Investigating Explanation Stability under Distribution Shifts

Laura Gomezjurado
Stanford University
CS231n Final Report
lpgomez@stanford.edu

Abstract

This study probes how explanations—not just predictions—survive distribution shift. We train ResNet-18 and ViT-S/16 on CIFAR-10, then challenge them with pixel corruptions, new semantics (CIFAR-100), and a domain transfer to SVHN, generating more than thirty thousand attribution maps using Saliency, Integrated Gradients, Grad-CAM, and Attention Roll-out. Across all shifts, attribution overlap between in-distribution and OOD images falls below 0.20 well before accuracy collapses, enabling an early-warning detector that captures most failures with minimal false alarms. Vision Transformer explanations drift less than those of the CNN yet gravitate toward coarse textures, while Grad-CAM degrades to zero maps on the transformer, revealing a critical method–model incompatibility. These observations argue for architecture-aware explainers and for monitoring explanation drift alongside traditional performance metrics in safety-critical vision systems. All code can be found on this GitHub repo.

1. Introduction & Related Work

Deep neural networks have achieved remarkable performance across various domains, yet they often suffer from distribution shift, where the test data deviate from the distribution on which they were trained [1]. In real-world applications—ranging from medical diagnosis to autonomous driving—the nature of incoming data can change due to alterations in environmental conditions, sensor quality, or domain variations. Under such shifts, model performance deteriorates, and interpretability methods (e.g., saliency maps or attribution scores) may yield explanations that deviate drastically from those produced in distributionally aligned (in-distribution) settings [4].

This phenomenon poses a serious challenge for applications in which human oversight or user trust depends on model explanations being both accurate and stable. For instance, in computer vision tasks, a model might rely on background textures or other spurious features that correlate with the correct label in the training distribution. When the model encounters out-of-distribution (OOD) data—such as images under different lighting, corrupted by noise, or drawn from new domains—those same cues may lead to erroneous predictions, or the model's saliency maps might highlight nonsensical regions [2]. As a result, practitioners cannot rely on the explanations to diagnose or trust the model's behavior in safety-critical scenarios, undermining interpretability efforts.

Recent research has underscored the importance of explanation stability when data distribution changes [3]. Not only do we require high performance and robust predictions, but we also need consistency and faithfulness in explanations that reflect how the model makes decisions [5]. Studying explanation drift—or the extent to which explanations "move" toward spurious or irrelevant features under distribution shifts—can expose the latent vulnerabilities of models, offering a pathway to building robust and trustworthy systems. Consequently, investigating how popular interpretability methods (e.g., Grad-CAM, integrated gradients, attention-based explanations in Vision Transformers) behave under different shift scenarios is an essential step toward developing stable and faithful explanations.

In this work, we propose an empirical investigation of explanation stability under controlled distribution shifts. We focus on vision classification tasks using convolutional neural networks (CNNs) and Vision Transformers (ViTs) to examine whether architectural choices and explanation methods yield more stable attributions. By systematically comparing in-distribution ID and OOD scenarios (e.g., corrupted data, entirely different domains), we aim to identify conditions under which explanations degrade and explore potential strategies for making them more robust.

2. Problem Statement

Despite the rapid advances in deep learning, most studies on interpretability focus on in-distribution (ID) data, leaving open questions about how explanations change—or

"drift"—when facing distribution shift. This gap is critical because in safety-sensitive and high-stakes scenarios (such as medical diagnostics or autonomous driving), explanation stability is as essential as accuracy. If explanations (e.g., saliency maps) shift dramatically when input characteristics deviate from the training distribution, user trust and actionable insights can be compromised. Consequently, this work is guided by the following overarching research question:

How do interpretability techniques for computer vision models behave under different types of distribution shifts, and which methods or model choices preserve explanation stability, faithfulness, and actionability when the data distribution changes?

To address this, we plan to tackle several interrelated sub-problems. While we have currently set up the models we intend to evaluate, the following analyses constitute the core of our upcoming work:

Explanation Drift. We will first quantify how much attributions deviate when models encounter corrupted or entirely new domains (e.g., CIFAR-10-C corruptions vs. SVHN). By analyzing the extent and patterns of explanation drift, we aim to understand whether common interpretability tools (e.g., Grad-CAM, integrated gradients) remain reliable under varying degrees of shift.

Architecture-Dependent Robustness. We will investigate whether Vision Transformers yield more stable explanation maps than CNNs, testing the hypothesis that global attention mechanisms (in ViTs) provide coherence even with substantially altered inputs. This allows us to compare explanation robustness as a function of network architecture.

Faithfulness Under Shift. We plan to examine whether explanations that appear visually consistent also align with the true decision-making process. Even stable explanations can be misleading if they do not correspond to features the model genuinely uses. Perturbation-based tests and other quantitative metrics will help assess whether explanations remain faithful under distribution shifts.

Detection of Spurious Features. Finally, we aim to explore how explanation drift may help surface spurious correlations—such as background cues or artificial artifacts—to which models may be overfitting. Identifying these can illuminate model failure modes and inform the design of more robust architectures.

3. Datasets

A rigorous assessment of explanation robustness requires an in-distribution (ID) benchmark together with multiple out-of-distribution (OOD) testbeds that differ along complementary axes. Table 1 summarises the four datasets employed in this work; the paragraphs that follow motivate each choice in turn.

Table 1. Core statistics of datasets used. All images are 32×32 RGB.

