Practical Work Report: OOD Detection & Neural Collapse Analysis

Course: Deep Learning in Computer Vision

Student Name: Boyuan ZHANG, Huanshan HUANG

February 20, 2026

Abstract

In this practical work, we explore the mechanisms of Out-of-Distribution (OOD) detection and the phenomenon of Neural Collapse (NC) in Deep Learning. We trained a ResNet-18 backbone on the CIFAR-100 dataset and achieved a 78.99% test accuracy. Through regularization, we reveal that our model enters the state of partial collapse. While the model exhibits strong decision-level consistency of 97.6% for NC4 and feature-classifier alignment of 0.949 for NC3, the geometric rigidity is relaxed, resulting in a non-zero within-class variance of 0.0252 for NC1. We further evaluated how this partial collapse affects NECO, a novel OOD detection method, across datasets of SVHN and MNIST. In the model regularized with Dropout, the lack of NC5 orthogonality causes geometric detection to fail. Simple inputs like MNIST lack the complex textures needed to generate orthogonality, causing their features to embed almost entirely within the In-Distribution (ID) principal subspace. Our study then demonstrates that removing Dropout allows the model to reach an ideal terminal state with more ideal NC5 orthogonality. Under this regime, NECO proves effective, successfully differentiating both high-energy (SVHN) and simple (MNIST) OOD data with notably higher AUC scores.

1 Introduction

Deep neural networks have revolutionized Computer Vision (CV), yet they often suffer from overconfidence on OOD data [1]. Ensuring reliability in open-world settings requires robust mechanisms to distinguish ID from OOD samples. Recent theoretical advancements suggest that deep networks undergo NC at the Terminal Phase of Training (TPT) [2]. In this state, within-class variability vanishes and class means form an Equiangular Tight Frame (ETF), offering a geometric basis for robust classification.

The main objectives of this practical work are:

- To train a ResNet-18 classifier on CIFAR-100 with optimized regularization.
- To implement and compare standard OOD scores including Maximum Softmax Probability (MSP), Energy, Mahalanobis, ViM [3], and NC-based methods.
- To quantitatively analyze the NC1 to NC5 phenomena with a detailed breakdown of the geometric properties.
- To investigate the impact of regularization like Dropout [4] on the perfect collapse theory.
- To simulate and analyze a novel OOD detection method named NECO [5].

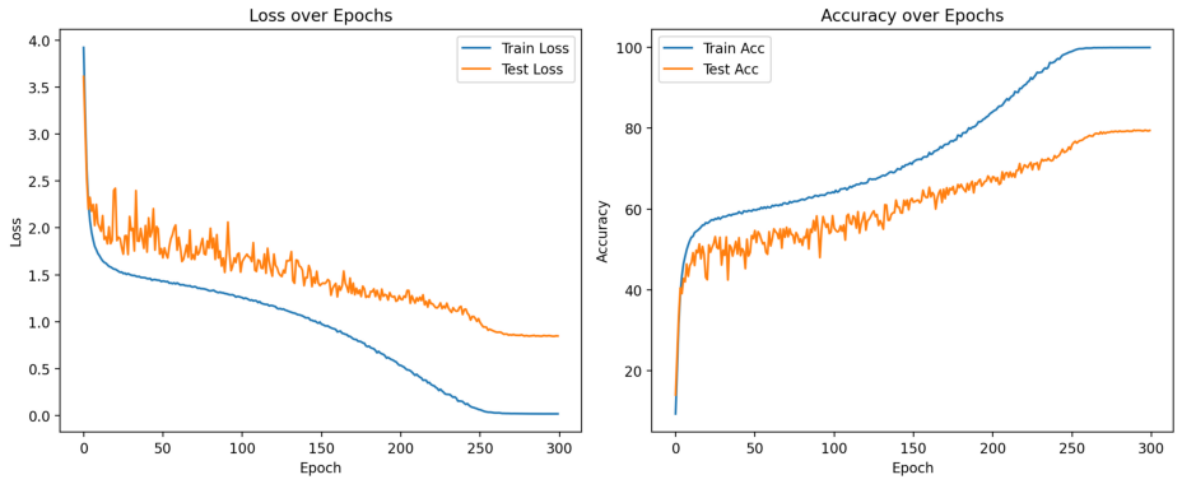


Figure 1: The Train curve.

2 Backbone Training

2.1 Experimental Setup

We trained a modified ResNet-18 architecture on the CIFAR-100 dataset. Specifically, to accommodate the 32x32 resolution of CIFAR-100 images, we replaced the standard 7x7 convolution (stride 2) and the initial max-pooling layer designed for ImageNet with a 3x3 convolution (stride 1). The training was conducted over 300 epochs using an initial learning rate of 5×10^{-2}, decayed by a weight decay of 1×10^{-3}. To enhance the generalization capability of our model, we introduced a dropout layer before the last layer (FC layer) of the standard ResNet-18 with a dropout rate of 0.3, as well as Stochastic Gradient Descent (SGD) with momentum 0.9 for the optimizer.

2.2 Training Results and Generalization

After 300 epochs, the model achieved a test accuracy of 78.99% while the training accuracy was 99.71%. The reduced train-test gap of approximately 21% confirms that our strong regularization strategy did not successfully control overfitting. While we tried multiple sets of hyper-parameters, we still see the gap, i.e., ResNet-18 always tends to overfit CIFAR-100, and we infer that the reason is that the parameter count of ResNet-18 is excessively large for the scale of the CIFAR-100 dataset. The gap between the high model capacity and the relatively limited data volume led to overfitting, and even strong regularization failed to fully bridge this gap. This provides a realistic partial collapse regime for analysis.

3 OOD Detection Methods

Before presenting the results, we briefly outline the OOD scoring mechanisms compared in this study:

- **MSP (Maximum Softmax Probability)**: Uses the maximum output probability as a confidence score. It serves as the baseline but is prone to overconfidence on OOD data. If the classifier assigns low maximum probability, it is unsure or the input may be OOD.

$$s_{max-prob}(x) = \max_c p(y = c|x) = \max_c softmax(z_c(x))$$

- **Max Logit**: Uses the maximum non-normalized logit directly, avoiding the compression effect of the Softmax function. Logits reflect raw model evidence before the softmax normalization, and avoid saturating effects of softmax.

$$s_{max-logit}(x) = \max_c \ z_c(x)$$

- **Energy Score**: Derives a scalar energy score from the logit distribution. It theoretically aligns better with the probability density of the inputs. Energy provides a scalar that correlates with the model's total evidence across classes. Lower energy (more negative) implies stronger evidence; higher energy (less negative or positive) can indicate OOD. Energy score derived from the logits: a common definition:

$$E(x) = -log(\sum_c e^{z_c(x)}) = -LSE(z(x))$$

With temperature $T > 0$ one can use:

$$E_T(x) = -T * log(\sum_c e^{z_c(x)/T})$$

- **Mahalanobis Distance**: A geometry-based method that calculates the distance of a sample to the nearest class centroid in the feature space, weighted by the inverse covariance matrix. For each class c, compute the class mean in feature space:

$$\mu_c = 1/N_c \sum_{i:y_i=c} f(x_i)$$

Then, estimate a shared covariance matrix Σ (or per-class covariance), typically the empirical covariance of features across the training set. Therefore, the Mahalanobis score calculated as:

$$d_{\text{Maha}}(x) = \min_c (f(x) - \mu_c)^\top \Sigma^{-1} (f(x) - \mu_c)$$

For each test sample, calculate the distance to all 100 class centers and take the minimum distance as the OOD score (the smaller the distance, the more similar to an in-distribution sample).

- **ViM(Virtual-logit Matching)** [3]:

$$s_{ViM}(x) = \alpha \cdot ||P_{residual}(f(x) - \mu)||_2 - \log \sum_{c=1}^{C} e^{z_c(x)},$$

where

$$\alpha = \frac{\sum_i \max_c z_c(x_i)}{\sum_i ||x_i^\perp||_2}$$

- **NECO**: A novel metric combining feature subspace geometry projection with prediction confidence:

$$NECO(x) = \frac{||P\ h_\omega(x)||}{||h_\omega(x)||} = \frac{\sqrt{h_\omega(x)^\top P P^\top h_\omega(x)}}{\sqrt{h_\omega(x)^\top h_\omega(x)}}$$

4 Neural Collapse Analysis

We investigated the geometric properties of the network at the Terminal Phase of Training (TPT). By systematically evaluating the five properties of Neural Collapse (NC), our analysis reveals a state of partial collapse, heavily influenced by the training dynamics and regularization strategies.

4.1 NC1: Variability Collapse

The NC1 property postulates that the within-class covariance converges to zero as training progresses. To evaluate this, we present two visualizations. The first is a t-SNE projection mapping high-dimensional features into a 2D space. The second is a statistical summary comprising a bar chart of variance per class and a histogram showing the distribution of these variances across all classes, where a red dashed line indicates the global mean baseline.

Observing the t-SNE plot, the feature clusters for different classes are distinguishable but remain somewhat dispersed rather than collapsing into singular, infinitely dense point-masses. This dispersion is quantitatively confirmed by the variance histogram, which shows a tightly grouped distribution peaking around a global mean of 0.0252, rather than approaching the perfect theoretical collapse threshold of 0.

This phenomenon indicates a state of partial NC1. The lack of perfect geometric collapse is a direct consequence of the strong Dropout layer set at a rate of 0.3. This regularization technique continuously injects noise during training, intentionally preventing the features from completely collapsing to single points in order to preserve the functional generalization capability of the model on unseen data. However, we note that most of the points on the picture are clustered clearly, which shows that our method primarily validates NC1. Although t-SNE visualizations that do not strictly preserve distances show dispersion, this lack of perfect collapse is rigorously confirmed by our quantitative variance measurements.

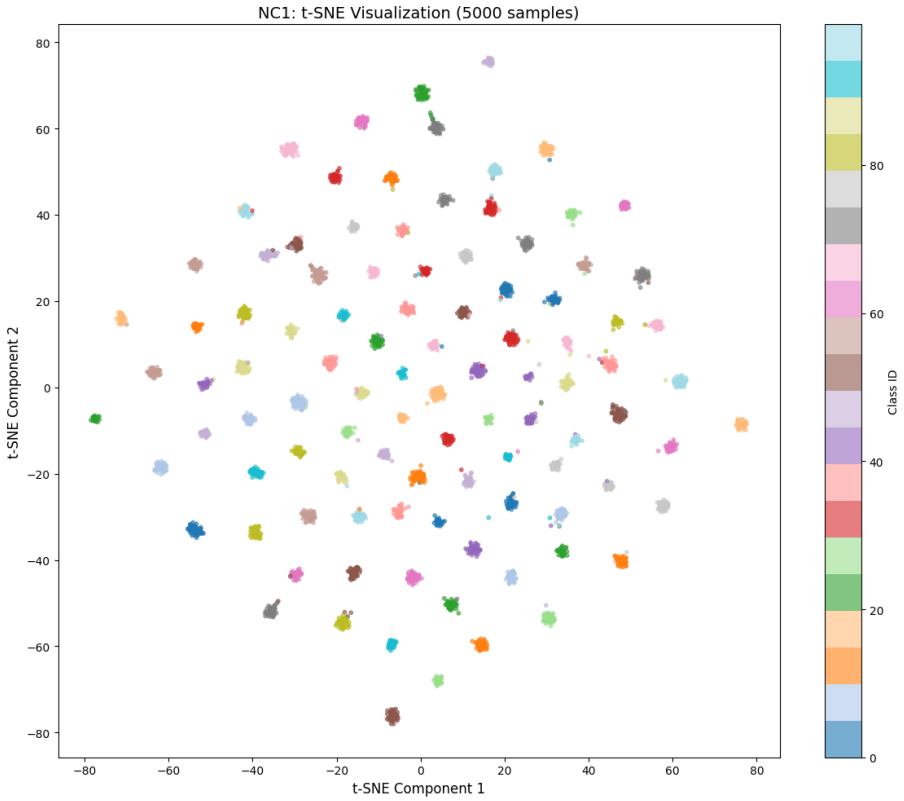


Figure 2: NC1: t-SNE Visualization of the feature space.

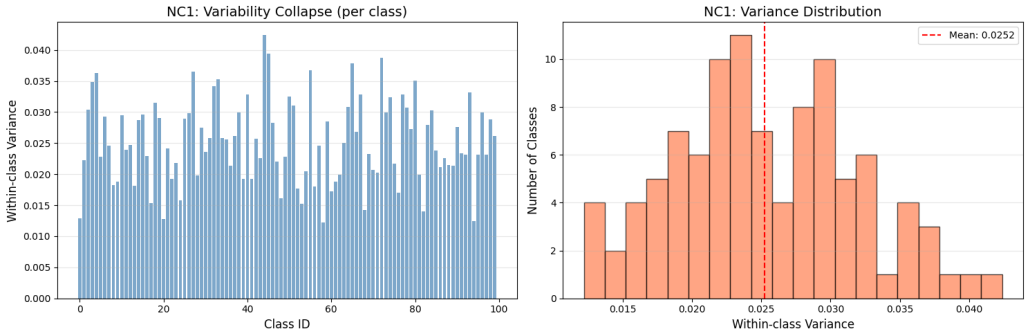


Figure 3: NC1: Within-class variance distribution showing a global mean at 0.0252.

