Revisiting Class-Incremental Learning in the Era of Foundation Models: A Gaussian Density Estimation Perspective

Anonymous authors
Paper under double-blind review

Abstract

Class-Incremental Learning (CIL) is commonly addressed by preventing catastrophic forgetting through regularization, replay buffers, or architectural adaptations. In this work, we revisit this formulation in the context of modern visual foundation models. We argue that once representations are frozen and sufficiently discriminative, the main challenge of CIL shifts from protecting model parameters to modeling class distributions in a stationary feature space. Building on this observation, we formulate incremental learning as recursive estimation of class-conditional Gaussian densities in foundation embeddings. To handle data scarcity, we introduce the Hybrid-Cov estimator, which decouples the Mahalanobis shape term from the log-normalization constant, providing a numerically stable alternative when the sample budget is small. We validate Gaussianity rigorously using formal statistical tests (D'Agostino-Pearson and Henze-Zirkler) and quantify inter-class separability via the Bhattacharyya distance, which provides a direct connection to the Bayes error rate. Despite its simplicity, the proposed approach achieves competitive performance on CIFAR-100, ImageNet-R, and Caltech-256, while requiring significantly lower computational cost and providing a principled probabilistic interpretation.

1 Introduction

Class-Incremental Learning (CIL) aims to extend a recognition system with new classes while maintaining performance on previously learned categories. The dominant view treats this problem as one of parameter interference: learning new classes requires updating network weights, which may overwrite previously acquired knowledge. Consequently, most existing methods attempt to control parameter updates through regularization strategies, replay buffers, or architectural adaptations. Although effective, these approaches share a common premise: incremental learning must be achieved through continuous, careful optimization of model parameters across the network.

The assumption that feature representations must remain plastic diverges significantly from the organizational principles of biological vision. The human visual system relies on early cortical regions responsible for feature extraction that mature and stabilize during the initial postnatal weeks or months Sluyters et al. (1990), providing highly stable feature representations that remain stationary throughout adulthood. However, despite this invariant sensory foundation, human lifelong learning persists. Higher-order cognitive systems seamlessly map these fixed perceptual representations onto newly encountered semantic categories, even if acquiring knowledge of a new object class can occasionally result in a slight, localized reduction in the ability to distinguish highly similar classes over time.

A similar structural separation has recently emerged in computer vision with the advent of large-scale visual foundation models. Encoders trained on massive datasets produce highly discriminative representations that remain stable when frozen. In such stationary feature spaces, representation drift—the primary source of catastrophic forgetting in deep neural networks—is eliminated by construction. Mirroring the biological paradigm, class-incremental learning can thus be reframed: it becomes a problem of modeling new class distributions within a fixed, robust embedding space, rather than continually adapting the underlying feature extractors.

In this work, we revisit a classical pattern recognition concept within this modern context: modeling class-conditional densities using Gaussian statistics. While Gaussian classifiers were historically constrained by weak or unstable representations, foundation models fundamentally transform this landscape. In their well-behaved embedding spaces, simple second-order statistical models provide a strong inductive bias that enables highly effective incremental recognition. We formulate CIL as the recursive estimation of class-conditional Gaussian densities in foundation embeddings. Each class is represented entirely by its sufficient statistics, which are updated incrementally without backpropagation, replay buffers, or exemplar storage. This approach yields a lightweight probabilistic classifier that eliminates the need for complex, parameter-heavy architectural adaptations.

Crucially, we supplement our empirical evaluation with a rigorous statistical validation of the Gaussian assumption in these embedding spaces. We employ two formal hypothesis tests alongside moment-based diagnostics (skewness and excess kurtosis): the D'Agostino–Pearson omnibus test D'Agostino (1971); D'Agostino & Pearson (1973) for univariate normality and the Henze–Zirkler (HZ) test Henze & Zirkler (1990) for multivariate Gaussianity. Furthermore, to quantify inter-class separability and establish theoretical limits, we utilize the *Bhattacharyya distance* D_B—a symmetric measure directly tied to the Bayes error rate. Our findings suggest that in the foundation model era, the central challenge of incremental learning lies not in preventing weight forgetting, but in properly exploiting the statistical structure of stable representation spaces.

In summary, the main contributions of this work are:

- **Reformulation.** We recast CIL as recursive probability density estimation in frozen foundation feature spaces, shifting the focus from parameter preservation to statistical modeling.
- **Probabilistic modeling.** We demonstrate that maintaining simple per-class first and second-order statistics suffices for competitive classification. To this end, we systematically evaluate multiple covariance parameterizations ranging from diagonal to full structures. Crucially, we introduce the Hybrid-Cov estimator, which decouples the Mahalanobis shape term from the log-normalization constant, providing a principled and numerically stable alternative when the sample budget is smaller than the feature dimensionality.
- **Rigorous Gaussianity analysis.** We validate the Gaussian assumption using formal univariate (D'Agostino–Pearson, Henze–Zirkler) and multivariate (diagonal HZ) normality tests, and quantify inter-class separability via the Bhattacharyya distance.
- **Competitive performance.** We achieve competitive accuracy on CIFAR-100, ImageNet-R, and Caltech-256, matching strong baselines with minimal overhead beyond the backbone.

2 Related Work

Regularization-based methods. To maintain model stability across sequential tasks, these methods limit updates to parameters identified as important for prior learning. Representative strategies include slowing updates to parameters estimated to be important for prior tasks Aljundi et al. (2018); Kirkpatrick et al. (2017), restricting learning in neurons critical for earlier tasks Jung et al. (2020), and introducing additional loss terms based on knowledge distillation Li & Hoiem (2018).

Rehearsal-based methods. These methods reuse samples from previously learned tasks. Some store a small set of exemplars Rebuffi et al. (2016); others maintain a limited memory buffer Belouadah & Popescu (2019); a third strategy uses stored samples to constrain gradient updates Lopez-Paz & Ranzato (2017); Chaudhry et al. (2019). Generative replay trains a generative model to produce samples from previous tasks Ostapenko et al. (2019); Shin et al. (2017); Wu et al. (2018).

Architecture-based methods. These methods modify the network architecture to increase or reorganize its learning capacity Hung et al. (2019); Jin & Kim (2022); Yan et al. (2021); Yoon et al. (2018).

Prompt-based methods. The emergence of foundation models has enabled parameter-efficient fine-tuning via learned prompts injected into frozen backbones. L2P Wang et al. (2021) introduces a prompt pool se-

lected by query-key similarity. DualPrompt Wang et al. (2022) combines task-specific and shared prompts. PGP Qiao et al. (2024) constrains gradient updates via prompt-based projection. CODA-Prompt Smith et al. (2023) proposes fully differentiable prompt construction. CPrompt Gao et al. (2024) addresses training–testing inconsistencies. ConvPrompt Roy et al. (2024) generates prompts via convolution over a shared embedding matrix. PromptFusion Chen et al. (2024) separates stability and plasticity using two prompt modules. APT Chen et al. (2025) directly injects learnable key-value tokens into transformer layers. RainbowPrompt Hong et al. (2025) enhances representational diversity through prompt-evolving.

Class prototype methods. Some methods are based on using distances to class prototypes (first order information), sometimes alongside second-order information. SimpleCIL Janson et al. (2022) is based on using first-session adaptation by using parameter-efficient fine tuning (PEFT) on the first task to adapt the backbone. Then, class prototypes are generated from the samples corresponding to the different tasks. RanPAC McDonnell et al. (2023) improves that method by applying a random projection followed by a non-linear activation, to make the features more separable in a larger feature space. Then it accumulates second-order information by using a shared Gram matrix learned incrementally, which enables compensating for the effect caused by correlations between features. Fetril Petit et al. (2023) stores centroid representations of past classes. When features from a new class arrive, the centroids are used to generate pseudofeatures, which correspond to the same new features, but translated using the stored centroids. Then, the pseudofeatures are used to train a linear classifier incrementally. FeCAM Goswami et al. (2023) explores the use of Mahalanobis distances instead of Euclidean distances commonly used in current prototype-based methods. It explores applying Tukey's ladder of Powers transformation to correct the skewness detected in its features, and then proposes to use shrinkage and then normalize the class covariance matrices by making their diagonal elements equal to 1, so that the Mahalanobis distances are comparable between classes. Despite the use of Mahalanobis distances being able to improve the results, the gaussian normalization factor is not considered and the covariance normalization is heuristic, which makes this work fail to follow a fully probabilistic approach.

