	RI _z	RD _z	Conf	RARS-lite	AUROC _R	AUROC _C
ResNet-18	0.562	0.381	0.213	0.512	0.698	0.472	0.812	0.742
ResNet-50	0.591	0.352	0.198	0.478	0.714	0.501	0.826	0.759
EfficientNet-B0	0.578	0.365	0.207	0.492	0.706	0.489	0.819	0.751

Table 13. Supplementary Table S13. Matched backbone and pretrained-encoder extension. The table’s purpose is to test whether representation-aware failure detection remains measurable across heterogeneous frozen encoders under the same evaluation logic and the recorded input preprocessing.

Backbone	Params	Dim.	AUROC _R	AUROC _C	Gain
ResNet-18	11.2M	512	0.812	0.742	+0.070
ResNet-50	23.5M	2048	0.818	0.751	+0.067
EfficientNet-B0	4.0M	1280	0.814	0.748	+0.066
DINOv2 ViT-S/14	21.6M	384	0.791	0.738	+0.053
DINOv2 ViT-B/14	86.6M	768	0.803	0.752	+0.051
CLIP ViT-B/32	87.8M	512	0.778	0.731	+0.047

Table 14. Supplementary Table S14. Generalization interpretation. This table records what each extension supports and what it does not establish, preventing overclaiming.

Test	Supported statement	Not established
Cross-dataset	Some transfer of ranking signal	Universal threshold invariance
Cross-backbone	Positive gain on 3 backbones	Broad architecture universality
Cross-corruption extension	Positive average association	Full unseen-domain guarantee
Bootstrap	Sampling stability	Multi-seed retraining stability

Table 15. Supplementary Table S15. Selected extension experiments from the supplied 20-experiment suite. The table’s purpose is to expose secondary checks rather than hide them in supplementary material.

Experiment	Setting	Result
E01	ρ (severity, MRI), six datasets	0.812, 0.836, 0.798, 0.783, 0.804, 0.762; all $p < 10^{-10}$
E02	All, ρ (severity, RARS)	-0.821
E03	All, ρ (MRI, flip)	0.487
E04	All, ρ (MRI, SC)	-0.712
E05	RI-failure by depth	$\rho_z = 0.469, \rho_h = 0.452, \rho_{final} = 0.413$ 0.452 vs 0.461
E07	Raw vs bounded mean cosine	
E08	Bootstrap ρ	0.461 [0.441, 0.482]
E09	ECE confidence vs RARS	0.147 vs 0.183
E10	AURC confidence vs RARS	0.198 vs 0.176
E13	Class-wise RARS SD	0.043
E14	Severity-confidence ρ	-0.742
E17	ρ (MRI, MRD)	0.723
E18	ρ without vs with semantic	0.421 vs 0.572
E19	Cross-corruption ρ train vs test	0.573 vs 0.501
E20	Median training time / dataset	88.5 min

Table 16. Supplementary Table S16. Representative RARS weight-sensitivity variants. The table’s purpose is to determine whether the reliability score is fragile to moderate reweighting of confidence, semantic consistency, MRI, and MRD.

Variant	α	β	γ	δ	AUROC
Baseline (equal)	1.0	1.0	1.0	1.0	0.812
MRI-heavy	1.0	1.0	1.5	1.0	0.814
MRI-light	1.0	1.0	0.5	1.0	0.808
MRD-heavy	1.0	1.0	1.0	1.5	0.809
MRD-light	1.0	1.0	1.0	0.5	0.810
Confidence-heavy	1.3	1.0	1.0	1.0	0.806
Confidence-light	0.7	1.0	1.0	1.0	0.815
Semantic-heavy	1.0	1.3	1.0	1.0	0.813
Semantic-light	1.0	0.7	1.0	1.0	0.809
All-mild	0.8	0.8	0.8	0.8	0.811
All-strong	1.2	1.2	1.2	1.2	0.810

Table 17. Supplementary Table S17. Runtime and parameter audit. The table’s purpose is to document computational cost, which is relevant to practical adoption and reproducibility.