4.2 NC2: Convergence to Simplex ETF

NC2 states that centered class means should organize themselves into an Equiangular Tight Frame (ETF), a highly symmetric geometric structure. To illustrate the debugging process of this metric, we provide two sets of plots showing pairwise cosine similarity distributions, pairwise angle distributions, and a cross-class cosine similarity heatmap. The red dashed lines represent the theoretical ETF standards, while the yellow dashed lines mark our empirical means.

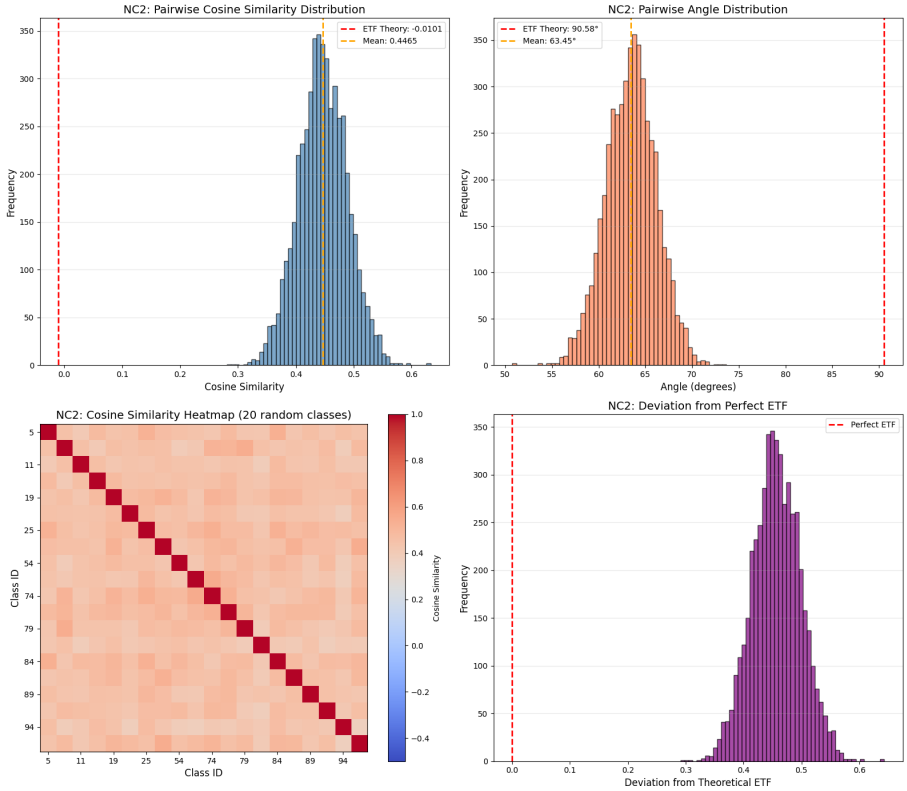


Figure 4: NC2: Uncentered ETF structure showing incorrect positive similarities (red off-diagonal).

In the uncentered plots, a clear anomaly is visible. The heatmap shows entirely red off-diagonal elements, indicating mathematically positive similarities across the board. Correspondingly, the mean cosine similarity is 0.4465 and the mean angle sits around 63.45 degrees, both failing to match the ETF theory. However, after applying global mean centering in the second set of plots, the off-diagonal elements of the heatmap correctly turn blue, signifying values near zero

or slightly negative. The empirical mean cosine similarity aligns with the theoretical baseline of -0.0101, and the mean angle correctly shifts to approximately 90 degrees.

This stark contrast highlights the mathematical necessity of removing the global feature bias before assessing ETF properties. Furthermore, even in the properly centered plots, the angle and similarity distributions exhibit a wide spread. This spread indicates a weak NC2 state, meaning that while the average geometry follows the simplex arrangement, the individual class means are not locked into a perfectly rigid crystal structure, which aligns with the noise introduced by the partial NC1 state.

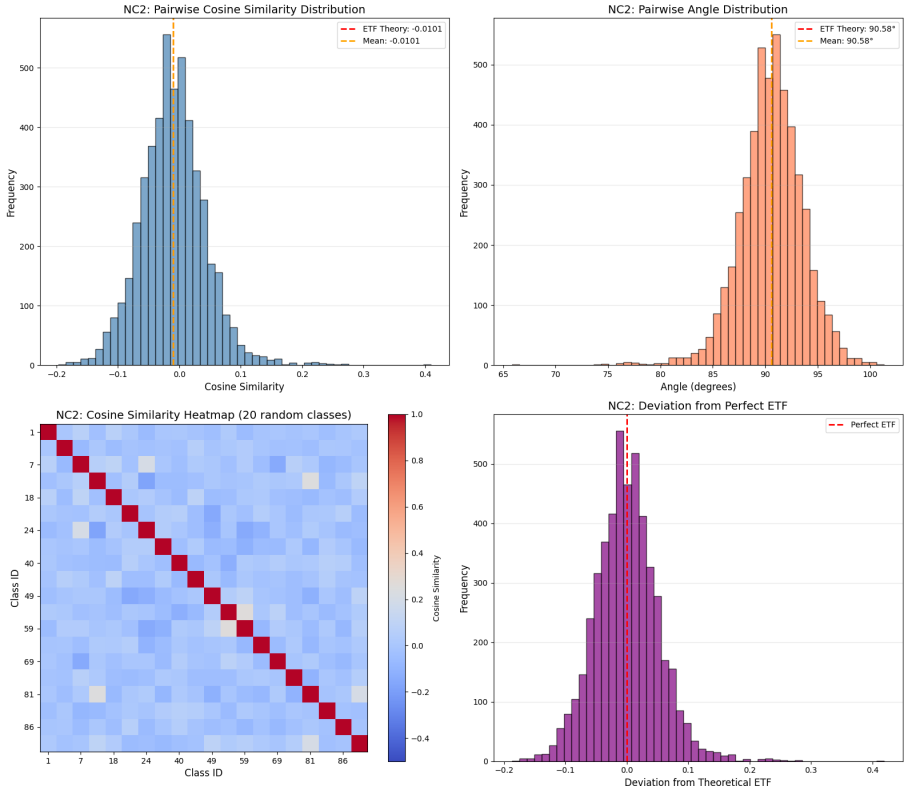


Figure 5: NC2: Corrected ETF structure after global mean centering (blue off-diagonal).

4.3 NC3: Self-Duality

NC3 evaluates the alignment between the empirical class means and the linear classifier weights. The visualization includes a dimensional alignment scatter plot against a perfect y equals x baseline, a histogram of Pearson correlations with a threshold for strong alignment, an angular deviation histogram, and a cross-correlation heatmap where the diagonal represents matching classes.

The data reveals a stark contrast between two different mathematical measurements. The Pearson correlation histogram is densely packed near the high end, yielding a global mean of 0.949, and the heatmap displays a sharp dark green diagonal representing near-perfect correlation. However, the angular deviation histogram tells a different story, showing a prominent peak around 45.3 degrees, which is far from the theoretical ideal of zero degrees.

This discrepancy is a known artifact of using Rectified Linear Unit activations, which push all feature vectors into a positive mathematical orthant, creating a massive global bias. Because the angular measurement relies on raw vectors, it measures the gap caused by this bias. In contrast, Pearson correlation intrinsically performs mean centering, revealing that the directional fluctuation of the features and the weights are highly synchronized. Therefore, once the non-informative global shift is discounted, the model demonstrates a strong NC3, proving that the

decision layer successfully targets the class prototypes.

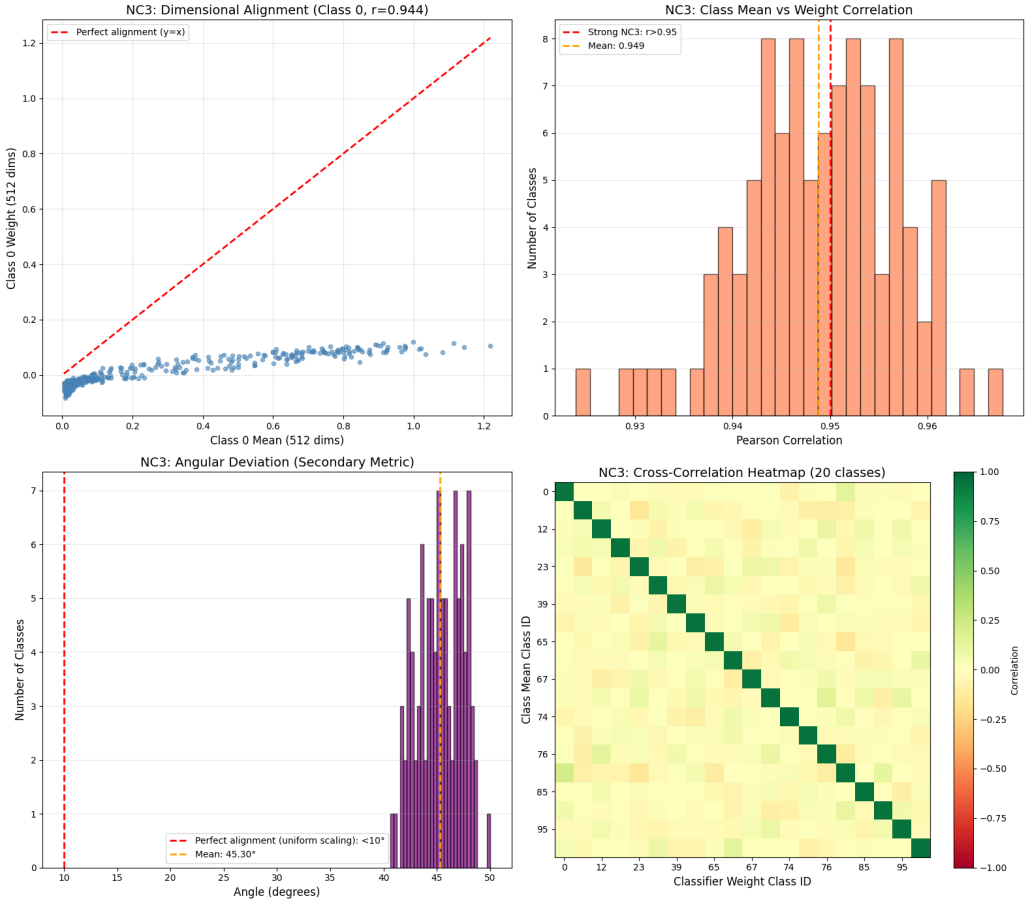


Figure 6: NC3: Alignment analysis showing high correlation despite angular bias.

4.4 NC4: Nearest Class Center Consistency

NC4 tests whether the complex neural network ultimately simplifies to a Nearest Class Center (NCC) decision rule. The figures display a per-class consistency bar chart with a red threshold line set at 95 percent, a distribution histogram of these consistencies, and a mismatch heatmap tracking true labels against model predictions.

Observing the metrics, the histogram is heavily skewed toward the right, indicating that the vast majority of samples align perfectly with the NCC rule. The mismatch heatmap is overwhelmingly white, containing only a few scattered light-red squares, which demonstrates an absence of systematic misclassification between specific class pairs. The overall mean consistency reaches a high value of 97.6 percent.

This strong consistency signifies a state of decision dominance. It proves that despite the loose geometric structure observed in the earlier variability and ETF analyses, the functional decision boundaries of the network have successfully simplified to the Voronoi tessellation of the class centroids, validating the use of distance-based metrics for inference.