The works most closely related to ours are FeCAM and RanPAC, which also operate on frozen feature spaces and exploit second-order information. However, FeCAM's use of the Mahalanobis distance omits the log-normalization term of the Gaussian density and applies ad-hoc covariance normalization, making it not a fully probabilistic classifier. RanPAC applies random projections to lift the feature space before accumulating a shared Gram matrix, introducing additional hyperparameters and breaking the one-to-one correspondence between class statistics and probability densities. More broadly, no existing work rigorously validates whether the Gaussian assumption is statistically justified in foundation model embedding spaces, nor provides a direct theoretical link between inter-class separability and achievable accuracy. The present work closes both gaps: it proposes a clean, fully probabilistic CIL framework with formal MAP inference, and validates the Gaussian assumption across three benchmarks using formal hypothesis tests and Bhattacharyya-based error bounds.

3 Proposed Method

Our CIL framework, as shown in Figure 1, is based on a highly modular decomposition consisting of a frozen visual foundation model for the features extraction and a probabilistic classification layer. The feature extractor maps input images to a fixed, high-dimensional embedding space where each object category is modeled as an independent, class-conditional, multivariate Gaussian distribution. These distributions are defined by mean vectors and covariance matrices estimated incrementally from observed samples. In inference, classification follows the maximum likelihood principle, assigning the input to the category that maximizes the likelihood under the stationary representation. By completely separating representation learning from statistical modeling, our framework enables backpropagation-free incremental updates; incorporating a novel category only requires estimating its distribution in the fixed feature space, leaving the representations of previously learned classes entirely intact.

3.1 Foundation Feature Space

A fundamental assumption of our framework is the availability of a pre-trained and frozen feature extractor. The representation model must be trained prior to the initialization of the incremental recognition system and

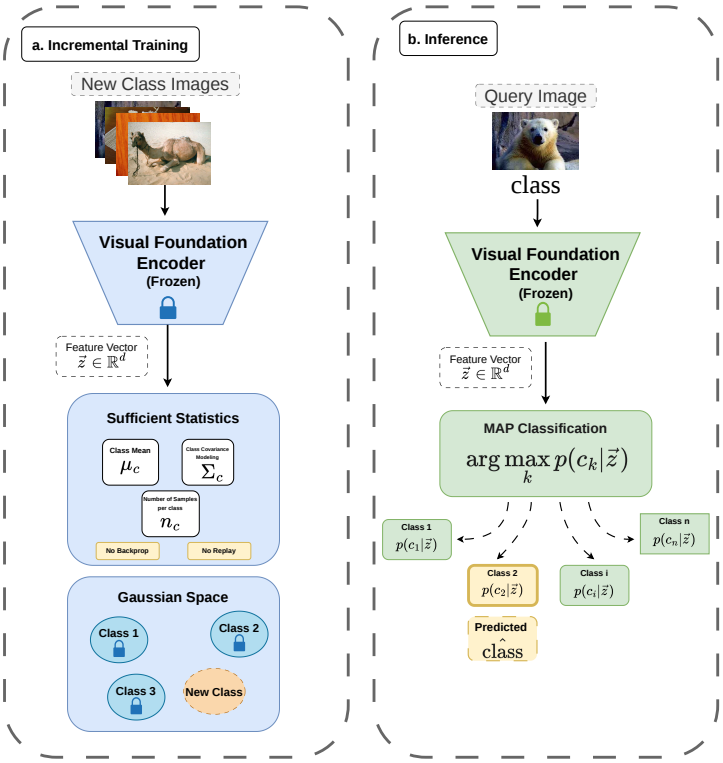


Figure 1: Overview of the proposed framework. A frozen foundation encoder maps each input image to a fixed embedding $\mathbf{z} \in \mathbb{R}^d$. Each class is represented by a Gaussian density (mean and covariance) updated incrementally without backpropagation. At inference, classification follows the MAP rule under the learned class-conditional densities.

remain fixed during all subsequent class-incremental updates. Moreover, the resulting embedding space must be sufficiently general and discriminative to characterize not only the classes observed during its pre-training phase, but also novel categories introduced over time.

This design is inspired by the hierarchical organization of the human visual system. During early development, exposure to diverse visual stimuli shapes stable, low-level feature detectors in the early visual cortex Sluyters et al. (1990). Once established, these foundational perceptual representations remain remarkably stable throughout adulthood, providing a constant sensory baseline, while higher-level cortical regions accommodate the acquisition of new semantic categories. Analogously, our framework assumes a stationary perceptual embedding space upon which incremental category learning is performed.

In practice, replicating such robust representations requires training in large-scale and diverse image corpora. Modern visual foundation models satisfy this requirement by leveraging supervised or self-supervised learning on massive datasets, yielding highly transferable and structured feature spaces. In our experiments, we employ Vision Transformer (ViT) Dosovitskiy et al. (2020) and DINOv2 Oquab et al. (2024) models as representative foundation encoders. These models are used exclusively as frozen feature extractors, without any fine-tuning during incremental learning.

3.2 Feature Representation

Given an input image $x \in \mathcal{X}$, we extract a fixed-dimensional embedding using a pretrained visual foundation model:

$$\mathbf{z} = \mathcal{F}(x) \in \mathbb{R}^d. \tag{1}$$

The encoder $\mathcal{F}$ remains frozen during training and inference. All learning is performed directly in the feature space $\mathbb{R}^d$.

3.3 Likelihood-Based Classification

We adopt a generative modeling framework and assume class-conditional Gaussian feature distributions:

$$p(\mathbf{z} \mid y = c) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{2}$$

where $\boldsymbol{\mu}_c \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_c \succ 0$.

Classification is performed via maximum a posteriori (MAP) inference:

$$\hat{y} = \arg\max_{c \in \mathcal{Y}} \left[\log p(\mathbf{z} \mid y = c) + \log \pi_c\right], \tag{3}$$

where $\pi_c = p(y = c)$ is the class prior. Under a Gaussian model this is equivalent to:

$$\hat{y} = \arg\min_{c} \left[(\mathbf{z} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{z} - \boldsymbol{\mu}_c) + \log|\boldsymbol{\Sigma}_c| - 2\log \pi_c\right]. \tag{4}$$

Where $(\mathbf{z} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{z} - \boldsymbol{\mu}_c)$ is the squared Mahalanobis distance, and $\log|\boldsymbol{\Sigma}_c|$ is the log-determinant.

Adding a new class requires only estimating $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ in the fixed feature space; previously learned class statistics are never modified.

3.4 The Small-Sample Rank-Deficiency Challenge for Covariance Modeling

Covariance conditioning The sample covariance $\hat{\Sigma}_c \in \mathbb{R}^{d \times d}$ estimated from n_c training samples is structurally rank-deficient whenever $n_c < d$. In our experimental setup that uses ViT and DINOv2 as foundation models, $d = 768$, while n_c ranges from ≈ 40 to 500, meaning that the empirical covariance matrices are consistently singular. This severe ill-conditioning prevents the stable computation of standard Mahalanobis distances and log-determinants due to near-zero eigenvalues. This statistical limitation directly motivates our alternative parameterizations— like the Hybrid Covariance model, which provides a robust probabilistic trade-off when the sample budget n_c is small.

We present four alternative covariance parameterizations designed to balance model capacity against estimation stability, followed by an extension of these methods to multimodal distributions via a Mixture of Gaussians framework. Specifically, we consider diagonal, full, hybrid, and shared covariance structures, analyzing their mathematical formulations, parameter complexities, and regularization strategies to ensure numerical robustness.

Diagonal Covariances In this configuration, we assume that the features within each class c are independent, leading to a diagonal covariance matrix $\Sigma_c = \mathrm{diag}(\sigma_{c,1}^2, \sigma_{c,2}^2, \ldots, \sigma_{c,d}^2)$. This simplification significantly reduces the number of parameters to estimate from $O(d^2)$ to $O(d)$, which is particularly advantageous when the dimensionality d is large relative to the number of samples per class. For each class $c \in \{1, \ldots, C\}$ containing training samples z, the mean vector μ_c and the diagonal elements of Σ_c are computed as:

$$\mu_c = \frac{1}{n_c} \sum_{i=1}^{n_c} z_i, \quad \sigma_{c,j}^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} (z_{i,j} - \mu_{c,j})^2 + \epsilon \tag{5}$$

where $\epsilon = 10^{-6}$ is a small constant added to ensure numerical stability and strictly positive variances.

Full Covariances To account for the dependencies between different dimensions of the feature vector, we estimate a full covariance matrix $\Sigma_c \in \mathbb{R}^{d \times d}$ for each class c. While this provides a richer representation of the class distribution, the estimation of d^2 parameters is prone to ill-conditioning when the number of samples n_c is limited.The mean vector μ_c and the empirical covariance Σ_c are calculated as:

$$\mu_c = \frac{1}{n_c}\sum_{i=1}^{n_c} z_i, \quad \Sigma_c = \frac{1}{n_c - 1}\sum_{i=1}^{n_c}(z_i - \mu_c)(z_i - \mu_c)^T \tag{6}$$

To ensure that the matrix is invertible and to improve numerical stability, a regularization is applied by adding a small identity term to the diagonal: $\hat{\Sigma}_c = \Sigma_c + \epsilon I$, with $\epsilon = 10^{-6}$.