Dataset	Train min	Shift min	Total min	Params	Train/Test
CIFAR10	74	22	96	166,026	15k/5k
CIFAR100	81	24	105	177,636	15k/5k
STL10	164	26	190	166,026	15k/5k
SVHN	58	20	78	166,026	15k/5k
FashionMNIST	96	25	121	166,026	15k/5k
DogsVsCats	223	28	251	164,994	15k/5k
Total	696	145	841	1,006,734	90k/30k

Table 18. Supplementary Table S18. Contribution map. The table’s purpose is to separate conceptual novelty, measurement design, empirical evidence, and operational validation so that no single performance number is asked to carry the entire contribution claim.

Contribution	Core claim	Evidence in supplied record
C1	Representation-aware reliability validation	RARS combines C, SC, MRI, and MRD.
C2	Incremental information beyond confidence	Partial $\rho = 0.2643$; MRI $\beta = 1.424$; RARS vs MSP +0.0551 AUROC.
C3	Multi-level representation analysis	Severity-RI correlation rises from 0.4312 to 0.8015 across levels.
C4	Broad empirical characterization within the evaluated scope	Six datasets, five corruptions, five severities, transfer, three backbones, ablations, selectivity, and runtime audit.

Table 19. Supplementary Table S19. Failure-mode interpretation. The table’s purpose is to state what RAVEN can and cannot infer from its internal signals.

Observed pattern	Interpretation
High confidence + high MRI	Confidence may be misleading; representation instability raises risk signal.
Low confidence + high MRI	Both output and representation evidence indicate caution.
Low RARS	Suitable candidate for selective rejection under the evaluated risk-coverage behavior.
High confidence + low MRI + wrong prediction	Important blind spot; representation stability does not guarantee correctness.

Table 20. Supplementary Table S20. Limitations and corresponding mitigation or future experiment. The table’s purpose is to make the validity boundary explicit for journal review.

Limitation	Mitigation / future work
One core backbone	Repeat full protocol across diverse CNNs and transformers.
Synthetic corruptions only	Add natural-shift benchmarks such as ImageNet-V2/ImageNet-R.
Single training seed	Retrain across multiple seeds and report between-seed intervals.
Limited pretrained-encoder scope	Evaluate DINOv2, CLIP, MAE, and Swin under the documented frozen-encoder, linear-probe, and preprocessing protocols.
RARS not a probability	Add explicit calibration if a probabilistic interpretation is required.
Stable-but-wrong failures	Investigate semantic/OOD diagnostics beyond representation stability.
Runtime > 11 h core	Profile feature extraction and evaluate lower-cost layer subsets.

Table 22. Supplementary Table S22. Complete Oxford-IIIT Pet breed-to-species mapping used for the DogsVsCats $\rightarrow$ Oxford-Pets natural-shift evaluation. The target’s 37 breed labels are collapsed to the source binary label space, with the first 12 breeds being cats and the remaining 25 breeds being dogs in the canonical dataset ordering. Source: Oxford-IIIT Pet dataset and Parkhi et al. [9].

ID	Oxford-IIIT Pet breed	Harmonized label
1	Abyssinian	Cat
2	Bengal	Cat
3	Birman	Cat
4	Bombay	Cat
5	British Shorthair	Cat
6	Egyptian Mau	Cat
7	Maine Coon	Cat
8	Persian	Cat
9	Ragdoll	Cat
10	Russian Blue	Cat
11	Siamese	Cat
12	Sphynx	Cat
13	American Bulldog	Dog
14	American Pit Bull Terrier	Dog
15	Basset Hound	Dog
16	Beagle	Dog
17	Boxer	Dog
18	Chihuahua	Dog
19	English Cocker Spaniel	Dog
20	English Setter	Dog
21	German Shorthaired	Dog
22	Great Pyrenees	Dog
23	Havanese	Dog
24	Japanese Chin	Dog
25	Keeshond	Dog
26	Leonberger	Dog
27	Miniature Pinscher	Dog
28	Newfoundland	Dog
29	Pomeranian	Dog
30	Pug	Dog
31	Saint Bernard	Dog
32	Samoyed	Dog
33	Scottish Terrier	Dog
34	Shiba Inu	Dog
35	Staffordshire Bull Terrier	Dog
36	Wheaten Terrier	Dog
37	Yorkshire Terrier	Dog

Table 21. Supplementary Table S21. Complete CIFAR-100 fine-to-coarse label mapping used for the CIFAR-100 $\rightarrow$ superclass natural-shift evaluation. Each superclass contains exactly five native fine labels; the mapping is fixed before evaluation and no target-domain refitting is performed. Source: CIFAR-100 dataset specification [6].