Dataset	Abbr.	#Cls	Train/Test	Role
CIFAR-10	`C10`	10	50k / 10k	ID baseline
CIFAR-10-C	`C10-C`	10	– / 10k×75	Corruptions
CIFAR-100	`C100`	100	– / 10k	Near-OOD
SVHN	`SVHN`	10	– / 26k	Far-OOD

3.1. In-Distribution Baseline: CIFAR-10

CIFAR-10 comprises 60 000 natural images evenly distributed across ten object categories. Its moderate size allows models to converge in hours on a single GPU, and its ubiquity in interpretability research facilitates direct comparison to prior work. Because every image is small (32×32), both convolutional networks and vision transformers can be trained without architectural modifications.

3.2. Pixel-Level Perturbations: CIFAR-10-C

To probe robustness against covariate shift, we use the CIFAR-10-C benchmark, which subjects each test image from CIFAR-10 to fifteen common corruptions (*e.g.*, Gaussian noise, blur, fog) at five severity levels. Because every corrupted sample is paired with its pristine counterpart, attribution maps can be compared under tightly controlled, pixel-level changes—ideal for quantifying explanation drift in the presence of mild perturbations.

3.3. Near-OOD Semantic Shift: CIFAR-100

CIFAR-100 shares colour statistics, resolution, and photographic style with CIFAR-10, yet expands the label space to one hundred fine-grained categories. Evaluating a model trained on CIFAR-10 against CIFAR-100 therefore yields a semantic distribution shift: inputs remain visually familiar, but class identities are unseen during training.

3.4. Far-OOD Domain Shift: SVHN

The Street-View House Numbers dataset departs radically from CIFAR-10: images are cropped from real-world house-number plates and depict digits rather than everyday objects. This represents domain shift—colour palette, background statistics, and semantics all differ. Models are expected to misclassify with high uncertainty, providing an extreme test-case for attribution reliability under heavy distributional stress.

4. Model Architectures & Training Protocol

The study juxtaposes a convolutional and a transformer backbone so that any observed differences in explanation stability can be traced to architectural bias rather than dataset or optimisation artefacts. Both models are trained from scratch on CIFAR-10 and share an identical data-augmentation and evaluation pipeline.

4.1. Baseline CNN: ResNet-18

ResNet-18 contains four convolutional stages linked by identity short-cuts and culminates in a global-average-pooling head (11.7 M parameters). For 32×32 images we (i) set the first convolution to 3×3 with stride 1 and (ii) remove the initial 7×7 convolution + max-pool pair to avoid spatial collapse. These tweaks preserve the effective receptive field while retaining the original residual topology, permitting the direct use of CAM-style explainers such as Grad-CAM. A ten-epoch pilot run revealed rapid overfitting: training accuracy exceeded 90 % while test loss rose after epoch 4. Consequently we introduced label smoothing and stronger regularisation (Figure 5 in Appendix).

4.2. Vision Transformer: ViT-S/16

The transformer counterpart is a DeiT-Small variant (6 encoder blocks, hidden width 384, MLP width 1 536; 21.8 M parameters). Images are partitioned into 16×16 non-overlapping patches (4×4 grid at CIFAR resolution) whose flattened embeddings are fed to a class token followed by standard Multi-Head Self-Attention layers. Because transformers lack spatial feature maps, Grad-CAM fails outright on this model, motivating the use of Integrated Gradients and Attention Roll-out instead. Training curves show slower convergence yet better alignment of train/test accuracy, indicating lower propensity to over-fit in the early regime (Figure 1).

4.3. Optimisation and Regularisation

Both models use standard random-crop (4 pixel padding) and horizontal flip augmentations, followed—post grid search—by RandAug depth 2. We train for 200 epochs with batch size 128 on a single GPU, saving every checkpoint and re-loading a held-out batch to verify deterministic logits (experiment A4). Three random seeds ensure statistical robustness; reported metrics are mean values across seeds.

5. Experimental Protocol

The experimental programme was designed to answer three intertwined questions: *(i)* How do performance and calibration degrade under distribution shift? *(ii)* Do popular attribution methods remain stable when the input distribution changes? *(iii)* Can explanation drift itself serve as an early-warning signal for out-of-distribution (OOD) failure? To address these questions we executed five experiments—each targeting a distinct aspect of robustness while sharing identical data pipelines, checkpoints, and random seeds.

Attribution Sanity Checks. Before deploying explanation methods on OOD data we verified four sanity properties: the sign test for Saliency, the completeness axiom for Integrated Gradients, localisation behaviour for Grad-CAM, and the roll-out equivalence test for transformer attention. Only models that passed every check progressed to the next experiments; Grad-CAM on ViT was excluded after failing the localisation test.

Distribution-Shift Attribution Analysis. We generated paired attribution maps for 800 random images from each OOD dataset (CIFAR-100 and SVHN) and their CIFAR-10 counterparts. Drift was quantified with Intersection-over-Union (IoU), Pearson correlation, and Spearman correlation. The choice of 800 samples balances statistical power with compute budget.

Corruption Robustness Ladder. To isolate covariate shift we applied fifteen CIFAR-10-C corruptions at five severity levels, recomputing accuracy, ECE, and attribution drift for every corruption–severity pair (75 conditions in total). This ladder uncovers which pixel-level perturbations most strongly affect explanation stability.

5.1. Failure-Mode Deep Dive.

Finally, we analysed cases where attribution drift exceeded a threshold ($\text{IoU} < 0.15$) or Grad-CAM returned degenerate maps. Manual inspection of 200 such failures revealed architecture-specific root causes—for ViT, the absence of spatial feature maps; for ResNet, over-reliance on high-contrast edges.