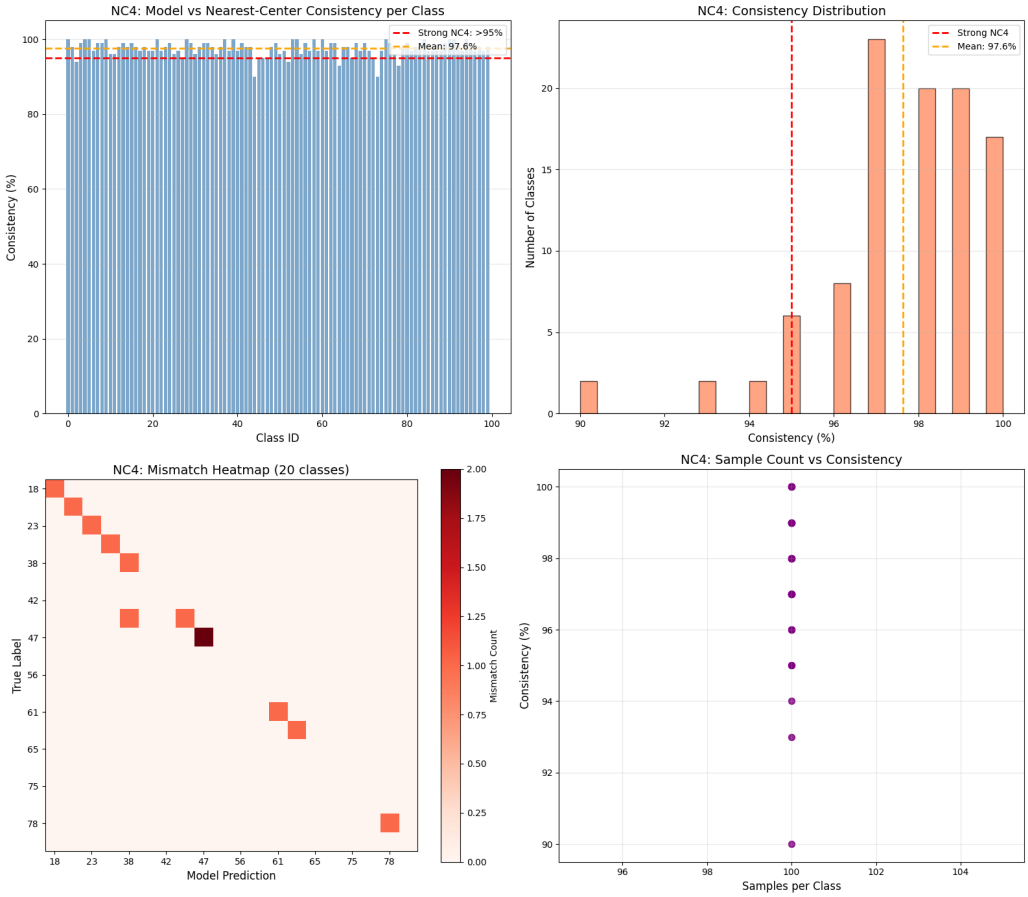


Figure 7: NC4: High consistency between softmax predictions and the NCC rule.

4.5 NC5: ID/OOD Orthogonality

NC5 postulates that Out-of-Distribution (OOD) data should ideally project into an orthogonal subspace relative to the In-Distribution (ID) data. The multi-panel figure contrasts CIFAR-100 against SVHN and MNIST, utilizing t-SNE projections, distance distribution histograms with vertical lines marking the distribution means, and feature norm density plots.

The visualizations depict a massive overlap between the datasets. In the t-SNE space, some OOD points intermingle with the ID points. The distance histograms reveal that the separation is relatively low, yielding a distance ratio of 1.56x for SVHN but a marginally lower 1.03x for MNIST. The feature norm distributions identically exhibit extensive intersection, failing to provide a clear threshold for separation for MNIST.

More specifically, we find the result interesting:

- **SVHN vs ID: High Distance and High Norm.** As shown in plot 3, this indicates that SVHN images, being rich in texture and color, induce strong activations in the ResNet model, which are similar in magnitude to the ID samples. Despite the close norm, SVHN samples are located further away from the ID class centers, as shown in plot 2. This confirms the NC5 orthogonality hypothesis. Since h_ω is similar for both groups, the increased Euclidean distance for SVHN is simply driven by directional misalignment. SVHN samples lie on the same high-dimensional hypersphere as ID data but fall into the space between class clusters. This indicates that our model has learned a certain semantic subspace for CIFAR-100, and while SVHN activates the network strongly, its feature vectors are orthogonal to the learned class directions.

- **MNIST vs ID: Low Distance and High Norm**. Plot 6 reveals that MNIST sam-

ples generate a substantial feature norm peaking around 7.5, close to the ID peak of 8.5. However, Plot 5 shows their distance to the nearest class center remains low. We infer that this is because MNIST lacks complex, while its high contrast edges strongly activate the neurons, but the resulting feature vectors are almost embedded within the learned ID subspace. Unseen textures that would generate activations in orthogonal directions, its features align perfectly with existing CIFAR-100 dimensions. Consequently, geometric checks based on distance seems to fail on MNIST, because it acts as a high-energy projection entirely subsumed within the ID subspace.

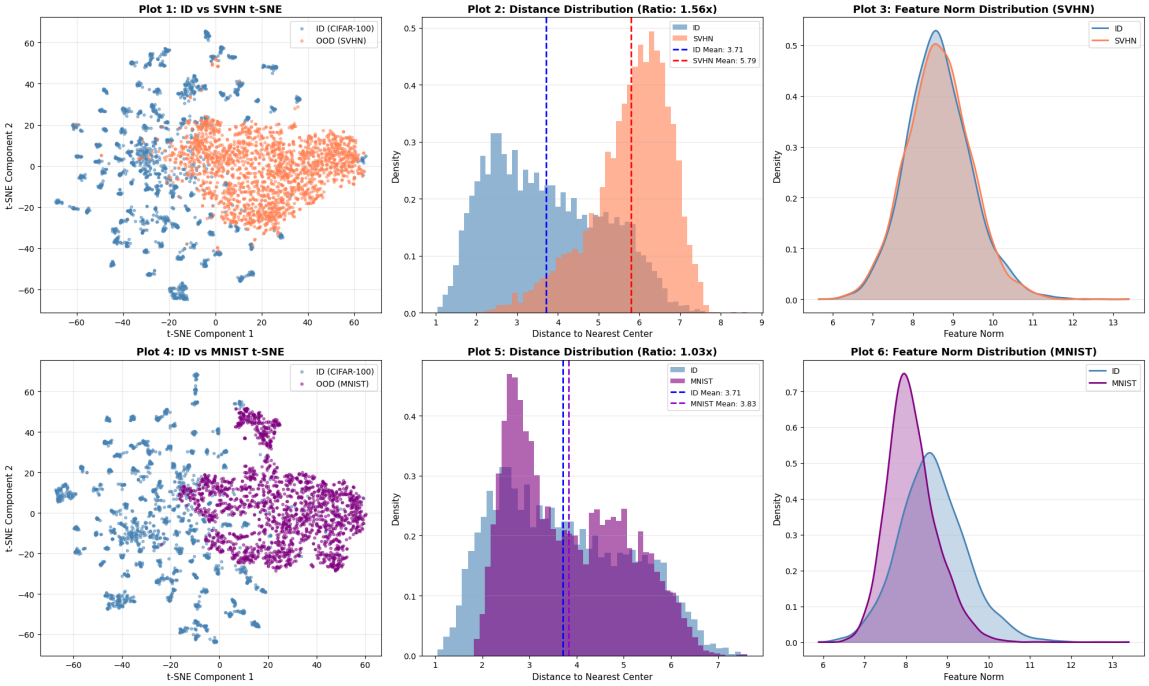


Figure 8: NC5: Comparison of SVHN and MNIST showing overlapping geometric distributions.

5 NECO Method

The NECO method [5] utilizes the orthogonality between OOD and ID data as indicated by NC5. The score is calculated using the following formula:

$$NECO(x) = \frac{||P\ h_\omega(x)||}{||h_\omega(x)||} = \frac{\sqrt{h_\omega(x)^\top P P^\top h_\omega(x)}}{\sqrt{h_\omega(x)^\top h_\omega(x)}}$$

We observe that while this geometric metric effectively detects SVHN, it struggles significantly with MNIST, as visualized in the NECO component analysis plots Fig.9. Detection performance was quantified using AUC:

- **SVHN** (AUC = 0.7834): The pure geometric NECO score works reasonably well because SVHN is a OOD type with high energy and orthogonal direction.
- **MNIST** (AUC = 0.6069): The score performs poorly, barely better than random guessing (0.5). This confirms that NECO, a geometric approach, is insufficient for collapsing low-complexity OOD datasets like MNIST.

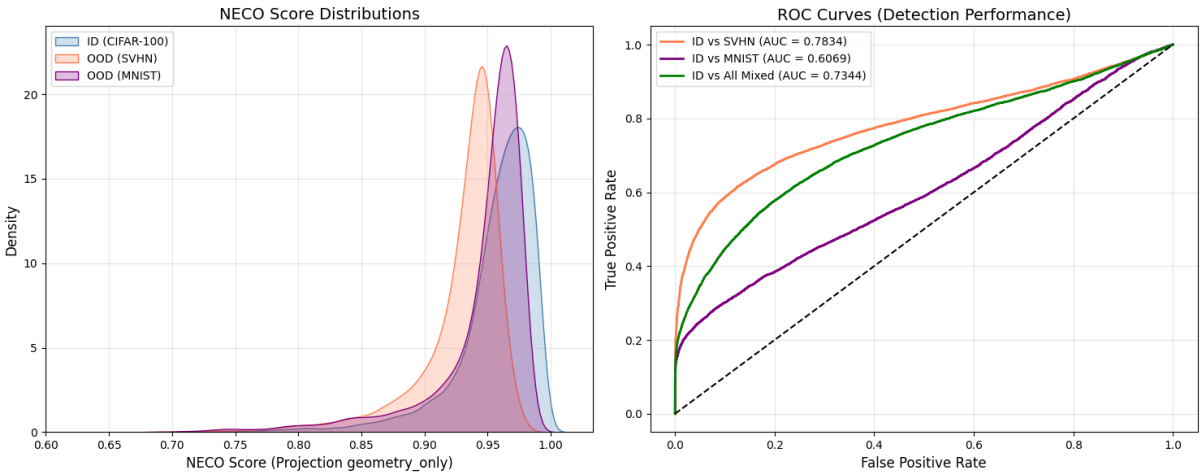


Figure 9: NECO scores and ROC curves.

The comparison analysis in Fig.10 consists of the following parts:

- **Geometric analysis.** From plot 2 in Fig.10, we observe that MNIST samples in purple cluster tightly in the high ratio region, mixing with ID samples in blue. Plot 3 further confirms this: the projection ratio distribution for MNIST almost overlaps with the ID distribution, near 1.0. This implies that $||Ph_\omega(x)|| \approx ||h_\omega(x)||$. SVHN in orange shifts to the left, showing that the standard NC5 orthogonality hypothesis where high-energy OOD data possess orthogonal components filtered out by the ID projection. Like we analyzed in NC5, NECO fails on MNIST due to subspace embedding. MNIST images possess simple, rigid structures without the complex background noise of natural images. As a result, their features span a much lower dimension space that is mostly subsumed by the dominant principal components of the CIFAR-100 subspace. Without any energy spilling over into the orthogonal residual space, the pure geometric projection metric is fundamentally blind to them.

- **Correlation analysis.** In plot 4, while MNIST lies on the diagonal line, looking closely at the magnitude revealed on the axes shows that MNIST samples are clustered closer to the lower left corner. They exhibit a lower norm relative to the extreme tails of SVHN and ID samples, though they still remain far from the origin. SVHN and ID samples extend much further out to higher norm values. Furthermore, the heatmap in plot 5 shows a strong correlation between MaxLogit and Norm. This suggests that the energy property may carry the critical discriminative information that the purely geometric NECO method misses.

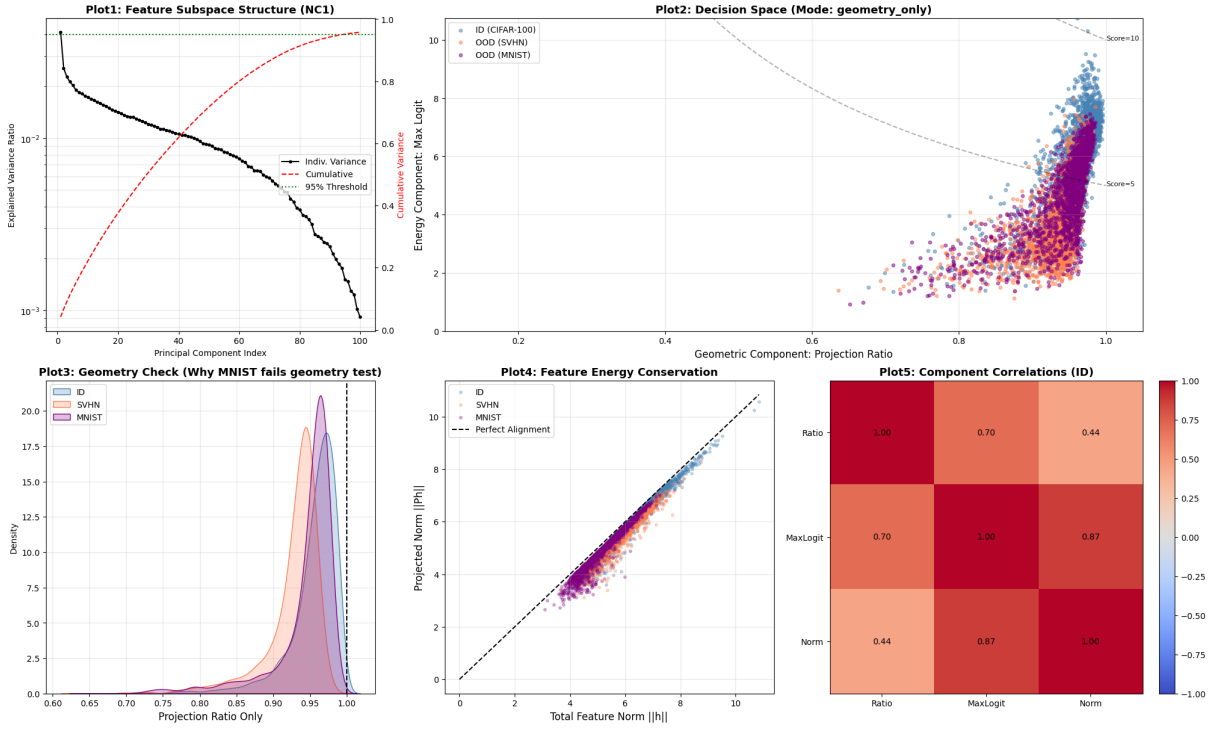


Figure 10: NECO: Comparison of SVHN vs MNIST.