Coarse superclass	Native CIFAR-100 fine classes
Aquatic mammals	beaver; dolphin; otter; seal; whale
Fish	aquarium fish; flatfish; ray; shark; trout
Flowers	orchids; poppies; roses; sunflowers; tulips
Food containers	bottles; bowls; cans; cups; plates
Fruit and vegetables	apples; mushrooms; oranges; pears; sweet peppers
Household electrical devices	clock; computer keyboard; lamp; telephone; television
Household furniture	bed; chair; couch; table; wardrobe
Insects	bee; beetle; butterfly; caterpillar; cockroach
Large carnivores	bear; leopard; lion; tiger; wolf
Large man-made outdoor things	bridge; castle; house; road; skyscraper
Large natural outdoor scenes	cloud; forest; mountain; plain; sea
Large omnivores and herbivores	camel; cattle; chimpanzee; elephant; kangaroo
Medium-sized mammals	fox; porcupine; possum; raccoon; skunk
Non-insect invertebrates	crab; lobster; snail; spider; worm
People	baby; boy; girl; man; woman
Reptiles	crocodile; dinosaur; lizard; snake; turtle
Small mammals	hamster; mouse; rabbit; shrew; squirrel
Trees	maple; oak; palm; pine; willow
Vehicles 1	bicycle; bus; motorcycle; pickup truck; train
Vehicles 2	lawn-mower; rocket; streetcar; tank; tractor

Table 23. Supplementary Table S23. Evidence-provenance ledger for the main empirical tables. The ledger maps each main table to its experiment block, population/protocol, metric, and statistical procedure so that a reviewer can trace a reported number without relying on inference from the table title.

Main table	Record block	Dataset / protocol	Metric	Statistical procedure
Core configuration	Six datasets; seed 42; frozen ResNet-18; 96 px; 5 corruptions $\times$ 5 severities	Configuration values	Direct protocol record	
Dataset-level summary	Six core datasets; sample-weighted aggregation	Acc, flip, MRI, MRD, confidence, RARS	Descriptive aggregation	
Multi-level representation	Six levels; paired clean-shift observations	ρ (severity, RI), ρ (RI, failure)	Spearman; 1,000 bootstrap resamples where reported	
Primary hypothesis tests	Pooled core evaluation; 750,000 observations	Spearman ρ	Rank correlation; bootstrap uncertainty	
Mixed-effect regression	Pooled shifted dataset/corruption random intercepts	β , SE, OR	Mixed-effect logistic regression	
Overall AUROC	Six-dataset pooled failure detection	AUROC	1,000 bootstrap resamples for reported CIs; paired 10,000 bootstrap for differences; DeLong secondary	
Component ablation	Same pooled failure-detection population	AUROC	Descriptive ablation comparison	
Selective prediction	Same pooled shifted population	AURC, risk@coverage	Risk-coverage evaluation	
Cross-dataset transfer	Six source $\rightarrow$ target pairs; no target refitting	AUROC, coverage, risk	Descriptive transfer evaluation	
Natural-shift evaluation	Six fixed source $\rightarrow$ target pairs; explicit label harmonization	Acc, MRI, RARS, AUROC, gain	Descriptive transfer evaluation	
Five-seed reproducibility	Six datasets; seeds {17, 42, 73, 101, 2024}	ρ , AUROC, AURC	Mean $\pm$ SD across retraining runs	
Heterogeneous encoder master	Eight frozen encoders; architecture-specific pre-processing; 5-epoch linear probe	AUROC, gain, MRI ρ	1,000 bootstrap CIs for AUROC; descriptive cross-encoder comparison	
MAE/Swin preprocessing ablation	Native-to-224 P2 plus P1, interpolation, 384-pixel, and five-crop variants	AUROC	Descriptive sensitivity analysis	

Table 24. Supplementary Table S24. Matched pretrained-encoder protocol audit. The table records the exact input contract, feature extraction rule, frozen policy, and probe protocol used for the heterogeneous-encoder extension. The distinction between the 96-pixel CNN/recorded DINOv2-CLIP path and the native-to-224 MAE/Swin path is intentional and is retained in the reported results.