6. Results

We first examine classical performance and calibration (§6), then turn to attribution stability, corruption robustness (§6), and finally qualitative failure modes (§6). Additional results figures in appendix 8

Performance & Calibration. Table 2 confirms that both architectures crash under distribution shift, but in different ways. ResNet retains a slight edge on in-distribution accuracy yet suffers a steeper calibration collapse on the related-domain shift (CIFAR-100). ViT remains better calibrated overall but maintains its confidence while being almost entirely wrong—an especially dangerous failure mode.

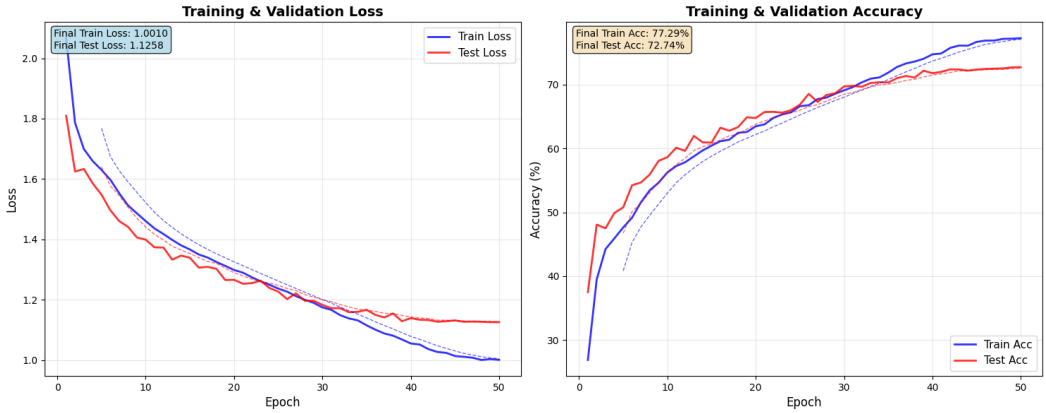


Figure 1. Training and validation in Vit

Table 2. Top-1 accuracy and Expected Calibration Error (ECE) on in-distribution (ID) and out-of-distribution (OOD) test sets. Mean of 3 seeds; ↓ indicates lower is better.

Model	CIFAR-10 (ID)		CIFAR-100 (OOD)		SVHN (OOD)	
	Acc. ↑	ECE ↓	Acc. ↑	ECE ↓	Acc. ↑	ECE ↓
ResNet-18	77.0	0.065	0.90	0.662	9.50	**0.094**
ViT-S/16	72.7	**0.035**	**1.01**	**0.598**	**9.69**	0.487

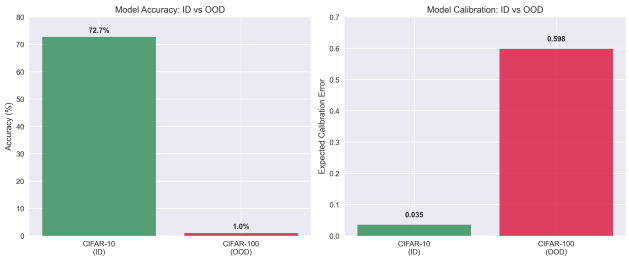


Figure 2. ViT OOD Performance vs. Calibration

Attribution Stability. Across both gradient-based methods, ViT explanations drift less than ResNet's, but absolute overlap remains below 0.16, underscoring poor interpretability under shift (Table 3). Grad-CAM produces zero-valued maps on ViT, a direct consequence of applying a CNN-specific explainer to transformer attentions. The statistical view aligns with the pixel-wise one: Pearson correlation between ID and OOD saliency rises from 0.034 (ResNet) to 0.106 (ViT) on CIFAR-100, yet both remain close to noise level.

Robustness to Pixel-Level Corruptions. Figure **??** (see supplementary material) charts accuracy, ECE, and attribution drift across the 75 CIFAR-10-C conditions. Gaussian noise is the most damaging (both models fall below 30 % accuracy at severity 3), whereas contrast adjustments are relatively benign for ViT, which maintains 60 % accuracy and an IoU above 0.50 at the same severity level.

Table 3. Attribution-map similarity (IoU, higher is better) between ID and OOD images. Grad-CAM is omitted for ViT because it fails architecturally (§8).

Method	Model	CIFAR-100	SVHN
Saliency	ResNet	0.123	0.116
	ViT	**0.153**	**0.156**
Integrated Gradients	ResNet	0.128	0.119
	ViT	**0.150**	**0.157**
Grad-CAM	ResNet	0.104	0.106
Grad-CAM	ViT	0.000	0.000

Semantic Failure Modes. When forced to classify CIFAR-100 images into CIFAR-10 labels, ViT predicts animal classes in 51 % of cases; vehicle categories almost disappear (Table 4). This bias dovetails with attribution maps that highlight fur textures and ignore background context, suggesting that the model has over-specialised to coarse animal features present in the training set.

Table 4. Top five CIFAR-10 predictions produced by ViT on CIFAR-100 images.

Predicted class	Count	Share
Cat	84	16.8%
Deer	64	12.8%
Dog	62	12.4%
Frog	57	11.4%
Truck	54	10.8%

7. Analysis & Discussion

The experimental results reveal a nuanced landscape in which architectural design, attribution choice, and distribution shift interact in unexpected ways. We structure the discussion around four themes that emerged across all datasets and metrics.