The weak NC5 signal is speculated to be due to the tight feature space of CIFAR-100 with 100 classes.

6 Explorations

6.1 Representation Bottleneck

In our early exploratory phase, before adopting the 3×3 convolution mentioned in Section 2, we directly applied the standard ResNet-18. At the beginning, we found that no matter how strong the regularization is, the model consistently underfits, while the train accuracy is stabilized around 75% and the test accuracy is stabilized around 60%. After lots of efforts, we located the problem, that is the input image size of CIFAR-100 is 32×32, while the ResNet-18 we use, defined in torchvision, is designed for ImageNet-1K, of which each image is 224×224. The first layer (7×7, stride=2) and subsequent maxpooling layer of the original ResNet were designed for bigger images of 224×224 to rapidly downsample and expand the receptive field. CIFAR-100 images are only 32×32, so directly applying $7 \times 7 + stride2 + maxpooling$ would compress the spatial dimensions excessively, causing information loss and hindering the learning of fine-grained features. We attach the results under this case in appendix A. An interesting phenomenon is that under this case the model seems to more reliable in distinguishing OOD samples that are from MNIST.

6.2 Model Without Dropout

The original ResNet-18 does not have dropout layers. Motivated by the idea that without dropout, the model loses generalization ability and thus NC5 being more easily to appear, we removed the dropout layer and get a rather ideal result, in which the OOD phenomenon is more significant. In NC5 analysis, SVHN and MNIST show similar behaviors not only in distance

distribution but also feature norm distribution, and in NECO, higher AUC scores are obtained. The results are shown in appendix B.

7 Conclusion

In this work, we demonstrated that NC is not a binary state but a spectrum influenced by training dynamics. Our ResNet-18 model on CIFAR-100 exhibits partial neural collapse. Heavy regularization with a Dropout rate of 0.3, intentionally prevents ideal geometric compression, resulting in a non-zero NC1 variance of 0.0252 and a weak NC2 ETF structure. Despite this geometric relaxation, the network achieves robust functional collapse at the decision level, evidenced by a bias-corrected NC3 weight-feature alignment of 0.949 and a 97.6% NC4 Nearest Class Center consistency.

Furthermore, our experiments demonstrate that the potential limitations of NECO, a geometric OOD detection method. On simple datasets, it can be influenced by regularization. In the model regularized with Dropout, simple inputs like MNIST do not collapse toward the origin. Their features almost embed within the ID principal subspace. Lacking the complex textures necessary to generate orthogonal components, they fail to separate from the ID data geometrically. However, removing Dropout allows the model to reach a rather ideal terminal state where the NC5 orthogonality property becomes more significant. Under this regime, NECO proves effective, successfully differentiating both high-energy (SVHN) and simple (MNIST) OOD data with notably higher AUC scores. This illustrates a trade-off: while strong regularization is necessary for model generalization, it potentially disrupts the rigid geometric subspace structures required for projection-based OOD detection.

References

[1] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[2] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117(40):24652–24663, 2020.

[3] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022.

[4] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[5] Mouin Ben Ammar, Nacim Belkhir, Sebastien Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

A Representation Bottleneck

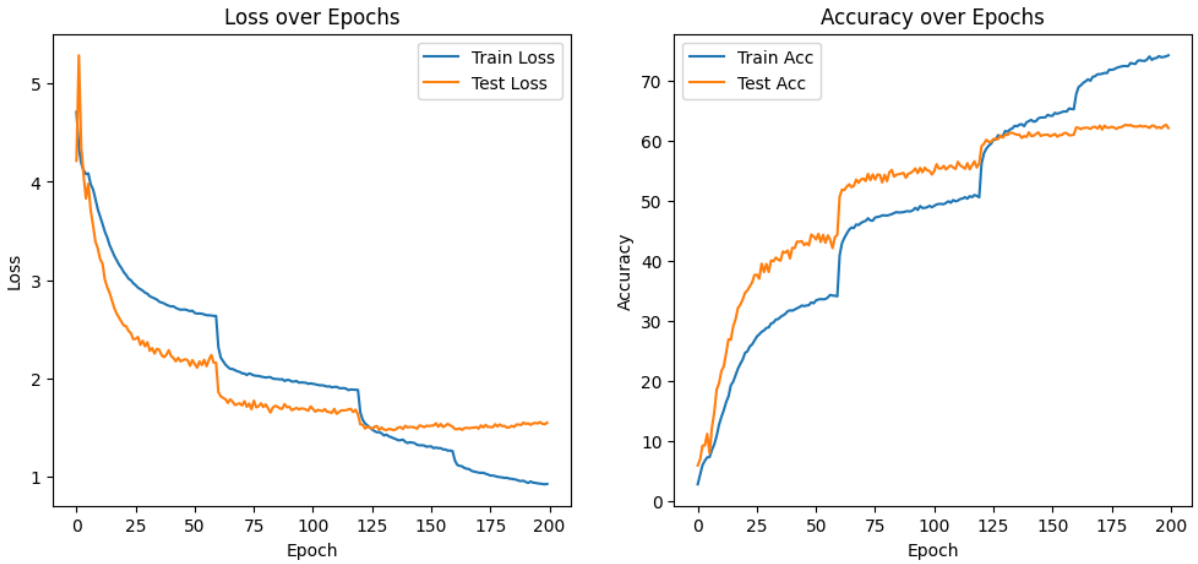
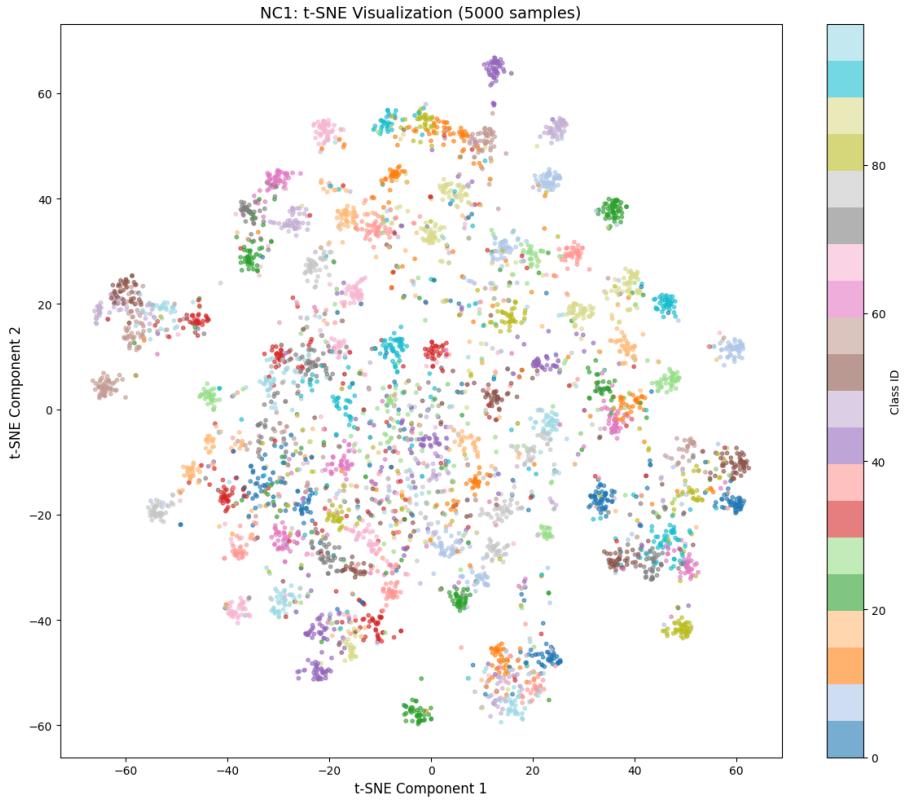
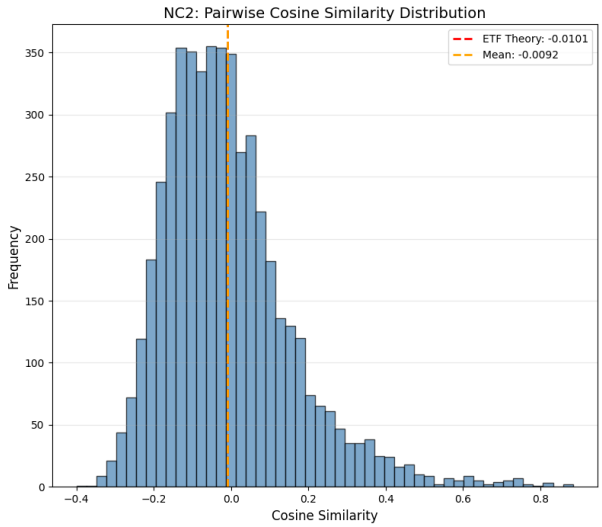
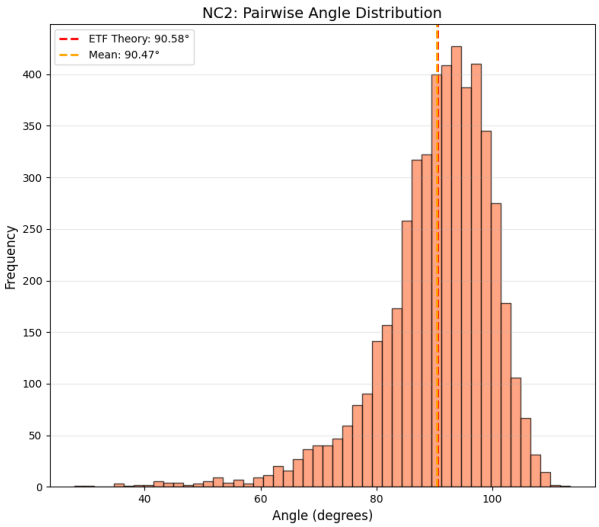
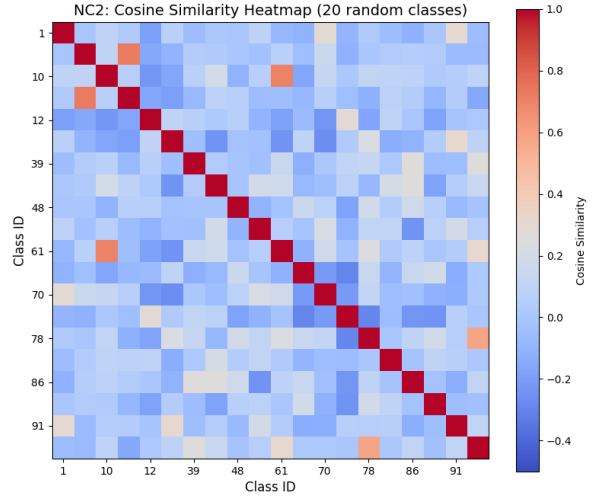


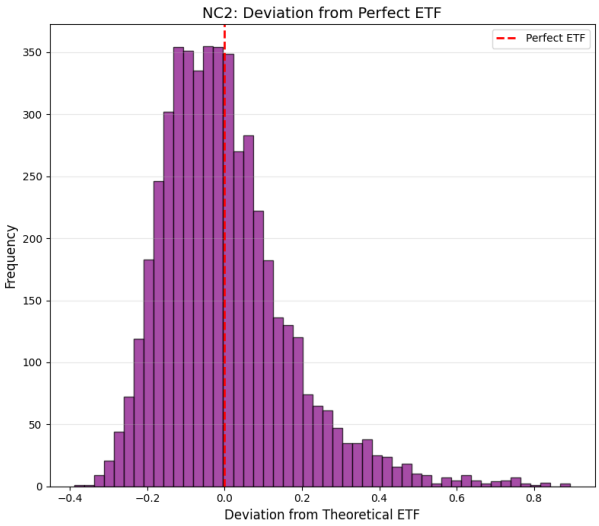
Figure 11: Train curve of underfitted model.