Encoder	Frozen	Input / preprocess-ing	Feature	Probe / optimization
ResNet-18 / 50	Yes	96 $\times$ 96 loader input	Final representation; dim 512/2048	Linear probe, 5 epochs; AdamW, lr 3×10^{-4} , wd 10^{-4} , batch 128
EfficientNet-B0	Yes	96 $\times$ 96 loader input	Final representation; dim 1280	Same linear-probe optimization
DINOv2 ViT-S/B	Yes	96-pixel loader input; model-internal resize to 224	Final representation; dim 384/768	Linear probe, 5 epochs; AdamW, lr 3×10^{-4} , wd 10^{-4} , batch 128
CLIP ViT-B/32	Yes	96-pixel loader input; model-internal resize to 224	Visual representation; dim 512	Linear probe, 5 epochs; AdamW, lr 3×10^{-4} , wd 10^{-4} , batch 128
MAE ViT-B/16	Yes	Native image $\rightarrow$ 224 $\times$ 224, bicubic (P2)	Final-block CLS; dim 768; patch 16 $\times$ 16	Linear probe 768 $\rightarrow$ C, 5 epochs; AdamW, lr 3×10^{-4} , wd 10^{-4} , batch 128
Swin-T	Yes	Native image $\rightarrow$ 224 $\times$ 224, bicubic (P2)	Global average pool after final stage; dim 768; patch 4 $\times$ 4, window 7 $\times$ 7	Linear probe 768 $\rightarrow$ C, 5 epochs; AdamW, lr 3×10^{-4} , wd 10^{-4} , batch 128

30. MAE ViT-B/16 – Detailed Results

The MAE extension uses the facebookresearch/mae checkpoint `mae_pretrain_vit_base`, loaded with `torch.hub.load("facebookresearch/mae", "mae_pretrain_vit_base")`. The encoder is frozen. Images are resized from native resolution to 224 $\times$ 224 using bicubic interpolation (P2), and the final-block CLS token (768 dimensions) is passed to a linear probe. The probe is trained for five epochs with AdamW, learning rate 3×10^{-4} , weight decay 10^{-4} , and batch size 128. The MAE pretraining objective is masked autoencoding with 75% masking; the classifier head is not imported from pretraining. The reported parameter count is 85.8M; the supplied checkpoint metadata reports 83.6% ImageNet fine-tuning top-1, although no classifier from that fine-tuning is used here. The extension was executed on CPU; throughput was not recorded.

Table 25. Supplementary Table S30. MAE ViT-B/16 failure-detection AUROC by dataset. Provenance: extension Section 30; 1,000 bootstrap resamples for the confidence intervals.

Dataset	AUROC _R	95% CI	AUROC _C	Gain
CIFAR10	0.774	[0.763,0.785]	0.724	+0.050
CIFAR100	0.741	[0.730,0.752]	0.691	+0.050
STL10	0.768	[0.757,0.779]	0.719	+0.049
SVHN	0.801	[0.790,0.812]	0.746	+0.055
FashionMNIST	0.786	[0.775,0.797]	0.731	+0.055
DogsVsCats	0.822	[0.812,0.832]	0.771	+0.051
Pooled	0.782	[0.780,0.784]	0.729	+0.053

Table 26. Supplementary Table S30b. MAE MRI-failure association. Confidence intervals are the supplied 95% bootstrap intervals; the very small pooled p -values are reported descriptively rather than used as the principal evidence.

Dataset	ρ	95% CI	p
CIFAR10	0.412	[0.406,0.418]	$< 10^{-300}$
CIFAR100	0.386	[0.380,0.392]	$< 10^{-300}$
STL10	0.408	[0.402,0.414]	$< 10^{-300}$
SVHN	0.445	[0.439,0.451]	$< 10^{-300}$
FashionMNIST	0.421	[0.415,0.427]	$< 10^{-300}$
DogsVsCats	0.438	[0.432,0.444]	$< 10^{-300}$
Pooled	0.418	[0.417,0.419]	$< 10^{-300}$

Table 27. Supplementary Table S30c. MAE RARS AURC across datasets, with multi-seed standard deviations from the extension record.

Dataset	AURC	SD
CIFAR10	0.198	± 0.009
CIFAR100	0.221	± 0.011
STL10	0.204	± 0.010
SVHN	0.178	± 0.008
FashionMNIST	0.190	± 0.009
DogsVsCats	0.168	± 0.008
Pooled	0.193	± 0.0091

Table 28. Supplementary Table S30d. MAE MRI-failure association by corruption family. The final column is the reported difference from ResNet-18.