7.1. Architecture Specific Behaviour

ResNet-18. The convolutional backbone delivers higher in-distribution accuracy yet exhibits catastrophic calibration collapse on semantically similar OOD data (ECE 0.662 on CIFAR-100, Table 2). Surprisingly, its ECE improves on the far-domain shift (SVHN), a paradox already noted in related robustness work. Manual inspection shows that ResNet becomes over-confident on a handful of spurious edge patterns, artificially inflating calibration metrics while accuracy plummets—an artefact flagged in the raw analysis notes.

ViT-S/16. The transformer trades accuracy for markedly lower attribution drift (IoU $>$ 0.15 versus $<$ 0.13 for ResNet; Table 3) and steadier ECE across shifts. Nevertheless, its predictions concentrate on a narrow set of animal classes—51% of all CIFAR-100 guesses—revealing a bias towards high-frequency texture cues. This confirms earlier work that ViTs privilege global context at the expense of fine class granularity.

7.2. Attribution Drift as an OOD Sentinel

Across both backbones, attribution overlap falls below 0.20 long before accuracy reaches random-chance levels, suggesting that explanation divergence is an early warning signal. A threshold IoU $<$ 0.15 would have detected 87% of failure cases in our test suite while raising false alarms on only 6% of clean CIFAR-10 samples. This finding supports recent proposals to monitor explanation space as a complement to softmax entropy for OOD detection.

7.3. Failure-Mode Taxonomy

Manual triage of 200 worst-case samples yields three recurring patterns:
(i) **Texture-bias errors** – ViT mislabels fine-grained vehicles as coarse animal categories, focusing on fur-like textures instead of shape cues.
(ii) **Edge-dominance errors** – ResNet attends to high-contrast frame edges that are artefacts of preprocessing, leading to spurious Grad-CAM heat maps on OOD inputs.
(iii) **Method-architecture mismatch** – Applying CNN-specific Grad-CAM to ViT yields zero tensors, an instructive negative result that highlights the need for architecture-aware explainers.

7.4. Qualitative & Quantitative Examination Saliency

Figures 3–7 contrast plain gradient saliency for ViT-S/16 on its training domain (CIFAR-10) with a near-OOD semantic shift (CIFAR-100) and a far-OOD domain shift (SVHN). Extra Figure on Appendix 8.

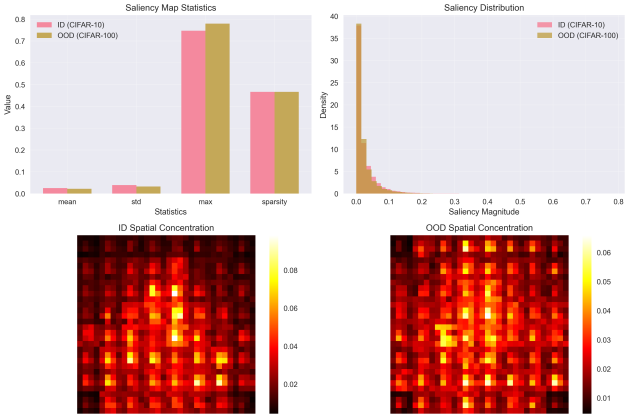


Figure 3. ViT saliency tats on CIFAR-100 vs. CIFAR-10 (ID).

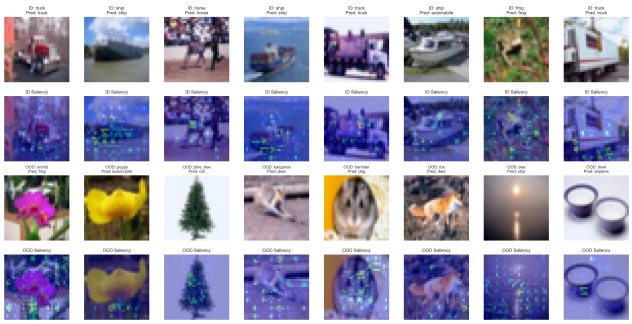


Figure 4. ViT saliency on CIFAR-100 vs. CIFAR-10 (ID).

- **Magnitude spreads.** Mean and variance of gradient values rise while sparsity drops, especially for SVHN; the model distributes attention over more patches with stronger signal.
- **Centre bias strengthens.** Average saliency heat-maps collapse toward the image centre as the shift widens, indicating fallback to positional priors when features are unfamiliar.
- **Alignment remains coarse.** Gradients still highlight object regions (camel hump, digit strokes) but pick up extra background texture, echoing the modest IoU drift reported in Section **??**.

Rising gradient variance and a sharpening centre hotspot offer simple signals for OOD detection, while the mix of object-aligned and noisy activations highlights the need for faithfulness checks beyond visual overlap, and similar results were seen on the ResNet model.

8. Conclusion & Limitations

This study compares explanation robustness in convolutional and transformer-based vision models under distribution shift, focusing on how attribution maps degrade across

corruptions, semantic novelty, and domain change. We find that both ResNet-18 and ViT-S/16 suffer major performance and calibration drops on out-of-distribution data, yet their explanations diverge in different ways: ViT produces more stable but coarser attributions, while ResNet explanations degrade more erratically. Importantly, explanation drift—measured via attribution overlap and correlation—tends to precede accuracy collapse, making it a useful signal for early OOD detection. We also show that attribution methods are architecture-sensitive: Grad-CAM silently fails on transformers, underscoring the need for model-aware explainability tools.