For obtaining the precision matrix Q_c (the inverse of the covariance matrix $\hat{\Sigma_c}$), an approach based on Cholesky decompositions is used. The covariance $\hat{\Sigma_c}$ is decomposed as $\hat{\Sigma_c} = C_c C_c^T$ with C_c lower triangular, and then the Cholesky precision matrix L_c is obtained by solving $C_c L_c = I$, with I the identity.

Then, the squared Mahalanobis distance and the log-determinant are computed from μ_c and L_c as:

$$d_M(z, \mu_c, L_c) = (zL_c - \mu_c L_c)^T (zL_c - \mu_c L_c) \tag{7}$$

$$\ln|\hat{\Sigma}_c| = 2\sum_{i=1}^{d} \ln(L_{c\,i,i}) \tag{8}$$

Hybrid Covariance The hybrid covariance estimator decouples the squared Mahalanobis distance from its associated normalization constant. For each class c, the squared Mahalanobis distance d_M is computed using the procedure described in the previous section (based on Cholesky decomposition). However, the log-determinant term is computed using only the diagonal elements (variances) of the covariance matrix:

$$\ln|\hat{\Sigma}_c|_{\text{diag}} = \sum_{j=1}^{d} \ln(\sigma_{c,j}^2) \tag{9}$$

where $\sigma_{c,j}^2$ denotes the j-th diagonal entry of $\hat{\Sigma}_c$. The resulting discriminant function is defined as:

$$g_c(z) = -\frac{1}{2}\left(d\ln(2\pi) + \sum_{j=1}^{d}\ln(\sigma_{c,j}^2) + (zL_c - \mu_c L_c)^T(zL_c - \mu_c L_c)\right) \tag{10}$$

Unlike shrinkage estimators (Ledoit-Wolf, OAS), which blend the empirical covariance with a structured target and thereby mix signal with noise across all eigenvalues, the Hybrid-Cov preserves the full Mahalanobis distance structure from the empirical covariance while replacing only the volume term — the component most sensitive to near-zero eigenvalues — with a diagonal approximation that is well-conditioned by construction.

Shared Covariance The Shared Covariance framework assumes that the feature variability is homoscedastic across all classes. In this paradigm, all classes share a single covariance matrix Σ, which captures the global structure of the feature space.

A key aspect of this framework in the continual learning setting is that Σ is re-estimated at the end of each task t using only the classes observed so far, $\mathcal{C}_t = \{1, \ldots, C_t\}$. Then, the same shared covariance Σ is used for all the classes (no individual class covariances are stored). The shared covariance can be computed as $\Sigma_{shared} = 1/(N-K)\sum_{c=1}^{K}(n_c - 1)\Sigma_c$, with N the total number of samples, K the number of classes and n_c the number of elements from the class c. Note that storing the sum $\sum_{c=1}^{K}(n_c - 1)\Sigma_c$ is sufficient for enabling the shared covariance to being updated with the covariance of the next class Σ_{K+1}.

Mixture of Gaussians Extension To capture multimodal feature distributions, we extend the class-conditional model to a mixture:

$$p(z \mid y = c) = \sum_{k=1}^{K_c} \alpha_{c,k} \mathcal{N}(z \mid \mu_{c,k}, \Sigma_{c,k}) \tag{11}$$

where $\alpha_{c,k} \geq 0$ and $\sum_k \alpha_{c,k} = 1$.

The MAP decision rule becomes

$$\hat{y} = \arg\max_c \left[\ln \left(\sum_{k=1}^{K_c} \alpha_{c,k} \mathcal{N}(z \mid \mu_{c,k}, \Sigma_{c,k}) \right) + \ln \pi_c \right] \tag{12}$$

the number of components $K_c \in \{1, 2, 3, 4\}$ can be selected per class by minimizing the Bayesian Information Criterion (BIC), using either diagonal, full, or hybrid covariance matrices per component.

3.5 Parameter Estimation and Incremental Updates

Given training data $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$, parameters are estimated using maximum likelihood. All sufficient statistics (sample count, mean, and sum of outer products) admit closed-form incremental updates; then class expansion for most Gaussian methods (except MOG4-F) requires only summary statistics and no storage of past samples. In the case of the Shared-Diag-Cov, and Shared-Full-Cov methods, the covariances of all known classes must be also stored for enabling incremental updates of known classes.

4 Experiments

We organize our empirical evaluation along two complementary axes. First, we benchmark the proposed Gaussian classifiers against prompt-based and feature-based CIL methods on three standard benchmarks and a new joint cross-dataset benchmark. Second, we rigorously examine the statistical foundations of our approach: we characterize the spectral geometry of the DINOv2 embedding space, then we validate the Gaussian assumption using formal hypothesis tests and quantify inter-class separability via the Bhattacharyya distance and its direct connection to the Bayes error rate. Together, these analyses provide both empirical evidence of competitive performance and a principled statistical explanation for why the approach succeeds.

4.1 Experimental Setup

Datasets and Protocol. We follow the standard class-incremental evaluation protocol on three benchmarks.

Split CIFAR-100 Krizhevsky (2009) consists of 100 object categories partitioned into 10 disjoint tasks of 10 classes each (400 training samples per class and 100 testing samples per class).

Split ImageNet-R Hendrycks et al. (2021) contains 200 ImageNet classes in diverse artistic styles, partitioned into 10 tasks of 20 classes each. Training samples are balanced to 40 per class. This is needed because some methods require the same number of samples per class because of rank constraints, and the smallest class has 40 training samples.

Split Caltech-256 Griffin et al. (2022) uses the first 250 of the original 256 categories, partitioned into 10 tasks of 25 classes each with 64 training samples per class.

In all settings, task identity is not provided at test time (single-head evaluation). The models are evaluated after each incremental phase on the union of all observed classes.

Joint Cross-Dataset Benchmark. We introduce a long-horizon benchmark that aggregates all 30 tasks from the three datasets (10 from each) into a single heterogeneous stream in randomized order. This setting evaluates resilience to catastrophic forgetting over 550 classes spanning natural images, artistic renditions, and high-resolution object photography.

Evaluation Metrics. We report Average Accuracy and Forgetting as defined in Chaudhry et al. (2018):

$$A_T = \frac{1}{T} \sum_{\tau=1}^{T} S_{T,\tau} \tag{13}$$

$$F_T = \frac{1}{T-1} \sum_{\tau=1}^{T-1} \big(\max_{k \leq T-1} S_{k,\tau} - S_{T,\tau} \big). \tag{14}$$

Methods. We evaluate five Gaussian variants: **Diag-Cov**, **Full-Cov**, **Hybrid-Cov**, **Shared-Diag-Cov** and **Shared-Full-Cov** (shared covariance, diagonal and full variants), and **MOG4-F** (mixture of up to 4 full-covariance Gaussians, BIC selection). Also, two extra gaussian methods using shrinkage **Oracle Approximating Shrinkage** (OAS) Chen et al. (2010) and **Ledoit-Wolf** (LW) Ledoit & Wolf (2004) are included. These shrinkage-based estimators regularize the empirical covariance matrix by blending it linearly with a structured target matrix, mitigating ill-conditioning and rank-deficiency artifacts when the number of samples per class is smaller than the feature dimensionality. Baselines include prompt-based methods: L2P Wang et al. (2021), DualPrompt Wang et al. (2022), DualPrompt-PGP Qiao et al. (2024), CODA-Prompt Smith et al. (2023), CPrompt Gao et al. (2024), ConvPrompt Roy et al. (2024), RainbowPrompt Hong et al. (2025), and APT Chen et al. (2025). Additionally, we compare against feature-based and analytic CIL methods, namely SimpleCIL Janson et al. (2022), FeCAM Goswami et al. (2023), and RanPAC McDonnell et al. (2023).

Visual Foundation Models. We evaluate using: (1) DINOv2-B/14 Oquab et al. (2024), (2) ViT-B/16 pre-trained on ImageNet-21k Dosovitskiy et al. (2020), and (3) ViT-Medium (ViT-M/16) Wightman (2019).

4.2 Class-Incremental Classification

Tables 1–3 report performance. DINOv2-B/14 is used as backbone for all our methods in Tables 1 and 3. Table 2 analyzes the backbone impact.

As shown in Table 1, the Gaussian-based methods achieve performance within a narrow margin of the strongest prompt-based approaches on CIFAR-100 and ImageNet-R, despite relying on a significantly simpler classification mechanism: no additional trainable parameters or optimization procedures are required during incremental learning. It should be noted that certain baselines (such as SimpleCIL, FeCAM, and RanPAC) do not report standard deviations or forgetting metrics in this table simply because these values were not provided in their original papers. In contrast, for our framework, the absence of variance and conventional forgetting is a structural property: accuracies are completely deterministic with respect to class presentation order, since only independent sufficient statistics are maintained, and previously learned class statistics are never modified.