Corruption	ρ	Δ vs ResNet-18
gaussian_noise	0.478	-0.060
brightness	0.382	-0.042
contrast	0.428	-0.040
blur	0.361	-0.037
pixel_dropout	0.442	-0.043

The MAE result is lower than the ResNet-18 reference for every listed corruption family in the supplied record. The interpretation is restricted to this benchmark: the result is consistent with, but does not by itself establish, a mismatch between reconstruction pretraining and corruption-sensitive reliability.

31. Swin-T – Detailed Results

Swin-T uses the torchvision ImageNet-1K checkpoint Swin_T_Weights.IMAGENET1K_V1; the encoder is frozen. Native images are resized to 224×224 with bicubic interpolation (P2). The final-stage representation is globally average-pooled to 768 dimensions and passed to a linear probe. The probe is trained for five epochs with AdamW, learning rate 3×10^{-4} , weight decay 10^{-4} , and batch size 128. The architecture uses patch size 4×4 and window size 7×7 ; the recorded parameter count is 28.3M. The supplied checkpoint metadata reports 81.5% ImageNet top-1, and the extension was executed on CPU; throughput was not recorded.

Table 29. Supplementary Table S31. Swin-T failure-detection AUROC by dataset. Provenance: extension Section 31; 1,000 bootstrap resamples for the confidence intervals.

Dataset	AUROC _R	95% CI	AUROC _C	Gain
CIFAR10	0.787	[0.776,0.798]	0.738	+0.049
CIFAR100	0.752	[0.741,0.763]	0.704	+0.048
STL10	0.781	[0.770,0.792]	0.732	+0.049
SVHN	0.812	[0.801,0.823]	0.762	+0.050
FashionMNIST	0.798	[0.787,0.809]	0.746	+0.052
DogsVsCats	0.838	[0.828,0.848]	0.784	+0.054
Pooled	0.795	[0.793,0.797]	0.744	+0.051

Table 30. Supplementary Table S31b. Swin-T MRI-failure association. Confidence intervals are the supplied 95% bootstrap intervals.

Dataset	ρ	95% CI	p
CIFAR10	0.431	[0.425,0.437]	$< 10^{-300}$
CIFAR100	0.401	[0.395,0.407]	$< 10^{-300}$
STL10	0.428	[0.422,0.434]	$< 10^{-300}$
SVHN	0.468	[0.462,0.474]	$< 10^{-300}$
FashionMNIST	0.441	[0.435,0.447]	$< 10^{-300}$
DogsVsCats	0.462	[0.456,0.468]	$< 10^{-300}$
Pooled	0.438	[0.437,0.439]	$< 10^{-300}$

Table 31. Supplementary Table S31c. Swin-T RARS AURC across datasets.

Dataset	AURC	SD
CIFAR10	0.189	± 0.008
CIFAR100	0.212	± 0.010
STL10	0.196	± 0.009
SVHN	0.171	± 0.007
FashionMNIST	0.182	± 0.008
DogsVsCats	0.161	± 0.007
Pooled	0.185	± 0.0084

Table 32. Supplementary Table S31d. Swin-T MRI-failure association by corruption family.

Corruption	ρ	Δ vs ResNet-18
gaussian_noise	0.508	-0.030
brightness	0.401	-0.023
contrast	0.445	-0.023
blur	0.378	-0.020
pixel_dropout	0.462	-0.023

Swin-T remains closer to ResNet-18 than MAE on the supplied corruption suite. The interpretation is deliberately architectural rather than causal: the results are compatible with the possibility that hierarchical feature aggregation retains more corruption-relevant structure, but the experiment does not isolate inductive bias as a causal factor.

32. Updated Pretrained-Encoder Master Table

The master table below is the compact reference for the heterogeneous-encoder extension. It is not intended to replace the detailed provenance and per-dataset tables above; its purpose is to let a reviewer audit the central comparison in one place.

Table 33. Supplementary Table S32. Updated master comparison across the eight evaluated encoders. The table preserves the supplied AUROC, confidence baseline, gain, MRI correlation, and AURC values.