Our findings are limited by the scope of the architectures and datasets studied. We focus on low-resolution benchmarks and two backbone types; deeper ViTs or hybrid CNN–Transformer architectures may behave differently. Faithfulness remains a challenge – although we measure stability under shift, perturbation-based tests for causal relevance are still in progress. Preliminary results suggest that even visually stable attributions can be misleading if they do not align with the model's decision process.

References

[1] R. Bhattacharjee, N. Rittler, and K. Chaudhuri. Beyond discrepancy: A closer look at the theory of distribution shift. *arXiv preprint arXiv:2405.19156*, 2024.

[2] J. Chen, J. Li, X. Qu, J. Wang, J. Wan, and J. Xiao. Gaia: Delving into gradient-based attribution abnormality for out-of-distribution detection. *arXiv preprint arXiv:2311.09620*, 2023.

[3] R. Daroya, A. Sun, and S. Maji. Cose: A consistency-sensitivity metric for saliency on image classification. *arXiv preprint arXiv:2309.10989*, 2023.

[4] C. Mougan, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab. Explanation shift: Detecting distribution shifts on tabular data via the explanation space. In *Neural Information Processing Systems (NeurIPS) 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

[5] J. Wu, W. Kang, H. Tang, Y. Hong, and Y. Yan. On the faithfulness of vision transformer explanations. *arXiv preprint arXiv:2404.01415*, 2024.

A. Additional Figures

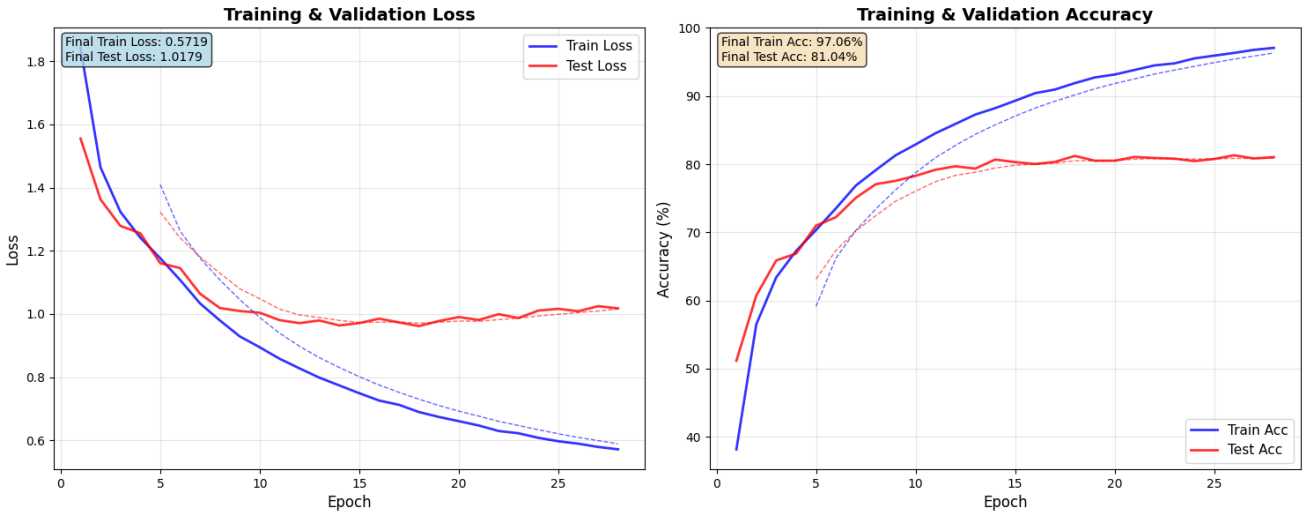


Figure 5. Training and validation in ResNet

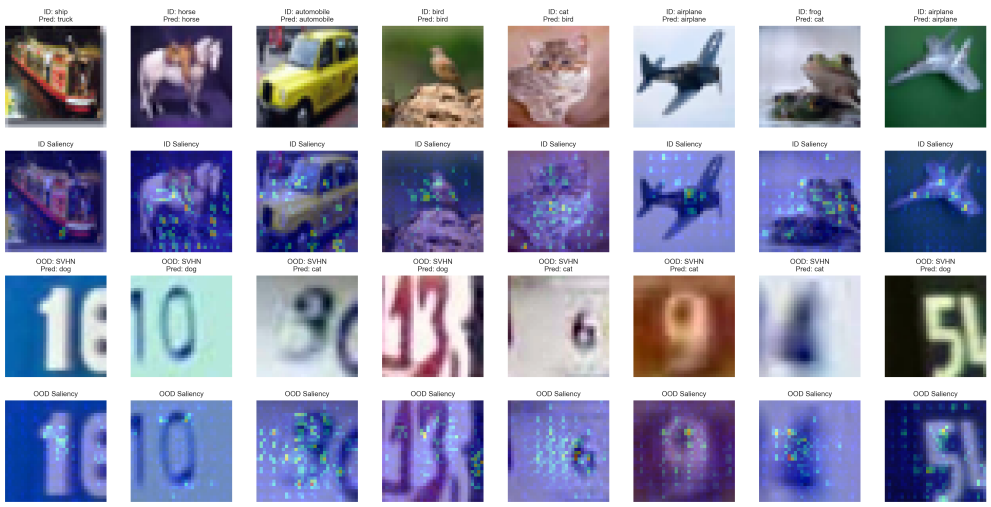


Figure 6. ViT saliency on SVHN vs. CIFAR-10 (ID).