CIFAR 100: In this dataset, the most gaussian-based performant method is MOG4-F (88.87%) followed by Shared-Full-Cov (88.72%), Full-Cov (88.38%) and Hybrid-Cov (88.38%). Notably, Full-Cov and MOG4-F match ConvPrompt on CIFAR-100 (88.87%) without any trainable parameters or optimization during incremental learning. This fact shows that using full covariance matrices is the best approach on CIFAR-100 because of the high number of training examples by class (400 training samples per class). The diagonal-based methods achieve lower accuracies (86%+), while shrinkage-based methods (OAS and LW) achieve much lower accuracies (84%+). In the case of MOG4-F, all of the classes have only one gaussian, and the differences in results respect to Full-Cov are due only to minor implementation details.

ImageNet-R: In this dataset, the most performant gaussian-based methods are Shared-Full-Cov (77.99%) followed by Diag-Cov (77.12%). Full-Cov, Hybrid-Cov, Shared-Diag-Cov and MOG4-F achieve lower accuracies (74%+), while shrinkage methods OAS and LW have the lowest accuracies (71%+). In this dataset, all the MOG4-F models have only 1 gaussian component. The methods Shared-Full-Cov and Diag-Cov are able to work with a lower number of training samples per class (40 in this dataset), and then they achieve the strongest results.

Caltech-256: In this dataset, the most gaussian-based performant methods are Full-Cov (94.93%), Hybrid-Cov (94.93%), and MOG4-F (94.39%). Shared-Full-Cov (93.62%) and Shared-Diag-Cov (92.97%) achieve slightly lower accuracies, while Diag-Cov achieves 90.06%. The shrinkage-based methods LW and OAS achieve 89.89% and 86.70% respectively. In this dataset, the methods which use full covariance matrices are the most performant despite the number of training samples per class being relatively low (64 training samples per class). Also, all of the MOG4-F models have only 1 gaussian component.

Also, the proposed methods improve over SimpleCIL, FeCAM, and RanPAC on both CIFAR-100 and ImageNet-R, except for the case of RanPAC in CIFAR-100. In this case, RanPAC achieves significantly higher performance over all the other methods. This can be explained because RanPAC uses parameter-efficient transfer learning (PETL) in the first session using AdaptFormer, and also uses random projections followed by a non-linearity to increase the dimensionality of the features in an efficient non-linear way, improving separability. RanPAC's advantage on CIFAR-100 stems from first-session PETL with AdaptFormer, which adapts the backbone to the specific dataset distribution; our method uses strictly frozen features without any dataset-specific adaptation, making the comparison not fully equivalent in terms of computational investment. Note that use of first session training in RanPAC makes the results dependent on the task presentation order. Finally, the OAS and LW shrinkage variants achieve lower performances than the other Gaussian-based methodologies in all the datasets.

For the case of the proposed methods, it is important to stress that the forgetting values primarily reflect the increased classification difficulty as the number of competing classes grows, rather than the degradation of previously learned representations.

In Table 2, the impact of the backbone is shown on two of the Gaussian-based methods. In the case of CIFAR-100 and ImageNet-R, both Diag-Cov and Full-Cov improve noticeably when using a DINOv2-B/14 backbone. This can be explained both because of the gaussianity of its feature space (discussed in Section 4.4) and because it was trained in an unsupervised way, using a large number of images (142M). However, in the case of Caltech-256, the differences between DINOv2-B/14 and ViT-M/16 are lower, with ViT-M/16 being the most performant backbone in the case of Diag-Cov, despite being trained only with 14M images. This fact can be explained for two reasons: (1) As shown in Sections 4.4.1 and 4.4.2, the Gaussianity of Caltech-256 classes is slightly worse on DINOv2-B/14, and (2) ViT-M/16 generates features with lower dimensionality (512-d versus 768-d). Then, there is a tradeoff between dimensionality and quality of the backbone: lower dimensionality enables using fewer training samples to estimate the gaussians reliably, but a backbone with higher quality enables the gaussians to be better separated in space. The best solution for this trade-off depends both on the method and on the dataset used. Then, the proposed probabilistic modeling framework is closely tied to the representational quality of the backbone, and may further benefit from future advances in large-scale visual representation learning.

In addition, results on the Joint Dataset Benchmark (introduced in this work) are presented. As already mentioned, the Joint Dataset Benchmark is formed by joining 10 tasks from CIFAR-100, ImageNet-R and Caltech-256 into a set of 30 tasks, which are presented in a randomized order. Then, this dataset enables to evaluate methods on extreme distribution shifts between natural, artistic, and high-resolution images, while also containing classes with different number of training samples. As shown in Table 3, Shared-Full-Cov achieves the highest accuracy (81.49%), followed by Shared-Diag-Cov (80.11%) and Diag-Cov (78.32%). In this case Full-Cov, Hybrid-Cov and MOG4-F achieve lower accuracies (59%+) because the number of training samples per class are different. Then, covariance matrices for the different classes have different ranks, and then the squared Mahalanobis distances are not directly comparable between. This result validates the theoretical analysis of Section 3.4: the Joint Benchmark spans classes with 40 to 500 training samples, and the best-performing methods (Shared-Full-Cov, Shared-Diag-Cov, Diag-Cov) are precisely those designed to handle the low-sample / imbalanced regime. The failure of Full-Cov and MOG4-F in this setting is not a weakness but a confirmation that covariance parameterization must match the available sample budget.

4.3 Spectral Analysis of the Foundation Embedding Space

To understand the underlying geometric structure of the foundation embedding space prior to density estimation, we analyze the eigenspectrum of the l_2-normalized representations. Figure 2 illustrates the singular

Table 1: Continual-learning performance on CIFAR-100, ImageNet-R, and Caltech-256. A_{10} and F_{10} in %. "—" indicates no published result on that benchmark.

Method	**CIFAR-100**		**ImageNet-R**		**Caltech-256**	
	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$
L2P Wang et al. (2021)	83.86 ±1.17	7.35 ±0.39	61.57 ±0.40	9.73 ±0.42	92.11 ±0.32	2.94 ±0.26
DualPrompt Wang et al. (2022)	86.51 ±0.49	5.16 ±0.20	68.13 ±0.52	4.68 ±0.18	92.50 ±0.16	2.50 ±0.16
DualPrompt-PGP Qiao et al. (2024)	86.92 ±0.05	5.35 ±0.19	69.34 ±0.05	4.53 ±0.04	—	—
CODA-Prompt Smith et al. (2023)	87.00 ±0.38	4.78 ±0.24	74.24 ±0.56	4.92 ±0.21	—	—
CPrompt Gao et al. (2024)	87.83 ±0.37	4.88 ±0.79	76.71 ±0.61	4.66 ±0.93	—	—
ConvPrompt Roy et al. (2024)	88.87 ±0.33	4.75 ±0.15	77.86 ±0.25	4.33 ±0.24	—	—
RainbowPrompt Hong et al. (2025)	**89.86** ±0.11	**3.44** ±0.26	79.09 ±0.13	3.90 ±0.23	—	—
APT Chen et al. (2025)	88.88 ±0.65	3.47 ±0.55	**79.40** ±0.47	**3.47** ±0.55	—	—
SimpleCIL Janson et al. (2022)	83.70	—	55.56	—	—	—
FeCAM Goswami et al. (2023)	85.70	—	63.70	—	—	—
RanPAC McDonnell et al. (2023)	92.20	—	77.90	—	—	—
Diag-Cov (ours)	86.07 ±0.00	4.30 ±0.12	77.12 ±0.00	7.44 ±0.17	90.06 ±0.00	**1.69** ±0.36
Full-Cov (ours)	88.38 ±0.00	4.12 ±0.00	74.71 ±0.00	5.16 ±0.02	**94.93** ±0.01	1.74 ±0.05
Hybrid-Cov (ours)	88.38 ±0.00	4.18 ±0.00	74.71 ±0.00	5.03 ±0.04	94.93 ±0.00	2.09 ±0.08
Shared-Diag-Cov (ours)	86.87 ±0.00	4.72 ±0.17	75.40 ±0.00	5.17 ±0.38	92.97 ±0.00	2.06 ±0.02
Shared-Full-Cov (ours)	88.72 ±0.00	4.68 ±0.16	77.99 ±0.00	5.87 ±1.44	93.62 ±0.00	2.39 ±0.09
MOG4-F (ours)	88.87 ±0.00	3.99 ±0.00	75.07 ±0.00	5.07 ±0.05	94.39 ±0.00	2.12 ±0.08
OAS (ours)	84.38 ±0.00	3.90 ±0.07	71.47 ±0.00	4.57 ±0.29	86.70 ±0.00	1.90 ±1.77
LW (ours)	84.78 ±0.00	3.93 ±0.09	71.91 ±0.00	4.54 ±0.43	89.89 ±0.00	1.50 ±0.08

Table 2: Impact of backbone architecture on Diag-Cov and Full-Cov across three benchmarks.