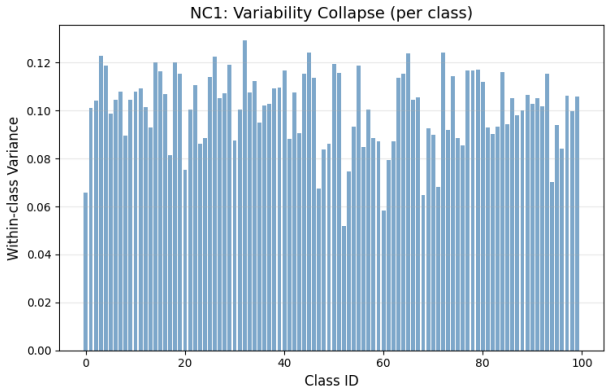


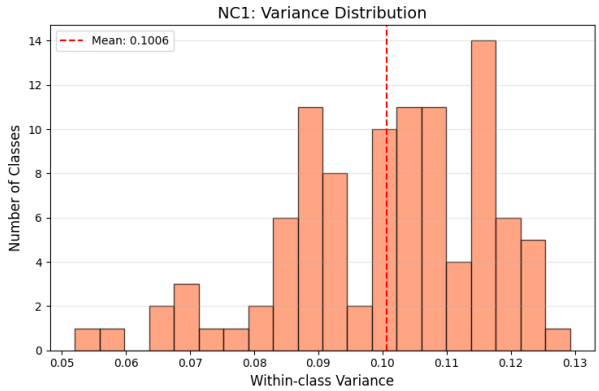

NC2: Pairwise Cosine Similarity Distribution
ETF Theory: -0.0101
Mean: -0.0092
Frequency
Cosine Similarity

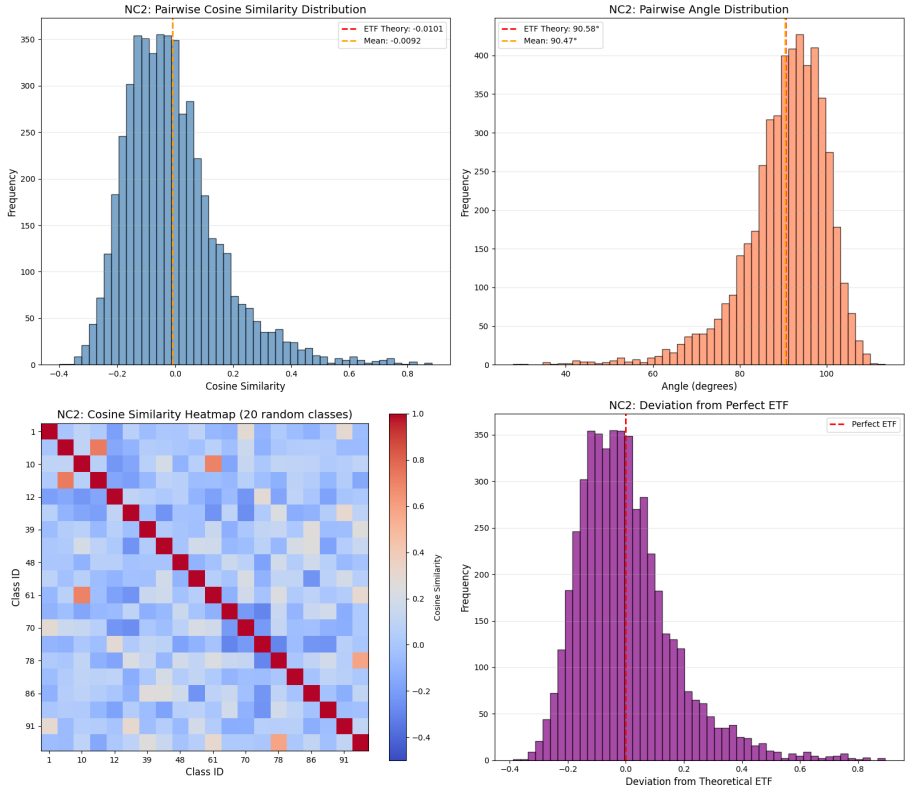

NC2: Pairwise Angle Distribution
ETF Theory: 90.58°
Mean: 90.47°
Frequency
Angle (degrees)

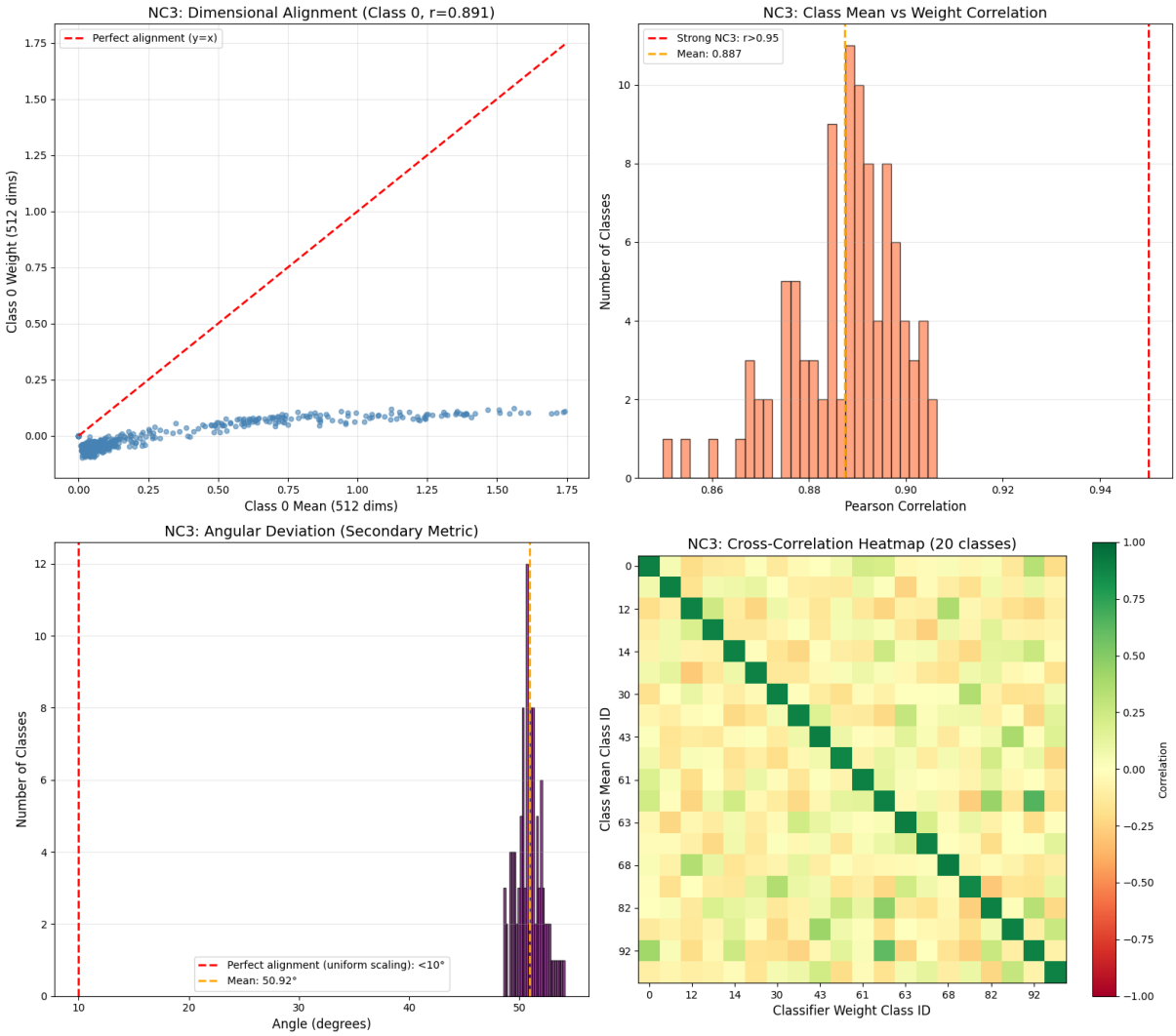

NC2: Cosine Similarity Heatmap (20 random classes)
Class ID
Cosine Similarity

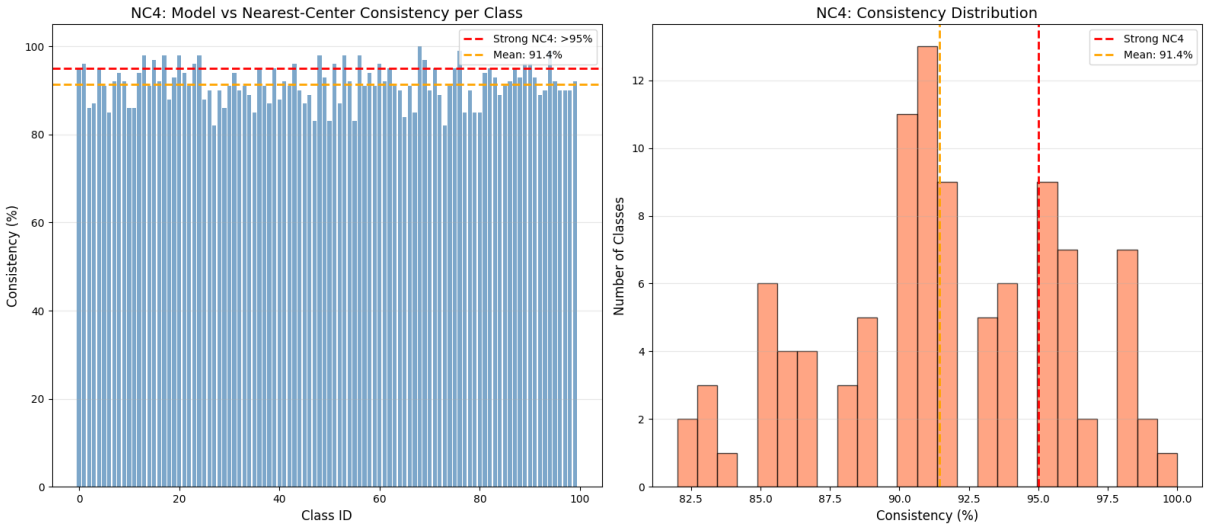

NC2: Deviation from Perfect ETF
Perfect ETF
Frequency
Deviation from Theoretical ETF