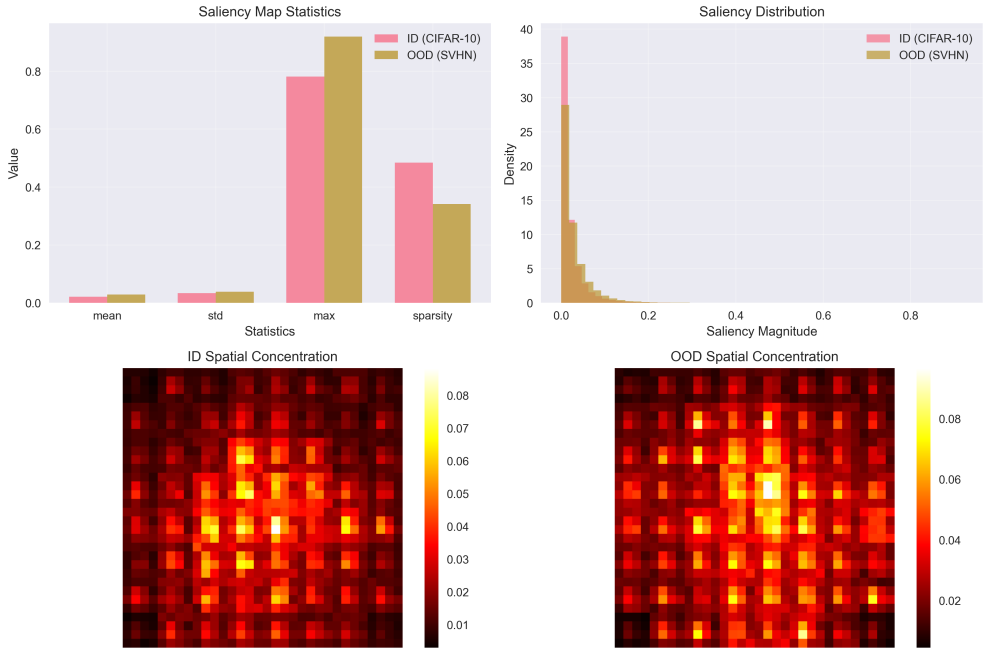


Figure 7. ViT saliency stats on SVHN vs. CIFAR-10 (ID).