Method	**Backbone**	**CIFAR-100**		**ImageNet-R**		**Caltech-256**	
		$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$	$A_{10}(\uparrow)$	$F_{10}(\downarrow)$
Diag-Cov	DINOv2-B/14	86.07 ±0.00	4.30 ±0.12	**77.12** ±0.00	7.44 ±0.17	90.06 ±0.00	1.69 ±0.36
	ViT-M/16	81.41 ±0.00	5.28 ±0.13	66.11 ±0.00	**4.79** ±0.14	**91.41** ±0.00	**1.63** ±0.12
	ViT-B/16	81.03 ±0.00	4.57 ±0.41	43.08 ±0.00	6.41 ±0.19	87.72 ±0.00	1.74 ±0.50
Full-Cov	DINOv2-B/14	**88.38** ±0.00	**4.12** ±0.00	74.71 ±0.00	5.16 ±0.02	94.93 ±0.01	1.74 ±0.05
	ViT-M/16	82.42 ±0.00	5.25 ±0.01	66.44 ±0.01	5.21 ±0.14	92.14 ±0.00	2.53 ±0.12
	ViT-B/16	82.40 ±0.00	5.27 ±0.36	46.23 ±0.00	6.34 ±0.12	87.13 ±0.00	1.92 ±0.80

value decay obtained via Principal Component Analysis (PCA) across the three evaluation benchmarks. The first principal component accounts for approximately 8%–10% of the total variance, followed by a sharp, exponential decay over the first ∼50 dimensions before plateauing. Crucially, despite the high-dimensional $d = 768$ ambient space, fewer than 130 leading components are required to capture 90% of the cumulative variance across all datasets. This rapid spectral decay demonstrates that the highly discriminative signal natively resides within a low-dimensional intrinsic subspace.

Statistical Implication for Small-Sample Rank Deficiency. This sharp variance concentration carries a critical statistical implication for our framework. While the ambient dimensionality of the foundation feature space is fixed at $d = 768$, our analysis demonstrates that the effective rank of the space is closer to

Table 3: Joint Dataset Benchmark (30 tasks: CIFAR-100 + ImageNet-R + Caltech-256, randomized order). A_{30} and F_{30} in %.

Method	$A_{30}(\uparrow)$	$F_{30}(\downarrow)$
Diag-Cov (ours)	78.32 ±0.21	6.05 ±0.12
Full-Cov (ours)	59.35 ±0.00	4.69 ±0.02
Hybrid-Cov (ours)	59.35 ±0.00	**3.86** ±0.04
Shared-Diag-Cov (ours)	80.11 ±0.00	4.89 ±0.05
Shared-Full-Cov (ours)	**81.49** ±0.00	5.49 ±0.05
MOG4-F (ours)	59.68 ±0.25	5.88 ±0.16

$\sim$130. Consequently, even though the feature extractor maps inputs into a high-dimensional space where empirical covariance matrices are structurally rank-deficient due to severe data scarcity ($n_c < d$), the discriminative signal is compressed within a tight active subspace. With a per-class sample budget ranging from $n_c \approx 40$ to 500, the underlying covariance structure within this low-dimensional active subspace can be estimated with reasonable statistical precision, directly justifying why second-order methods remain highly informative despite the overarching rank constraints detailed in Section 3.4.

Cross-Dataset Spectral Consistency. As visible in Figure 2, a notable characteristic of this spectral decay is its extreme invariance across entirely distinct image distributions. The singular value decay curves for Split CIFAR-100, Split ImageNet-R, and Split Caltech-256 overlap almost perfectly. This structural consistency explicitly reveals that DINOv2's feature spacing and variance distributions are highly robust to severe semantic and domain shifts—behaving identically whether processing low-resolution natural images, artistic renditions, or high-resolution object photography. This invariant behavior removes any requirement for dataset-specific dimensionality scaling, validating the use of a uniform probabilistic modeling layer across highly heterogeneous continual tasks.

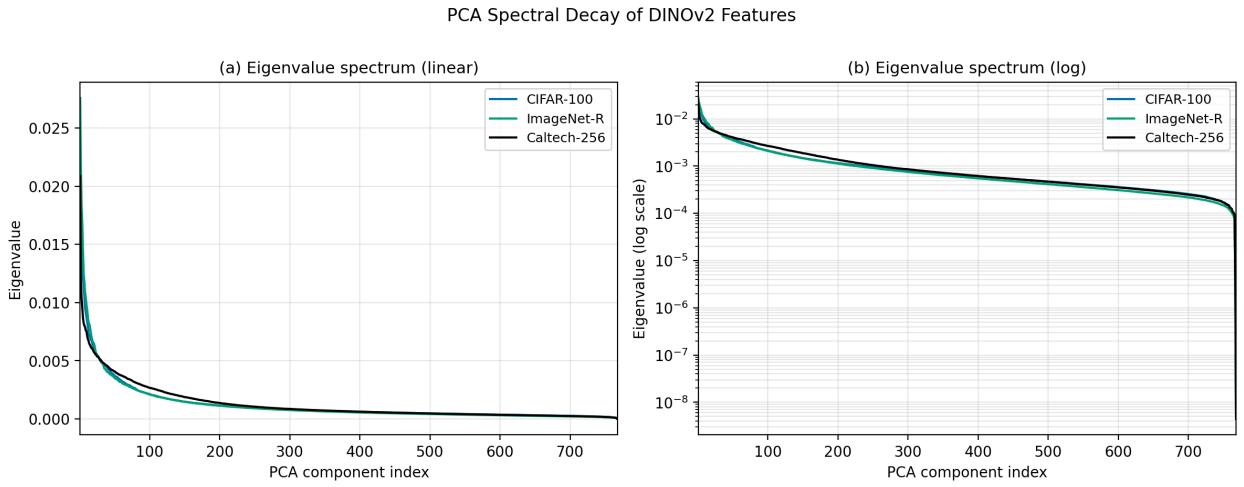


Figure 2: PCA spectral decay of ℓ_2-normalized DINOv2 features across the three benchmarks. Left: Individual per-component explained variance showing a steep initial drop. Right: Cumulative explained variance, highlighting that the 90% threshold is achieved within the first 130 components.

4.4 Analysis of Gaussianity Assumptions

In this section, we examine whether the embedding space of a visual foundation model natively satisfies the Gaussian assumption. To this end, we conduct a rigorous statistical evaluation using the training split of each benchmark. Our analysis specifically focuses on DINOv2; because it is trained via self-supervised learning on a massive, diverse dataset without the semantic constraints of manual labels, it yields a highly continuous, transferable, and well-behaved representation space.

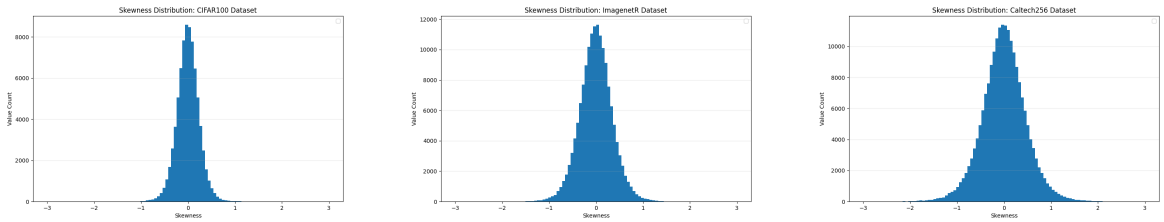
To demonstrate that these self-supervised features are consistently Gaussian, we organize our validation into two structural levels: (1) Univariate Gaussianity: We evaluate the marginal distributions of individual feature dimensions using moment-based diagnostics (skewness and excess kurtosis Pearson (1895)), the D'Agostino–Pearson formal omnibus normality test D'Agostino (1971); D'Agostino & Pearson (1973), and a per-coordinate application of the Henze–Zirkler (HZ) test Henze & Zirkler (1990). (2) Multivariate Gaussianity: We evaluate the joint distribution in its native high-dimensional space by formulating a specialized diagonal-covariance adaptation of the HZ test Henze & Zirkler (1990) to circumvent sample-size rank deficiencies.