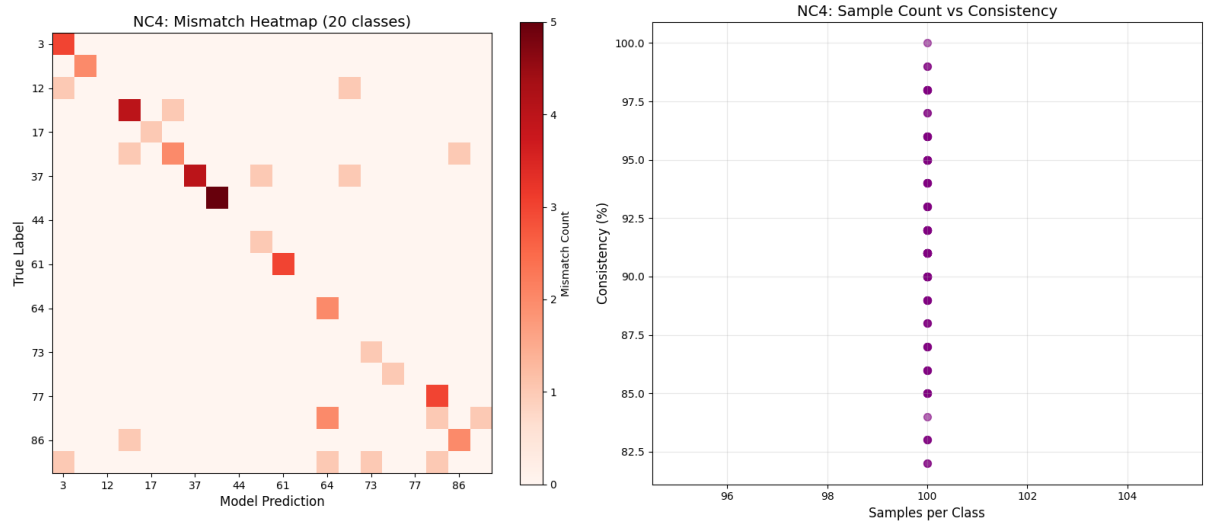

NC1: Variability Collapse (per class)
Within-class Variance
Class ID

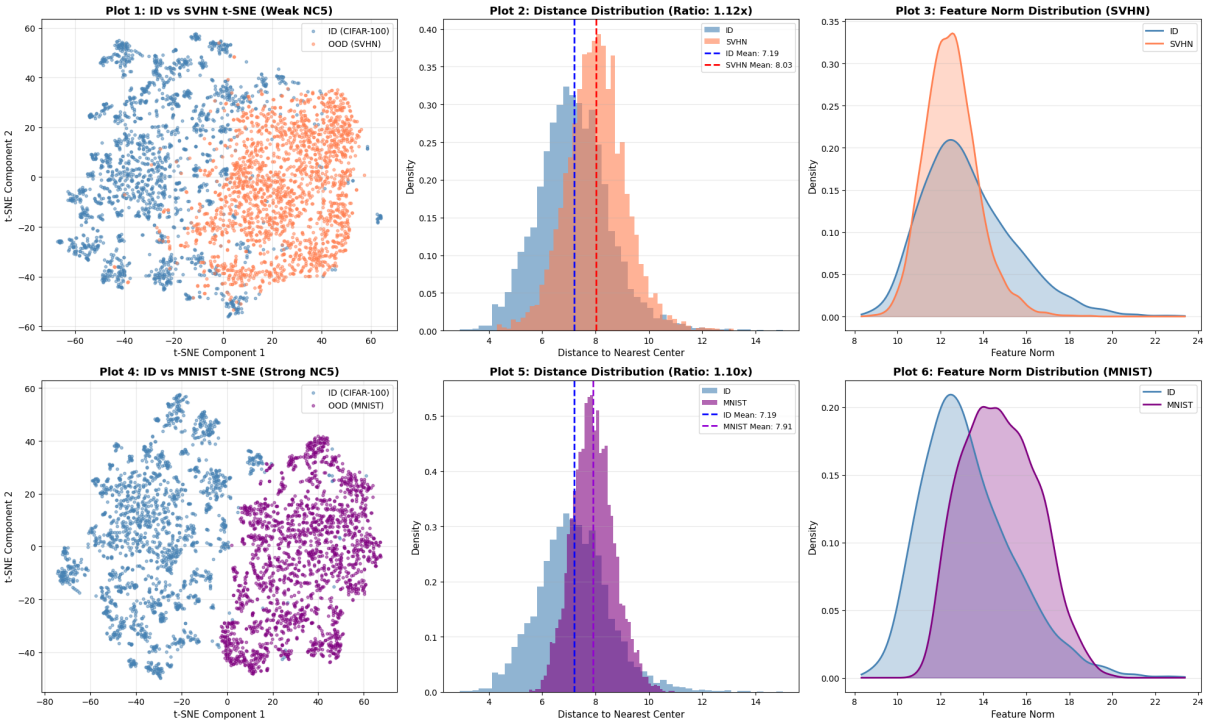

NC1: Variance Distribution
Mean: 0.1006
Number of Classes
Within-class Variance

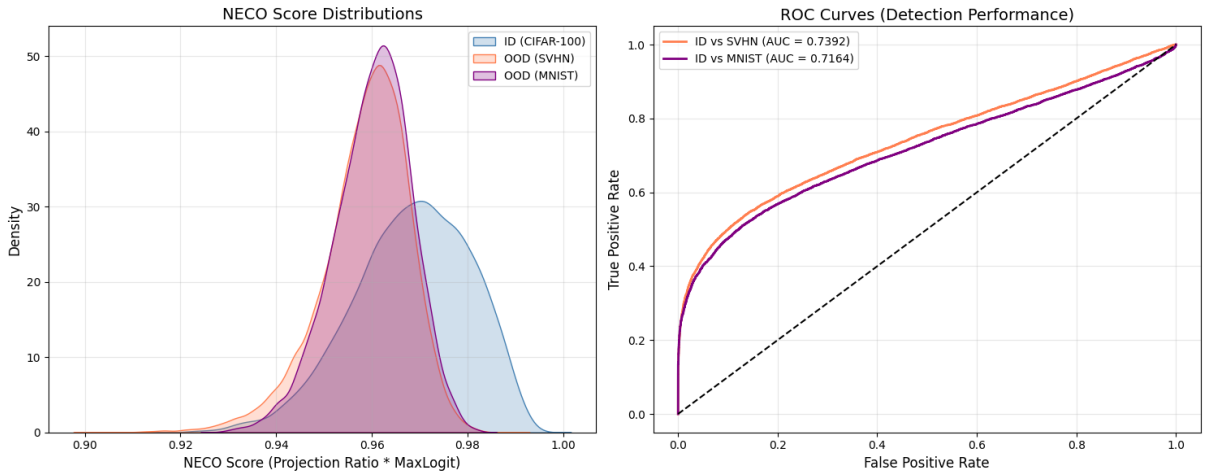

NC2: Pairwise Cosine Similarity Distribution
ETF Theory: -0.0101
Mean: -0.0092
Frequency
Cosine Similarity
NC2: Pairwise Angle Distribution
ETF Theory: 90.58°
Mean: 90.47°
Frequency
Angle (degrees)
NC2: Cosine Similarity Heatmap (20 random classes)
Class ID
Class ID
Cosine Similarity
NC2: Deviation from Perfect ETF
Perfect ETF
Frequency
Deviation from Theoretical ETF

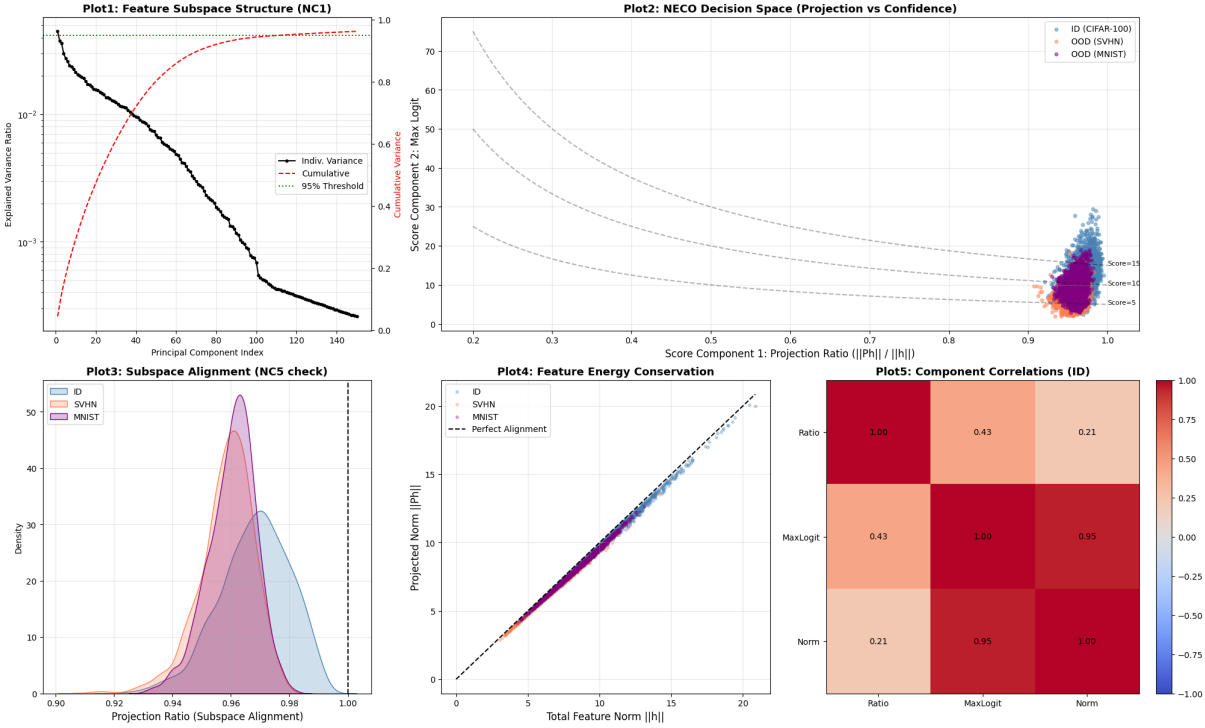

NC3: Dimensional Alignment (Class 0, r=0.891)
Perfect alignment (y=x)
Class 0 Weight (512 dims)
Class 0 Mean (512 dims)
NC3: Class Mean vs Weight Correlation
Strong NC3: r>0.95
Mean: 0.887
Number of Classes
Pearson Correlation
NC3: Angular Deviation (Secondary Metric)
Perfect alignment (uniform scaling): <10°
Mean: 50.92°
Number of Classes
Angle (degrees)
NC3: Cross-Correlation Heatmap (20 classes)
Class Mean Class ID
Classifier Weight Class ID
Correlation


NC4: Model vs Nearest-Center Consistency per Class
Strong NC4: >95%
Mean: 91.4%
Consistency (%)
Class ID
NC4: Consistency Distribution
Strong NC4
Mean: 91.4%
Number of Classes
Consistency (%)

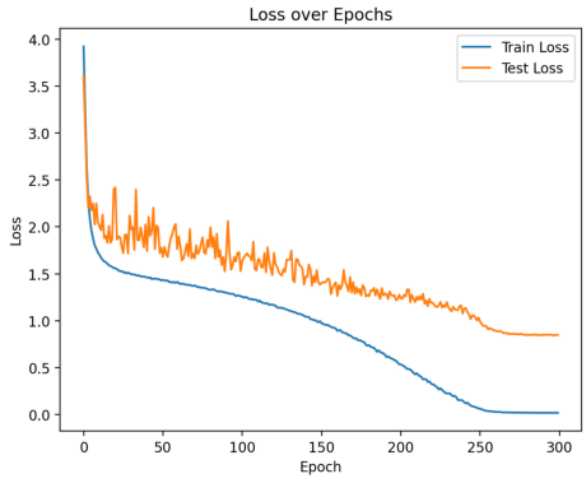

NC4: Mismatch Heatmap (20 classes)
True Label
Model Prediction
Mismatch Count
NC4: Sample Count vs Consistency
Consistency (%)
Samples per Class

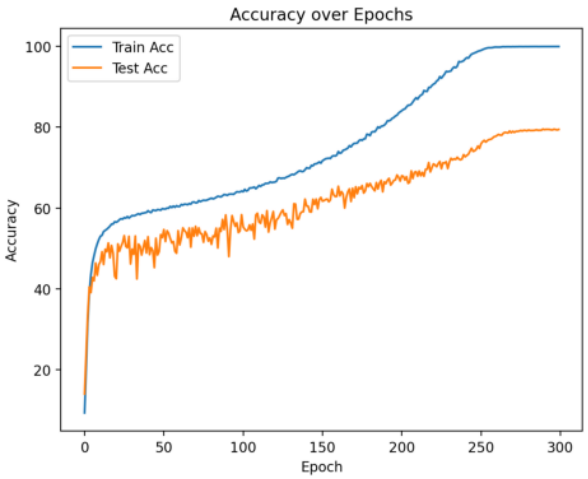

Plot 1: ID vs SVHN t-SNE (Weak NC5)
ID (CIFAR-100)
OOD (SVHN)
t-SNE Component 1
t-SNE Component 2
Plot 2: Distance Distribution (Ratio: 1.12x)
ID
SVHN
ID Mean: 7.19
SVHN Mean: 8.03
Density
Distance to Nearest Center
Plot 3: Feature Norm Distribution (SVHN)
ID
SVHN
Density
Feature Norm
Plot 4: ID vs MNIST t-SNE (Strong NC5)
ID (CIFAR-100)
OOD (MNIST)
t-SNE Component 1
t-SNE Component 2
Plot 5: Distance Distribution (Ratio: 1.10x)
ID
MNIST
ID Mean: 7.19
MNIST Mean: 7.91
Density
Distance to Nearest Center
Plot 6: Feature Norm Distribution (MNIST)
ID
MNIST
Density
Feature Norm

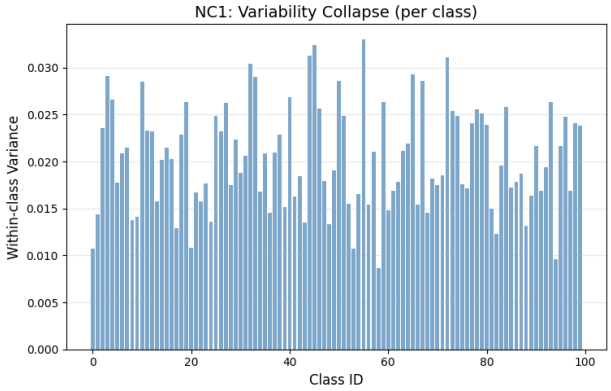

NECO Score Distributions
ID (CIFAR-100)
OOD (SVHN)
OOD (MNIST)
Density
NECO Score (Projection Ratio * MaxLogit)
ROC Curves (Detection Performance)
ID vs SVHN (AUC = 0.7392)
ID vs MNIST (AUC = 0.7164)
True Positive Rate
False Positive Rate

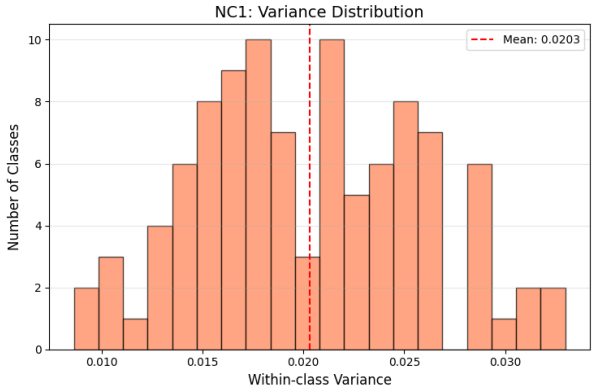

Plot1: Feature Subspace Structure (NC1)
Explained Variance Ratio
Cumulative Variance
Indiv. Variance
Cumulative
95% Threshold
Principal Component Index
Plot2: NECO Decision Space (Projection vs Confidence)
ID (CIFAR-100)
OOD (SVHN)
OOD (MNIST)
Score Component 2: Max Logit
Score Component 1: Projection Ratio (||Ph|| / ||h||)
Score=15
Score=10
Score=5
Plot3: Subspace Alignment (NC5 check)
ID
SVHN
MNIST
Density
Projection Ratio (Subspace Alignment)
Plot4: Feature Energy Conservation
ID
SVHN
MNIST
Perfect Alignment
Projected Norm ||Ph||
Total Feature Norm ||h||
Plot5: Component Correlations (ID)
Ratio
MaxLogit
Norm
1.00
0.43
0.21
0.43
1.00
0.95
0.21
0.95
1.00