Backbone	Params M	Dim	AUROC _R	95% CI	AUROC _C	Gain	MRI ρ
ResNet-50	23.5	2048	0.818	[0.816,0.820]	0.751	+0.067	0.472
EfficientNet-B0	4.0	1280	0.814	[0.812,0.816]	0.748	+0.066	0.465
ResNet-18	11.2	512	0.812	[0.810,0.814]	0.742	+0.070	0.462
DINOv2 ViT-B/14	86.6	768	0.803	[0.801,0.805]	0.752	+0.051	0.446
Swin-T	28.3	768	0.795	[0.793,0.797]	0.744	+0.051	0.438
DINOv2 ViT-S/14	21.6	384	0.791	[0.789,0.793]	0.738	+0.053	0.432
MAE ViT-B/16	85.8	768	0.782	[0.780,0.784]	0.729	+0.053	0.418
CLIP ViT-B/32	87.8	512	0.778	[0.776,0.780]	0.731	+0.047	0.401

The gain over confidence is positive for all eight encoders in the supplied record. The table should not be read as a universal ranking of pretraining paradigms: the models differ in architecture, input contract, pretraining objective, and feature dimensionality even though the downstream probe protocol is held fixed. The result supports a narrower statement that representation-aware failure detection remains measurable across this heterogeneous set.

33. Preprocessing Ablation – P1 versus P2

The preprocessing ablation quantifies whether the MAE/Swin backbone comparison could be driven by the input path rather than the representation itself. P2 is the main condition: native image resized to 224×224 using bicubic interpolation. P1 is the earlier 96-pixel loader path followed by bicubic resize to 224. Additional conditions vary interpolation, resolution, or test-time cropping while leaving the frozen encoder and probe unchanged.

Table 34. Supplementary Table S33a. MAE ViT-B/16 preprocessing ablation.

Protocol	Input	Interp.	AUROC _R	Δ vs P2
P2 (main)	Native $\rightarrow 224$	BICUBIC	0.782	0.000
P1	96 $\rightarrow 224$	BICUBIC	0.775	-0.007
P2-BILINEAR	Native $\rightarrow 224$	BILINEAR	0.779	-0.003
P2-384	Native $\rightarrow 384$	BICUBIC	0.784	+0.002
P2-multicrop	Native $\rightarrow 224 \times 5$	BICUBIC	0.786	+0.004

Table 35. Supplementary Table S33b. Swin-T preprocessing ablation.

Protocol	Input	Interp.	AUROC _R	Δ vs P2
P2 (main)	Native $\rightarrow 224$	BICUBIC	0.795	0.000
P1	96 $\rightarrow 224$	BICUBIC	0.789	-0.006
P2-BILINEAR	Native $\rightarrow 224$	BILINEAR	0.792	-0.003
P2-384	Native $\rightarrow 384$	BICUBIC	0.797	+0.002
P2-multicrop	Native $\rightarrow 224 \times 5$	BICUBIC	0.798	+0.003

Table 36. Supplementary Table S33c. Interpretation of preprocessing sensitivity.

Tested variation	Observed AUROC effect
Native $\rightarrow 224$ versus 96 $\rightarrow 224$	+0.006 to +0.007
BICUBIC versus BILINEAR	+0.003
384 versus 224	+0.002
Five-crop TTA versus single crop	+0.003 to +0.004

The largest tested preprocessing effect is approximately 0.007 AU-ROC, whereas the full encoder spread in the master comparison is approximately 0.040. This does not prove preprocessing irrelevance; it bounds the sensitivity within the tested alternatives and justifies retaining P2 as the main MAE/Swin condition because it matches the native input contract of the pretrained checkpoints more closely than the 96-to-224 path. The comparison is therefore explicitly conditioned on the documented preprocessing protocol rather than presented as a preprocessing-free property of the encoders.

References

- [1] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 1321–1330 (2017).
- [2] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017).
- [3] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems 31 (2018).
- [4] W. Liu, X. Wang, J.D. Owens, Y. Li, Energy-based out-of-distribution detection. In: Advances in Neural Information Processing Systems 33 (2020).
- [5] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30 (2017).
- [6] A. Krizhevsky, Learning multiple layers of features from tiny images. Technical Report, University of Toronto (2009).
- [7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16000–16009 (2022).
- [8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022 (2021).
- [9] O.M. Parkhi, A. Vedaldi, A. Zisserman, C.V. Jawahar, Cats and Dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3498–3505 (2012).