4.4.1 Univariate Gaussianity: Moments and Formal Tests

Shape moments. Following Pearson (1895); D'Agostino (1971); D'Agostino & Pearson (1973), we compute sample skewness and excess kurtosis for every (*class* c, *dimension* j) pair:

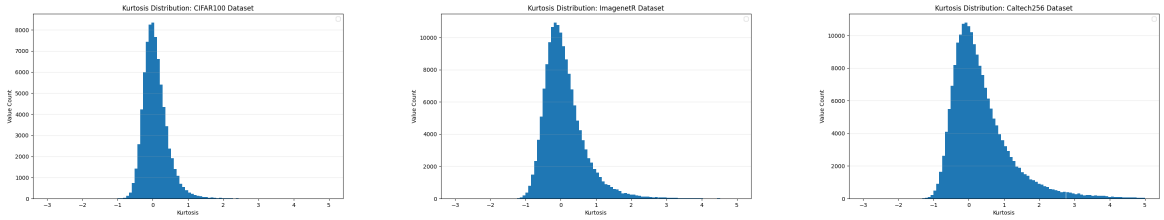
$$\widehat{\mathrm{Skew}}_{c,j} = \frac{\frac{1}{n_c}\sum_i (x_{i,j} - \mu_{c,j})^3}{\hat{\sigma}^3_{c,j}}, \qquad \widehat{\mathrm{Kurt}}_{c,j} = \frac{\frac{1}{n_c}\sum_i (x_{i,j} - \mu_{c,j})^4}{\hat{\sigma}^4_{c,j}} - 3, \tag{15}$$

so that both quantities vanish exactly for a Gaussian. Table 4 reports aggregate statistics. Mean skewness is essentially zero on all benchmarks ($|\overline{\mathrm{Skew}}| \leq 0.002$). The fraction of pairs satisfying $|\mathrm{Skew}| < 0.5$ drops from 96.1% on CIFAR-100 to 74.7% on Caltech-256, reflecting the smaller per-class sample count and greater visual diversity of the latter. Figures 3 and 4 show per-dataset skewness and excess-kurtosis distributions; both concentrate near zero on all three benchmarks.



(a) CIFAR-100 (b) ImageNet-R (c) Caltech-256

Figure 3: Empirical skewness distribution over all ($C \times d$) (class, coordinate) pairs of the frozen DINOv2 embeddings. The distributions are sharply centered near zero on all three benchmarks, confirming that the marginal feature distributions are symmetric about their means.



(a) CIFAR-100 (b) ImageNet-R (c) Caltech-256

Figure 4: Empirical excess-kurtosis distribution over all ($C \times d$) (class, coordinate) pairs of the frozen DINOv2 embeddings. The distributions concentrate near zero, consistent with marginal Gaussianity.

Table 4: Per-dimension Gaussianity statistics over all ($C \times d$) (class, dimension) pairs per dataset ($d = 768$). $|\mathrm{Skew}|_\mu$ and $|\mathrm{Kurt}|_\mu$ are mean absolute values; $\mathbb{P}[\cdot]$ is the fraction of pairs satisfying each cut-off.

Dataset	**pairs**	$\lvert\mathrm{Skew}\rvert_\mu$	Skew_σ	$\mathbb{P}[\lvert\mathrm{Skew}\rvert<0.5]$	$\lvert\mathrm{Kurt}\rvert_\mu$	$\mathbb{P}[\lvert\mathrm{Kurt}\rvert<1]$
CIFAR-100	76,800	0.181	0.236	0.961	0.264	0.982
ImageNet-R	153,600	0.269	0.348	0.859	0.432	0.924
Caltech-256	192,000	0.360	0.472	0.747	0.659	0.822

D'Agostino–Pearson omnibus test. The omnibus test D'Agostino (1971) combines the two shape statistics into:

$$K^2 = Z_1(\widehat{\mathrm{Skew}})^2 + Z_2(\widehat{\mathrm{Kurt}})^2 \;\sim\; \chi^2_2 \quad \text{under } H_0, \tag{16}$$

where Z_1 and Z_2 are normalized transformations of the skewness and kurtosis estimators. We reject H_0 at $\alpha = 0.05$ and report the fraction of pairs passing (Table 5). With $\bar{n}_c \approx 400$, 120, and 90 for CIFAR-100, ImageNet-R, and Caltech-256 respectively, the test is conservative (even minor departures accumulate into

significant K^2 values for the larger datasets); nonetheless 70–81% of all pairs pass the strict test, and the relaxed criterion ($|\text{Skew}| < 1$ and $|\text{Kurt}| < 1$) yields 82–98%.

Table 5: Per-feature normality tests on the training split (fractions of $(C \times d)$ pairs). *normaltest*: D'Agostino–Pearson χ^2_2 at α=0.05. *relaxed*: $|\text{Skew}| < 1$ and $|\text{Kurt}| < 1$ simultaneously.

Dataset	**Normaltest (%)**	**Relaxed (%)**	**Both (%)**
CIFAR-100	70.0	97.8	70.0
ImageNet-R	80.8	91.6	79.7
Caltech-256	73.4	82.3	71.3

Henze–Zirkler univariate test per dimension. As a stricter, hypothesis-based normality criterion, we apply the Henze–Zirkler (HZ) test Henze & Zirkler (1990) independently to every individual coordinate space ($d = 1$) for each class-dimension pair (c, j). Under the null hypothesis H_0 of exact univariate normality, the finite-sample expectation of the statistic is given by $\mathbb{E}_{H_0}[T_{\text{HZ}}] = 1 - a^{-1/2}\left(1 + \frac{b^2}{a} + \frac{3b^4}{4a^2}\right)$, where $a = 1 + 2b^2$ and $b^2 = \frac{1}{2}(4/3)^{0.4} n_c^{-0.4}$ represents the smoothing parameter or bandwidth conditioned on the per-class sample size n_c

Table 6: Summary results from computing the univariate Henze–Zirkler test statistic over all $(C \times d)$ class-dimension pairs on the training split. Here, $\mathbb{E}_{H_0}[T_{\text{HZ}}]$ denotes the exact finite-sample theoretical expectation under the null hypothesis of perfect normality , determined by the per-class sample size n_c via the data-adaptive bandwidth $b^2 = \frac{1}{2}(4/3)^{0.4} n_c^{-0.4}$.

Dataset	$\tilde{n}_c$	$\mathbb{E}_{H_0}[T_{\text{HZ}}]$	T_{HZ} mean $\pm$ std
CIFAR-100	500	≈ 0.04	0.00052 ± 0.00095
ImageNet-R	137	≈ 0.06	0.00134 ± 0.00210
Caltech-256	100	≈ 0.07	0.00234 ± 0.00379

Summary of Univariate Findings. Taken together, these marginal diagnostics reveal that the individual coordinates of the foundation feature vectors exhibit symmetric shapes sharply centered near zero skewness and minimal excess kurtosis. The fact that the observed Henze-Zirkler scores fall orders of magnitude below their theoretical H_0 expected values confirms that the univariate feature dimensions can be safely modeled as independent marginal Gaussians without violating statistical normality assumptions.

4.4.2 Multivariate Gaussianity: Diagonal Henze–Zirkler Test

Per-dimension normality is necessary but not sufficient for the joint d-dimensional distribution to be Gaussian. We extend our analysis to evaluate the distribution in its native $d = 768$ space. Crucially, computing a standard full-covariance Henze–Zirkler (HZ) test Henze & Zirkler (1990) is mathematically intractable in this setting. Because the number of available samples per class is smaller than the feature dimensionality, the empirical full covariance matrix becomes singular and non-invertible, which prevents the calculation of standard Mahalanobis distances.

To circumvent this rank deficiency without resorting to destructive dimensionality reduction techniques (such as PCA), we formulate a diagonal-covariance adaptation of the HZ test. This variant computes distance metrics using explicit per-coordinate sample variances, effectively regularizing the space.

Under exact multivariate normality, $T_{\text{HZ}} \to 1$ as $d \to \infty$ Henze & Zirkler (1990); values close to 1 indicate near-Gaussianity. Table 7 confirms that all three benchmarks concentrate tightly near the Gaussian reference $T_{\text{HZ}} = 1$, validating the Gaussian assumption at the multivariate level under a diagonal model.

Summary of Multivariate Findings. It is important to explicitly clarify that because our modified HZ test enforces a diagonal matrix to handle the small-sample regime, it does not mathematically guarantee full, arbitrary multidimensional Gaussianity where cross-dimensional covariances are fully unconstrained. However, given the severe data limitations per class relative to the embedding dimensions ($n_c < d$), this

Table 7: Summary results from computing the diagonal HZ test statistic Henze & Zirkler (1990) over all classes in the ambient foundation feature space ($d = 768$), using all available training samples. Values concentrated near $T_{\mathrm{HZ}} \approx 1.00$ indicate convergence toward empirical multivariate Gaussianity under the feature-independence assumption.