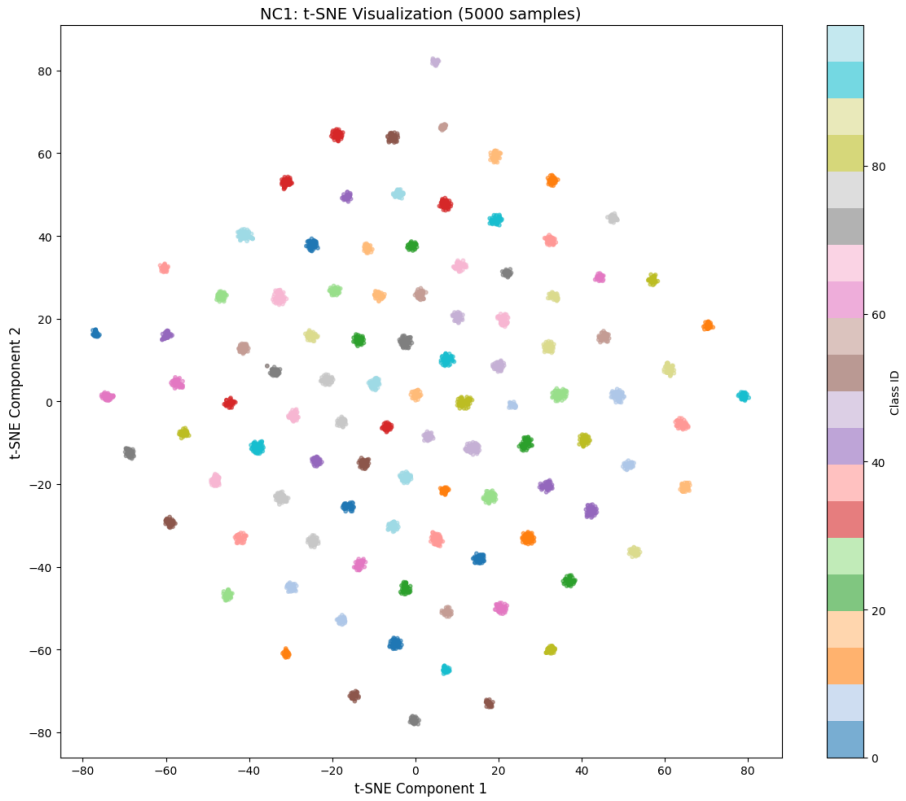
B Model Without Dropout

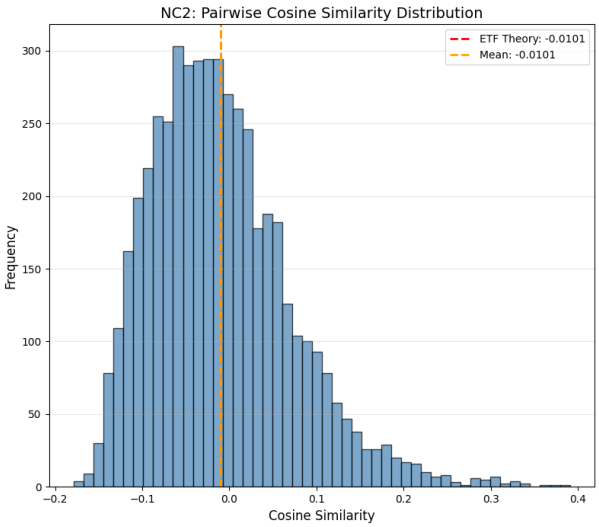


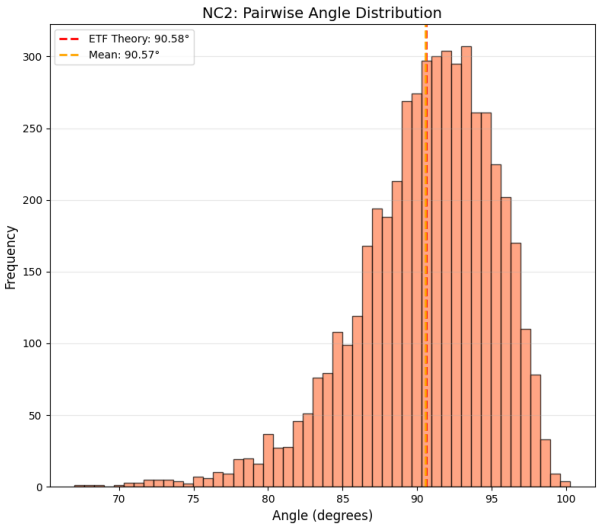


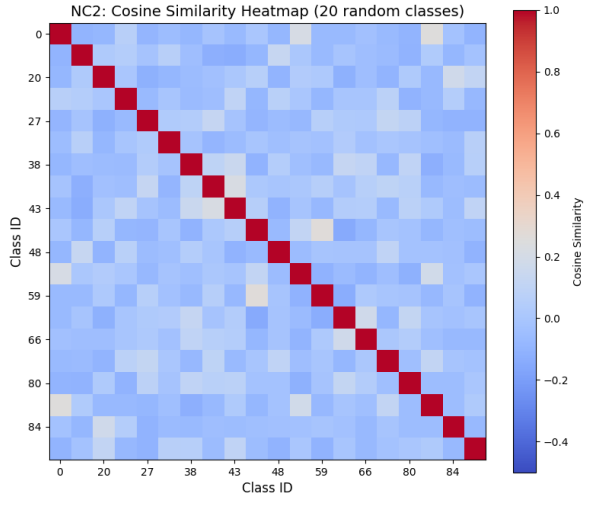


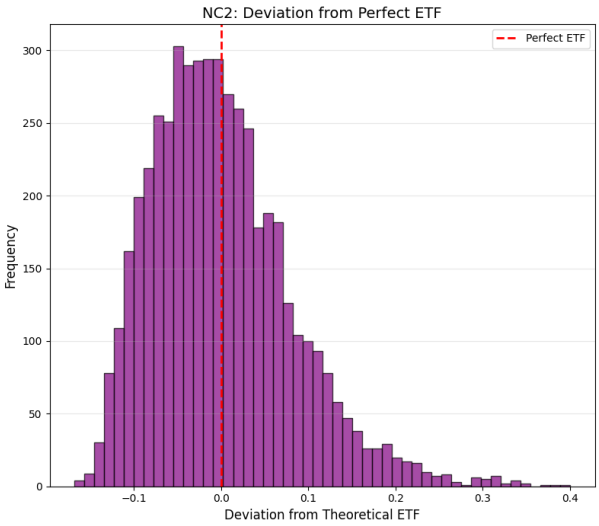


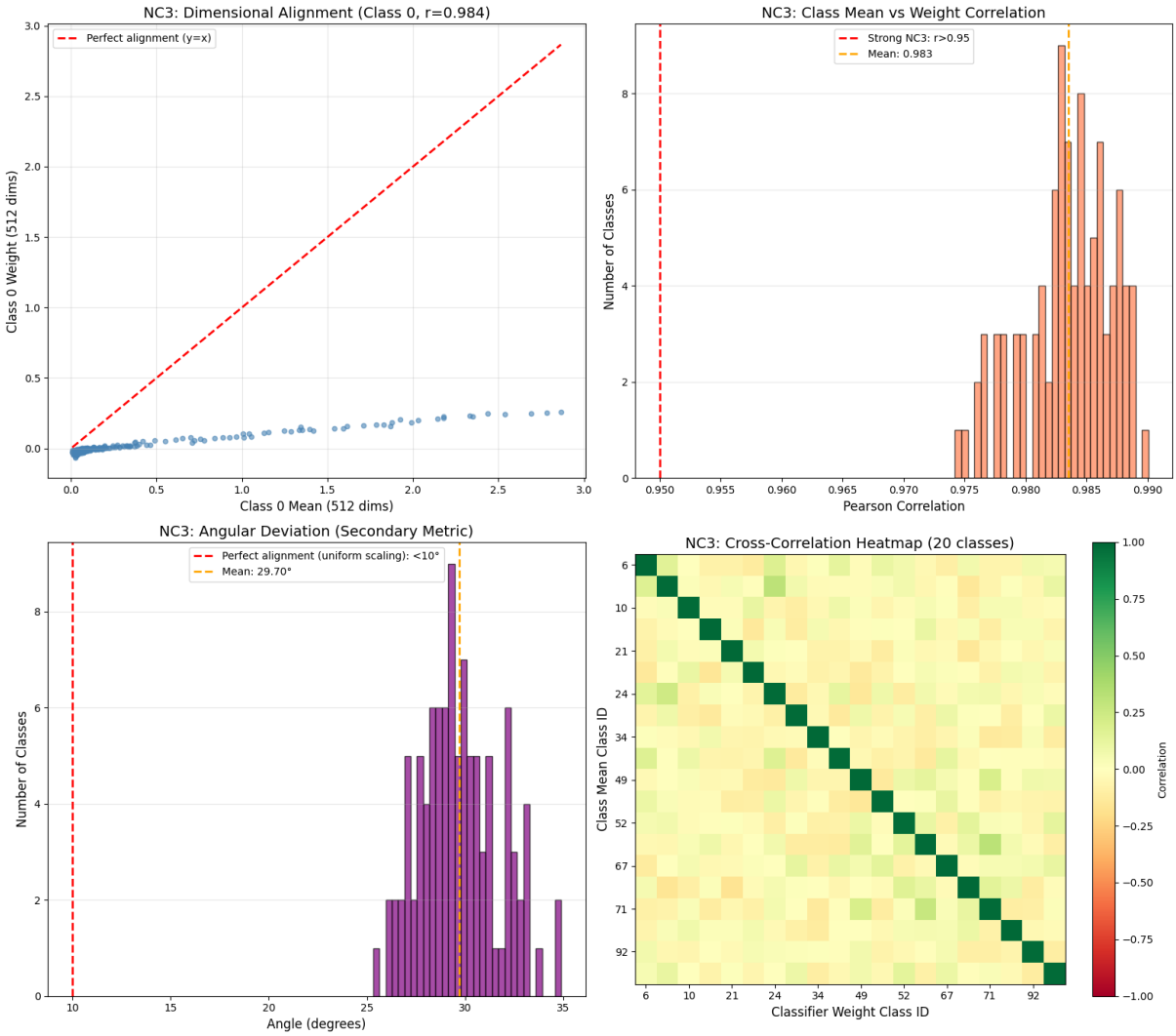

NC1: t-SNE Visualization (5000 samples)
t-SNE Component 1
t-SNE Component 2
Class ID
80
60
40
20
0
−20
−40
−60
−80
−80
−60
−40
−20
0
20
40
60
80