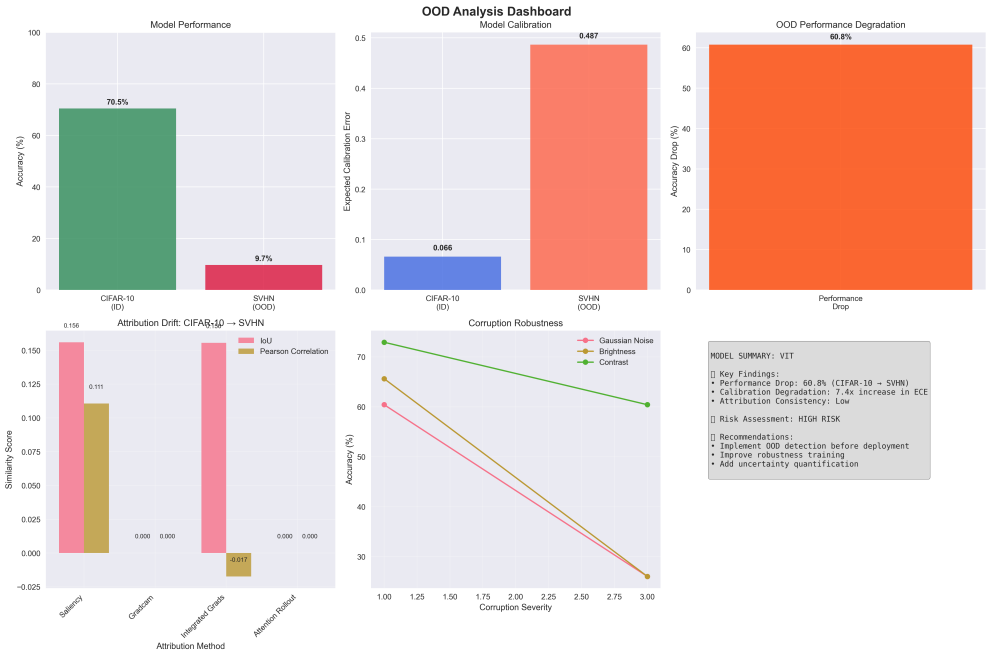
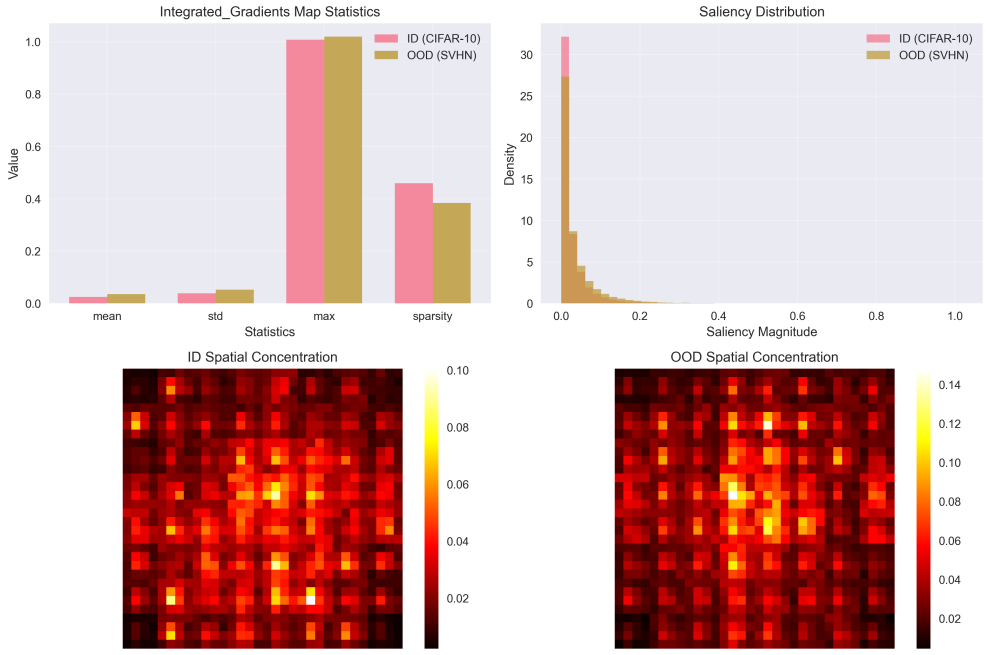
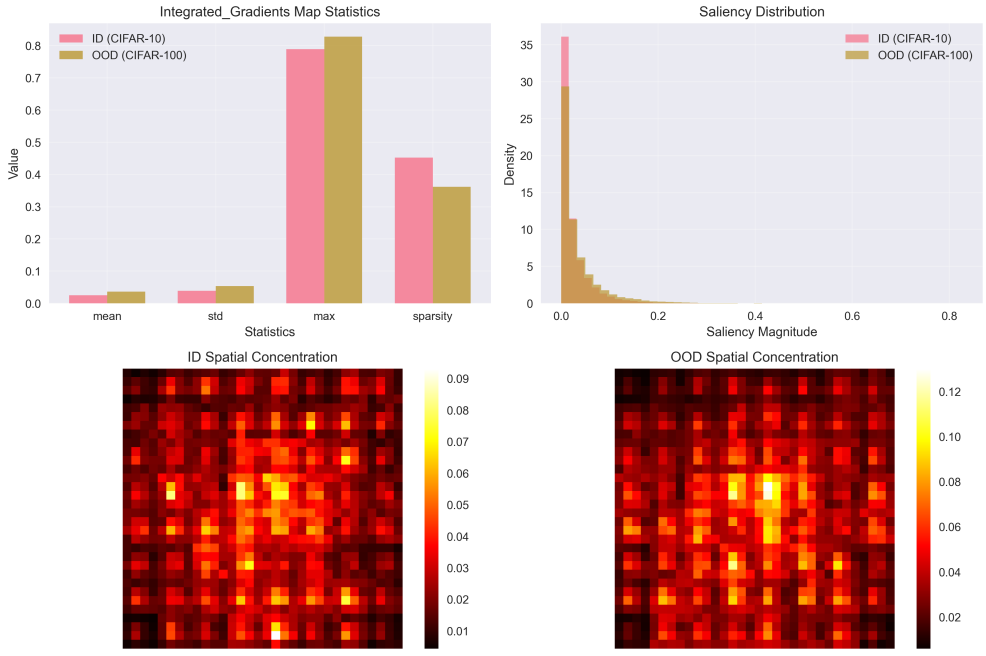