Dataset	C	T_{HZ} mean $\pm$ std
CIFAR-100	100	1.0002 ± 0.0020
ImageNet-R	200	1.0033 ± 0.0073
Caltech-256	250	1.0016 ± 0.0089

evaluation serves as a highly robust and reliable proxy approximation. The clean multivariate convergence to $T_{\mathrm{HZ}} \approx 1.00$ across all three datasets confirms that the joint space scales properly under independent assumptions, providing a sound statistical justification for deploying independent-feature probabilistic classifiers in visual foundation spaces.

4.5 Inter-Class Separability via Bhattacharyya Distance

To evaluate the statistical limits of our model and understand why performance scales with the number of classes, we examine the structural separability of the embedding space. Specifically, to quantify how distinguishable any two class-conditional Gaussians are from one another, we compute pairwise *Bhattacharyya distances*. Unlike simple geometric metrics, the Bhattacharyya distance is symmetric and has a direct, closed-form connection to the theoretical Bayes error rate via the Bhattacharyya coefficient $\rho_{ij} = e^{-D_B(i,j)}$, allowing us to rigorously map distributional overlap to actual classification risk. Under our independent feature assumption, this measure simplifies significantly. Specifically, under the diagonal-Gaussian parameterisation, the distance is defined as:

$$D_B(i,j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \bar{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\ln\left(\frac{\det \bar{\boldsymbol{\Sigma}}}{\sqrt{\det \boldsymbol{\Sigma}_i \, \det \boldsymbol{\Sigma}_j}}\right), \tag{17}$$

where $\bar{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)/2$. Because both covariance matrices are diagonal, all operations reduce to element-wise arithmetic:

$$\det \boldsymbol{\Sigma} = \textstyle\prod_d \sigma_d^2, \qquad \bar{\boldsymbol{\Sigma}}^{-1} = \mathrm{diag}\left(\frac{2}{\sigma_{i,d}^2 + \sigma_{j,d}^2}\right).$$

This keeps computation in $\mathcal{O}(C \cdot d)$ per class pair, avoiding dense matrix inversions.

We characterize inter-class separability through two complementary lenses: the *Bhattacharyya distance* D_B, which jointly accounts for differences in both means and covariances, and the plain *Euclidean distance* $d_E = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2$ between class mean vectors, which isolates mean separation alone.

Bhattacharyya distance. Figure 5 characterizes D_B as the class set grows. The mean pairwise D_B tends to remain stable when the number of training classes increases, at full class-set size: 89.43 (CIFAR-100), 84.81 (ImageNet-R), 220.61 (Caltech-256). This indicates that, on average, the distances between gaussians remain almost constant. However, the minimum pairwise D_B decreases sharply with the number of classes, with minimum pairwise D_B: 3.57 (CIFAR-100), 11.54 (ImageNet-R), and 20.96 (Caltech-256), which shows that there appear overlapping classes when new classes are added. Figure 6 shows that the number of pairs with low D_B increases when more classes are added. Pairs with $D_B < 50$ represent only 4.77% (CIFAR-100), 10.40% (ImageNet-R), and 0.15% (Caltech-256), confirming that the vast majority of class pairs are well-separated in the probabilistic sense. However, the amount of pairs with $D_B < 50$ is related to the accuracy that can be achieved, which for Gaussian methods is 88.87% (CIFAR-100), 77.99% (ImageNet-R) and 94.36% (Caltech-256). Then, it can be concluded that the existence of overlapping classes is the cause of the drop of accuracy in the Gaussian methods.

Euclidean distance between class means. Figures 7 and 8 show the analogous curves for d_E. Mean d_E also tends to remain stable as the class set grows, at full class-set size: 0.749 (CIFAR-100), 0.727 (ImageNet-R), 0.978 (Caltech-256). Minimum pairwise d_E decreases sharply with the number of classes, achieving:

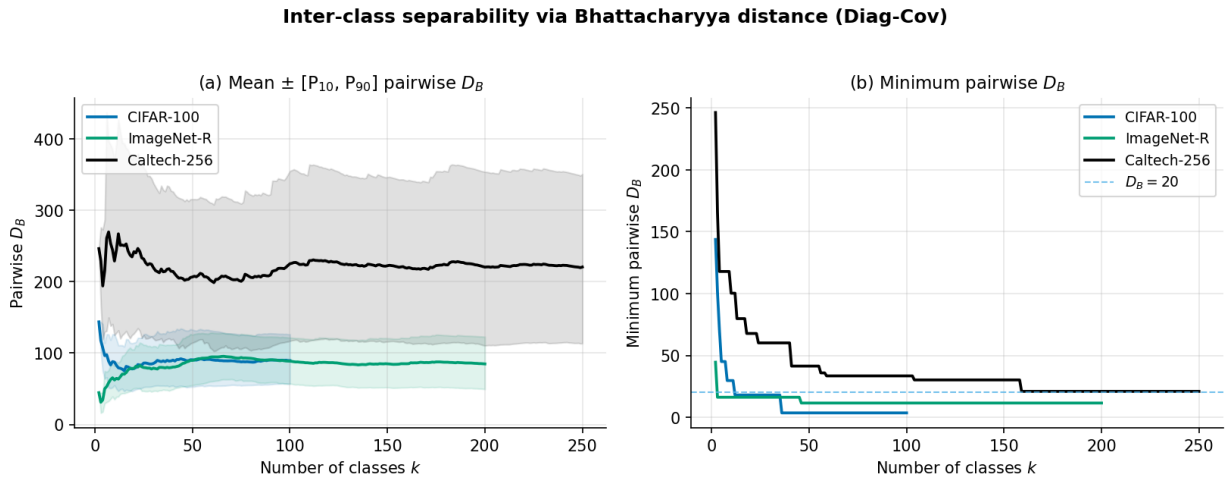


Figure 5: Inter-class separability under the Diag-Cov Gaussian model as the class set grows. **(a)** Mean pairwise Bhattacharyya distance D_B with shaded $[\mathrm{P}_{10}, \mathrm{P}_{90}]$ band. **(b)** Minimum pairwise D_B; the dashed line marks $D_B = 20$.

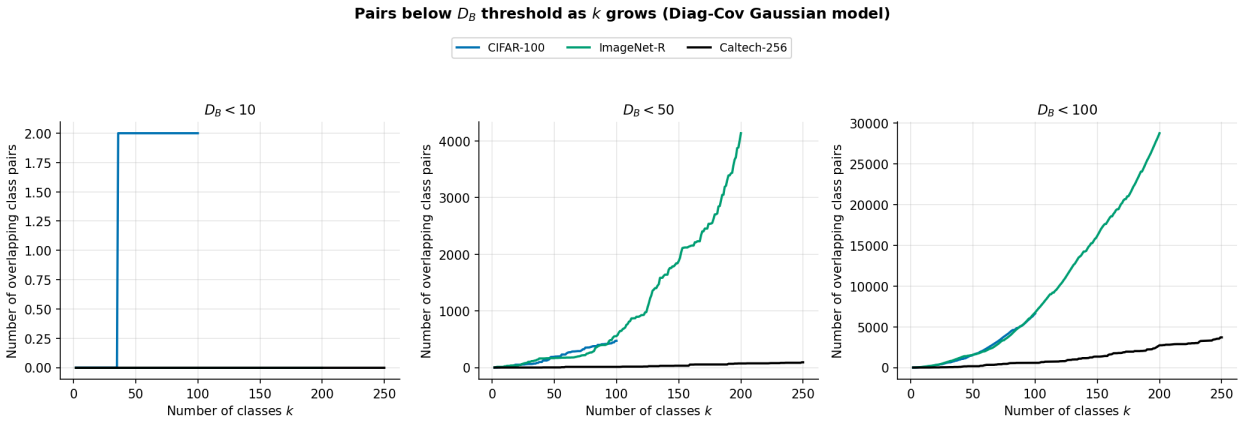


Figure 6: Count of class pairs below three overlap thresholds ($D_B < 10$, $D_B < 50$, $D_B < 100$) as the class set grows.

0.144 (CIFAR-100), 0.205, (ImageNet-R) and 0.275 (Caltech-256). Pairs with $d_E < 0.5$ represent only 0.91% (CIFAR-100), 2.20% (ImageNet-R), and 0.04% (CIFAR-100) of the pairs. Then, the analysis done before between pairwise D_B values and accuracies also extends to d_E, indicating that the drop of accuracy is caused by overlapping classes — the rare, visually similar pairs that also carry the lowest D_B values. The two metrics are complementary: d_E isolates mean separation in the embedding space while D_B additionally penalizes covariance dissimilarity, which is why D_B ranks pairs such as *boy/girl* as the most confused (small $d_E = 0.145$ *and* overlapping covariances), while d_E is more interpretable as a geometric separation threshold.