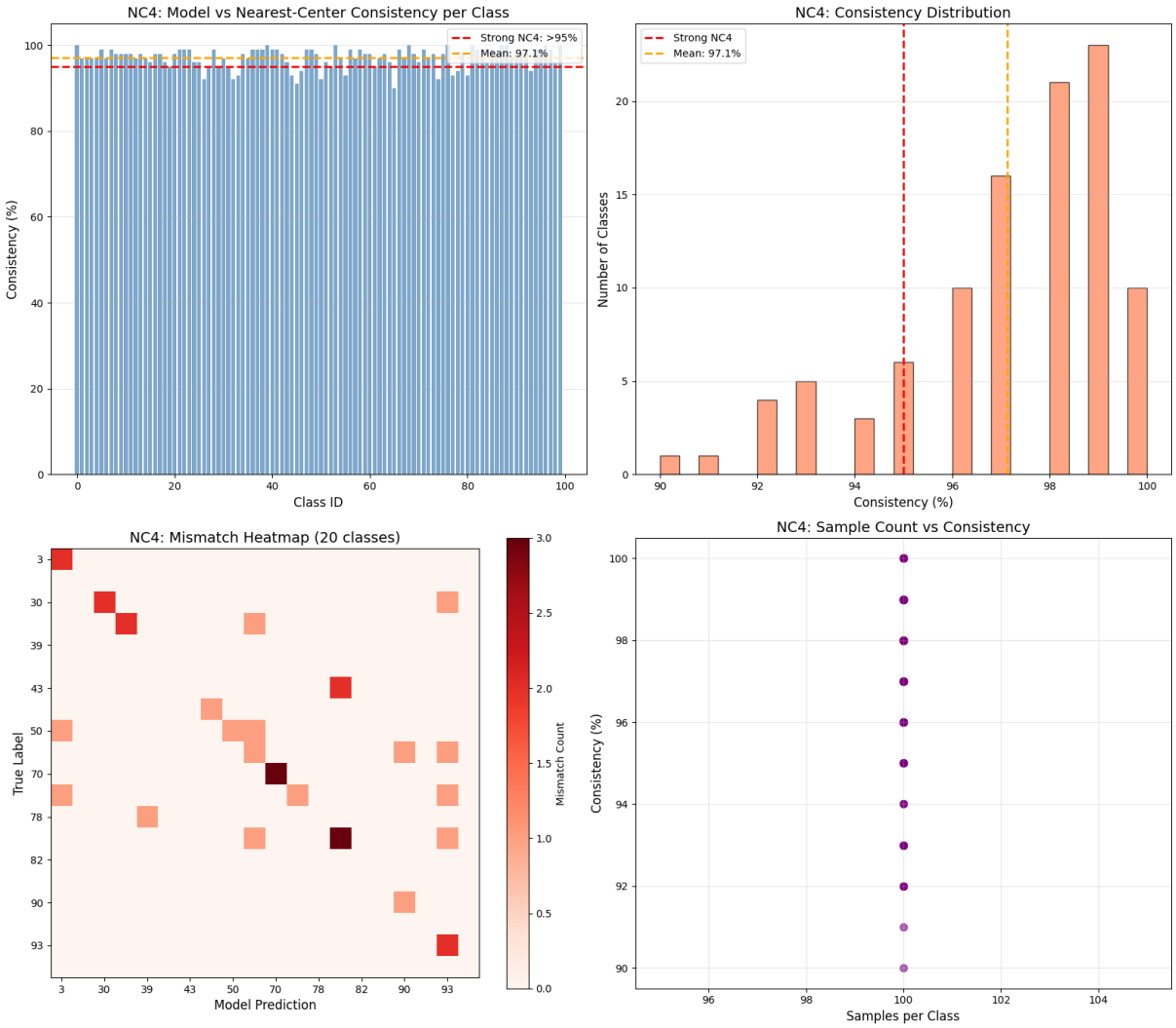

NC2: Pairwise Cosine Similarity Distribution
ETF Theory: -0.0101
Mean: -0.0101
Frequency
Cosine Similarity

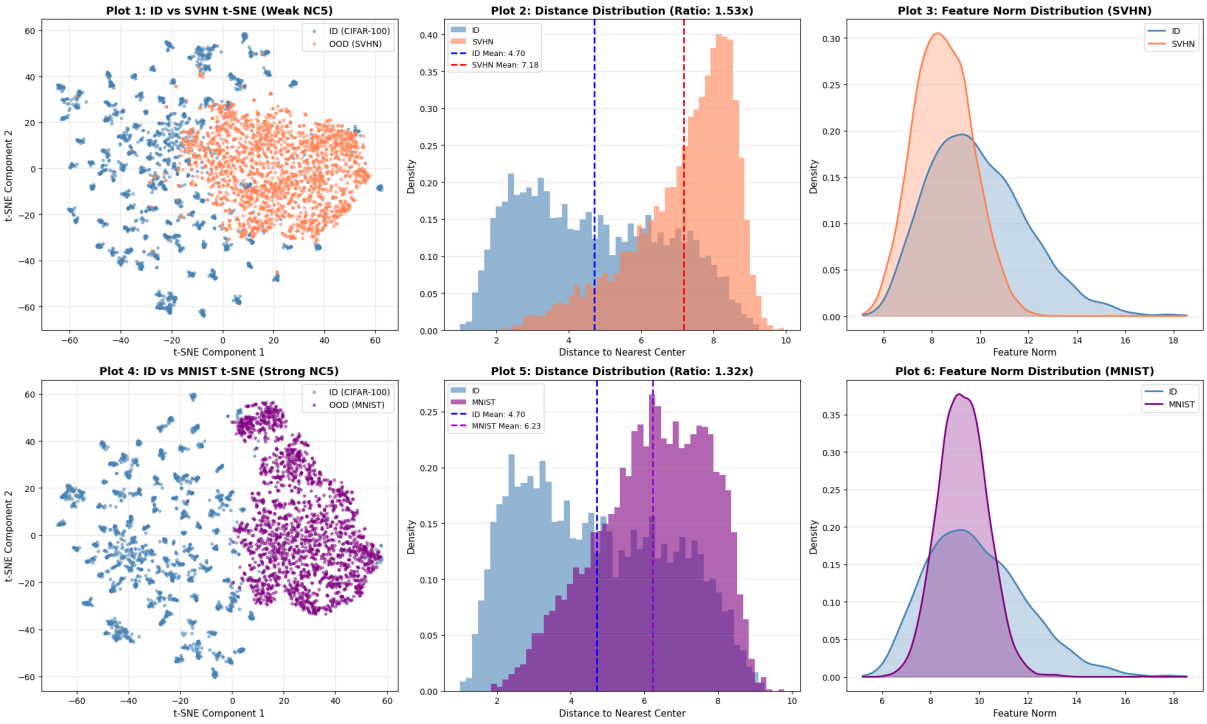

NC2: Pairwise Angle Distribution
ETF Theory: 90.58°
Mean: 90.57°
Frequency
Angle (degrees)

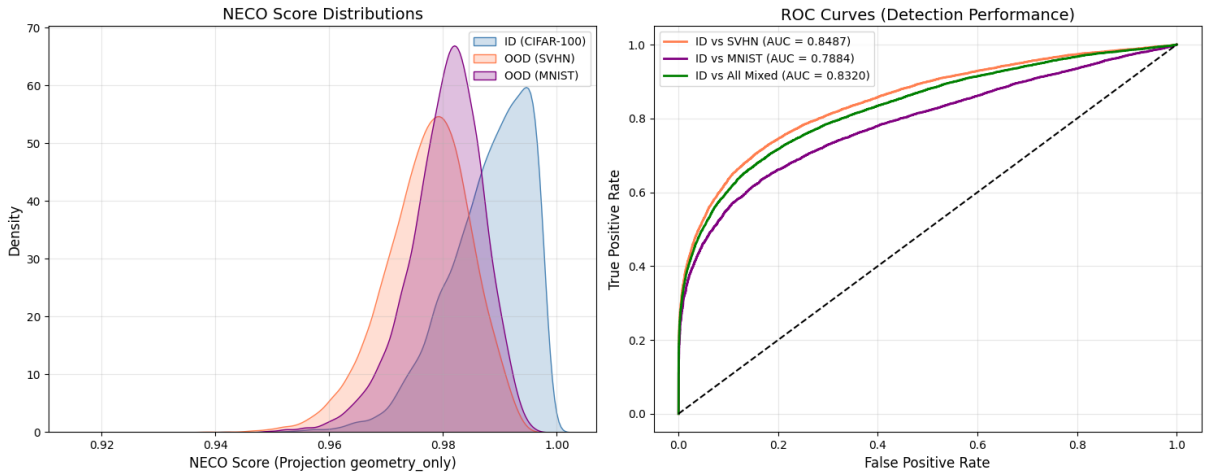

NC2: Cosine Similarity Heatmap (20 random classes)
Class ID
Class ID
Cosine Similarity

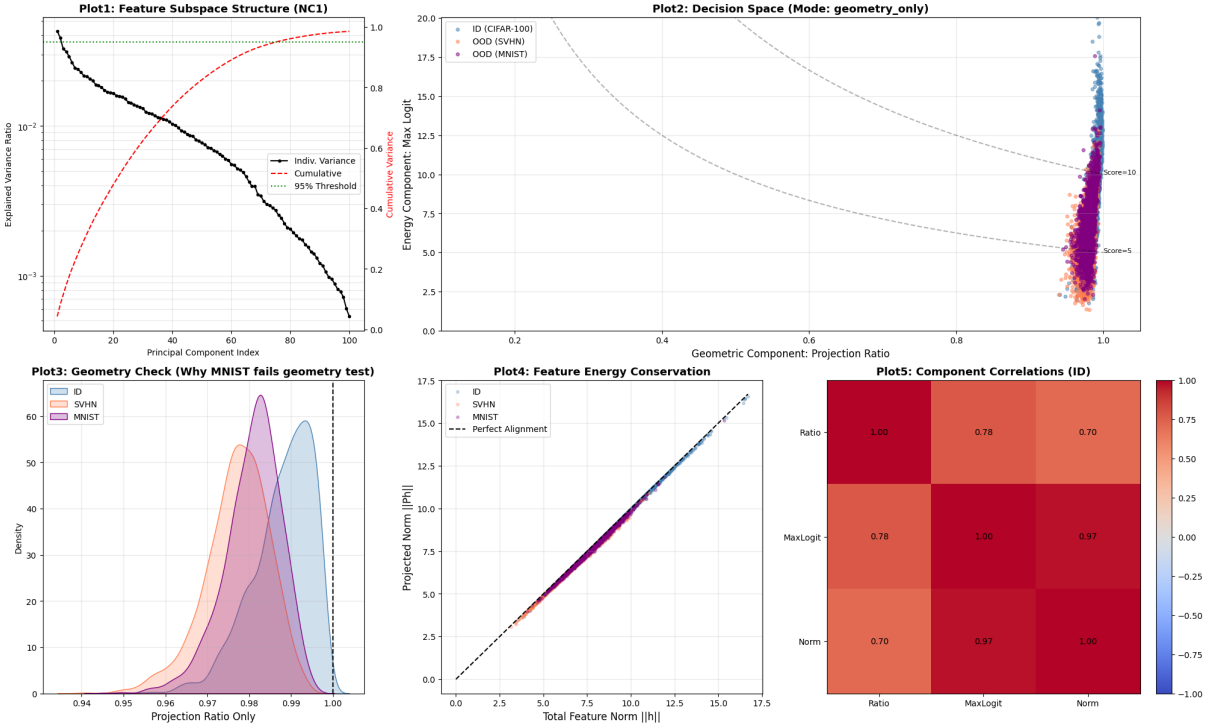

NC2: Deviation from Perfect ETF
Perfect ETF
Frequency
Deviation from Theoretical ETF


NC3: Dimensional Alignment (Class 0, r=0.984)
Perfect alignment (y=x)
Class 0 Weight (512 dims)
Class 0 Mean (512 dims)
NC3: Class Mean vs Weight Correlation
Strong NC3: r>0.95
Mean: 0.983
Number of Classes
Pearson Correlation
NC3: Angular Deviation (Secondary Metric)
Perfect alignment (uniform scaling): <10°
Mean: 29.70°
Number of Classes
Angle (degrees)
NC3: Cross-Correlation Heatmap (20 classes)
Class Mean Class ID
Classifier Weight Class ID
Correlation


NC4: Model vs Nearest-Center Consistency per Class
Strong NC4: >95%
Mean: 97.1%
Consistency (%)
Class ID
NC4: Consistency Distribution
Strong NC4
Mean: 97.1%
Number of Classes
Consistency (%)
NC4: Mismatch Heatmap (20 classes)
True Label
Model Prediction
Mismatch Count
NC4: Sample Count vs Consistency
Consistency (%)
Samples per Class


Plot 1: ID vs SVHN t-SNE (Weak NC5)
ID (CIFAR-100)
OOD (SVHN)
t-SNE Component 1
t-SNE Component 2
Plot 2: Distance Distribution (Ratio: 1.53x)
ID
SVHN
ID Mean: 4.70
SVHN Mean: 7.18
Density
Distance to Nearest Center
Plot 3: Feature Norm Distribution (SVHN)
ID
SVHN
Density
Feature Norm
Plot 4: ID vs MNIST t-SNE (Strong NC5)
ID (CIFAR-100)
OOD (MNIST)
t-SNE Component 1
t-SNE Component 2
Plot 5: Distance Distribution (Ratio: 1.32x)
ID
MNIST
ID Mean: 4.70
MNIST Mean: 6.23
Density
Distance to Nearest Center
Plot 6: Feature Norm Distribution (MNIST)
ID
MNIST
Density
Feature Norm


NECO Score Distributions
ID (CIFAR-100)
OOD (SVHN)
OOD (MNIST)
Density
NECO Score (Projection geometry_only)
ROC Curves (Detection Performance)
ID vs SVHN (AUC = 0.8487)
ID vs MNIST (AUC = 0.7884)
ID vs All Mixed (AUC = 0.8320)
True Positive Rate
False Positive Rate


Plot1: Feature Subspace Structure (NC1)
Indiv. Variance
Cumulative
95% Threshold
Explained Variance Ratio
Cumulative Variance
Principal Component Index
Plot2: Decision Space (Mode: geometry_only)
ID (CIFAR-100)
OOD (SVHN)
OOD (MNIST)
Score=10
Score=5
Energy Component: Max Logit
Geometric Component: Projection Ratio
Plot3: Geometry Check (Why MNIST fails geometry test)
ID
SVHN
MNIST
Density
Projection Ratio Only
Plot4: Feature Energy Conservation
ID
SVHN
MNIST
Perfect Alignment
Projected Norm ||Ph||
Total Feature Norm ||h||
Plot5: Component Correlations (ID)
Ratio
MaxLogit
Norm
1.00
0.78
0.70
0.78
1.00
0.97
0.70
0.97
1.00