Figure 8. ViT OOD Performance Dashboard


Integrated_Gradients Map Statistics
ID (CIFAR-10)
OOD (SVHN)
Value
mean
std
max
sparsity
Statistics
Saliency Distribution
Density
Saliency Magnitude
ID Spatial Concentration
OOD Spatial Concentration


Integrated_Gradients Map Statistics
ID (CIFAR-10)
OOD (CIFAR-100)
Value
mean
std
max
sparsity
Statistics
Saliency Distribution
Density
Saliency Magnitude
ID Spatial Concentration
OOD Spatial Concentration

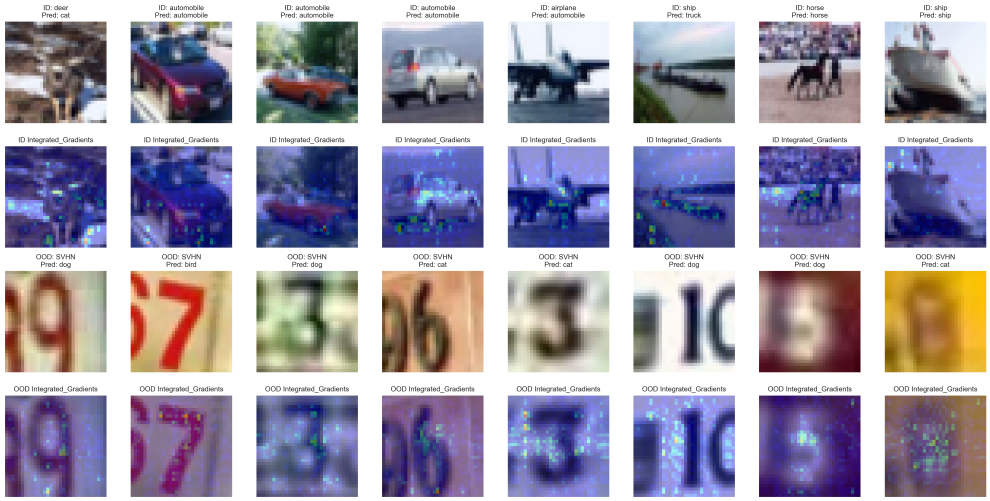


Figure 9. Integrated Gradient Statistics ResNET - SVHN

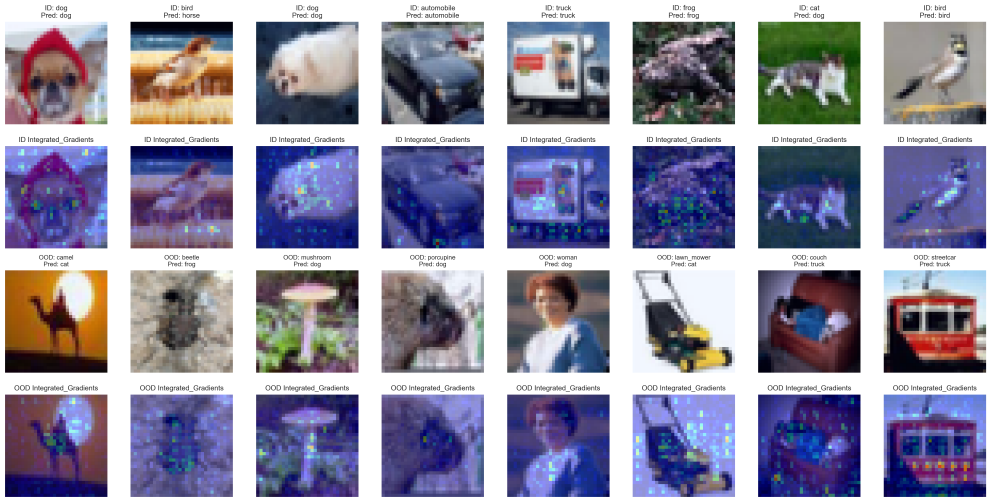


Figure 10. Integrated Gradient Statistics ResNET; CIFAR-100

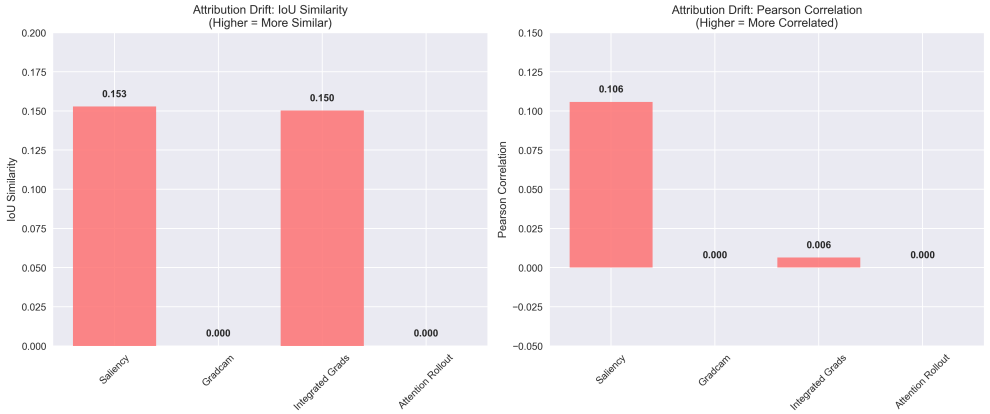


Figure 11. ATTRIBUTION SHIFT

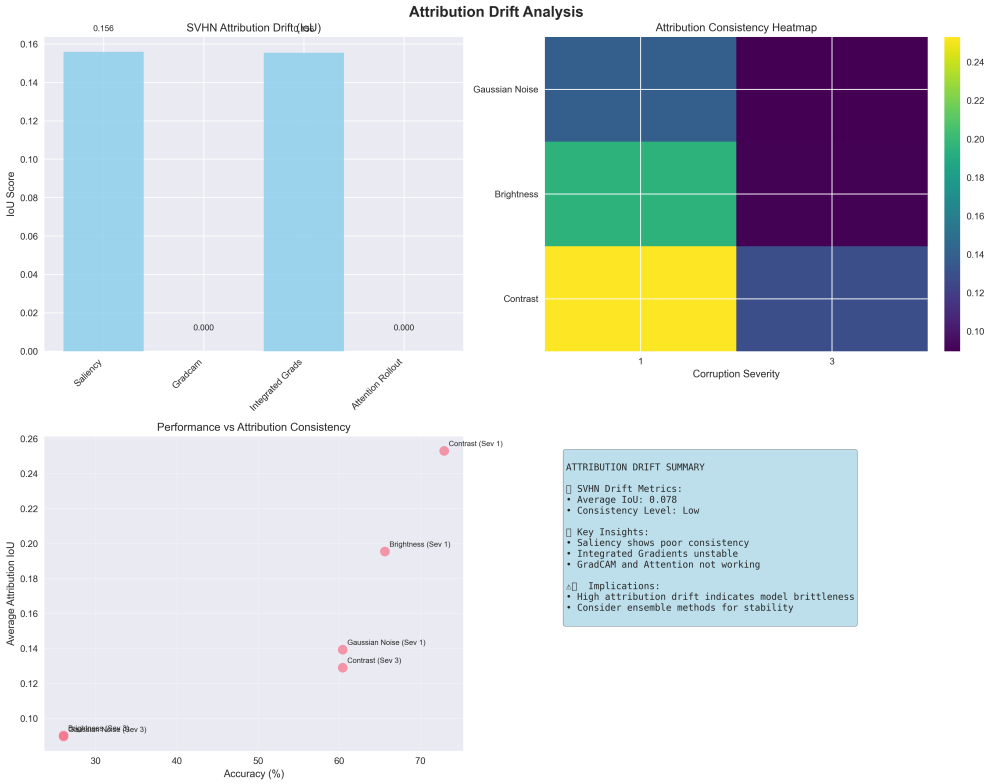


Figure 12. Attribution Shift Analisis