Most-Confused Classes and Error Analysis. Both distance metrics converge on the same structurally difficult class pairs, providing deep insight into the true failure modes of the model. Table 8 lists the five lowest-recall classes per benchmark under the `Diag-Cov` parameterization. To quantify these error modes precisely, we report the main confusion target along with the confusion percentage (Conf.%), which indicates the proportion of misclassified samples from the base class that were erroneously assigned to the dominant confusion target.

Furthermore, we contrast the localized pair metrics against the global statistics of each dataset. Specifically, we compare the pairwise Bhattacharyya distance (D_B) against the dataset's average macro-pairwise Bhat-

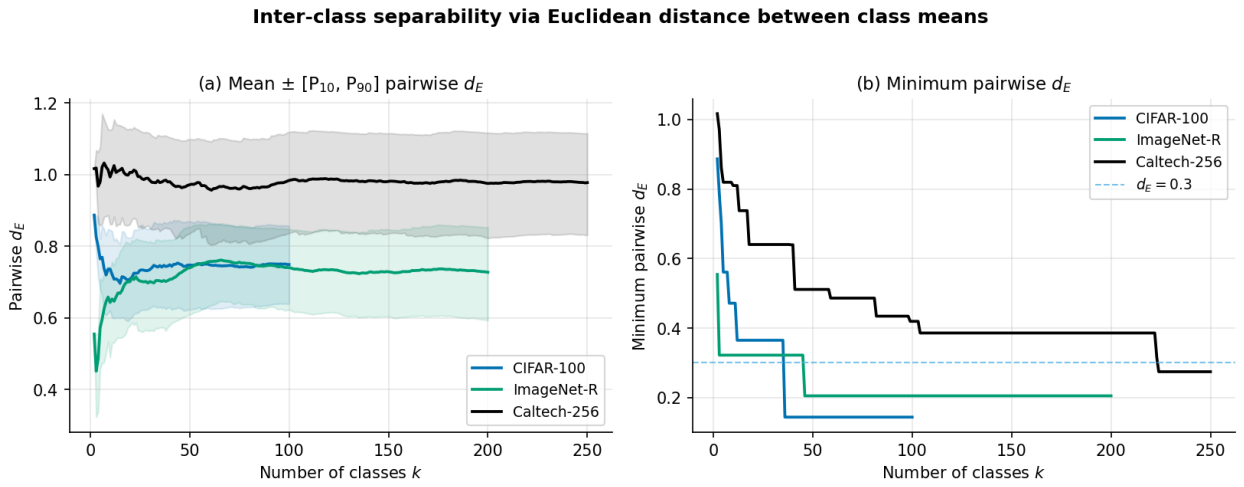


Figure 7: Inter-class separability via Euclidean distance between class mean vectors as the class set grows. **(a)** Mean pairwise d_E with shaded $[\mathrm{P}_{10}, \mathrm{P}_{90}]$ band. **(b)** Minimum pairwise d_E; the dashed line marks $d_E = 0.3$.

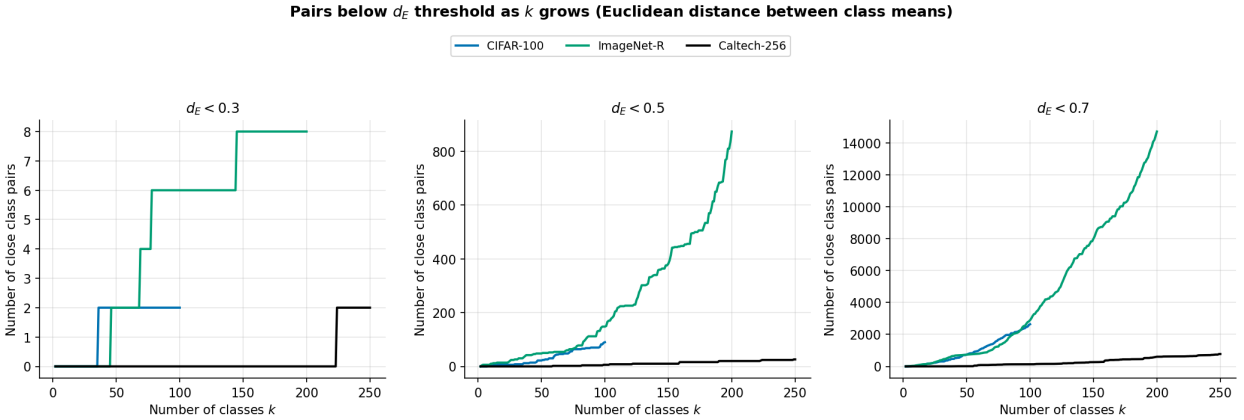


Figure 8: Count of class pairs below three Euclidean distance thresholds ($d_E < 0.3$, $d_E < 0.5$, $d_E < 0.7$) as the class set grows.

tacharyya distance ($\overline{D_B}$), and the pairwise Euclidean distance between class mean vectors (d_E) against the macro-average mean distance ($\overline{d_E}$).

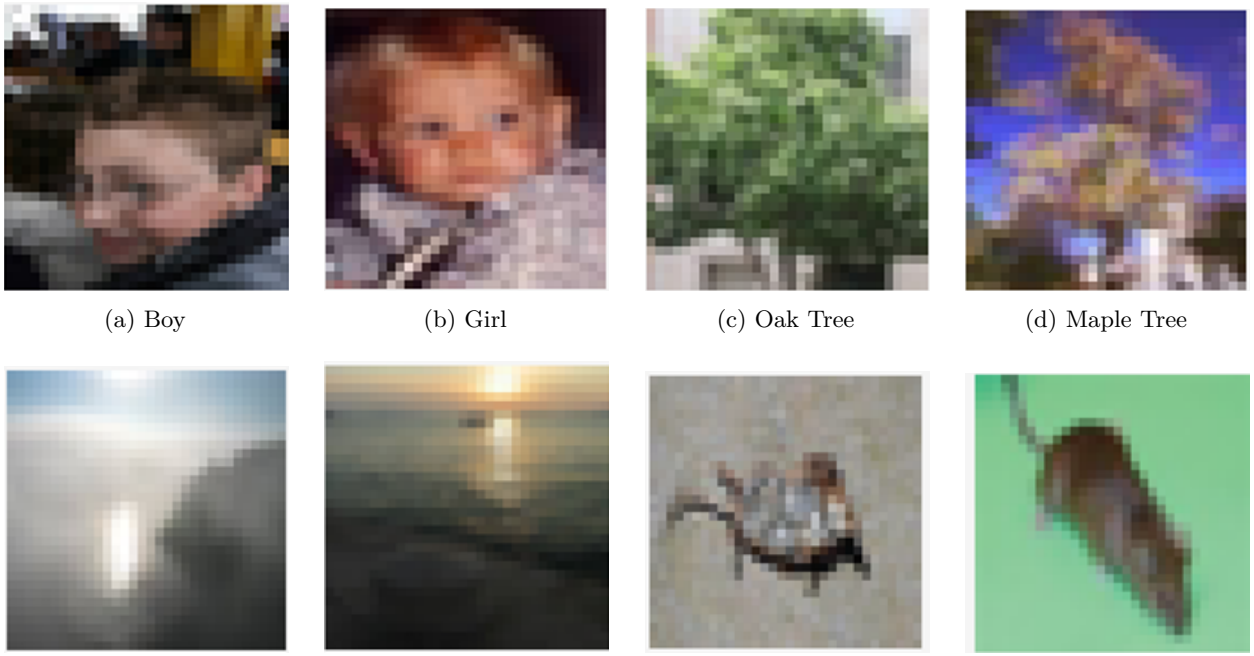
Every hard pair consistently exhibits values that are orders of magnitude lower than their respective dataset averages. For instance, we observe extreme representation overlap in *boy/girl* ($D_B = 3.45$ vs. $\overline{D_B} = 89.43$; $d_E = 0.145$ vs. $\overline{d_E} = 0.749$), *bow_tie/scarf* ($D_B = 16.57$ vs. $\overline{D_B} = 84.81$), and *top-hat/cowboy-hat* ($D_B = 29.75$ vs. $\overline{D_B} = 220.61$). This drastic contrast confirms that severe localized distributional overlap—rather than task ordering or parameter degradation—is the primary driver of classification error in this framework. Figures 9–11 display the representative images closest to each class mean in the DINOv2 space, visually illustrating the inherent semantic and visual ambiguities that dominate these failure points.

4.6 Computational Requirements

Gaussian model evaluation mainly involves matrix–vector multiplications. For Diag-Cov, the total cost is approximately 0.46 MFLOPs on CIFAR-100, 0.92 MFLOPs on ImageNet-R, and 1.15 MFLOPs on Caltech-256. For Full-Cov, the corresponding costs are approximately 0.12, 0.24, and 0.29 GFLOPs. These are negligible compared to the $\approx$ 17.56 GFLOPs required for a forward pass of DINOv2 ViT-B/14, so the classification stage introduces only marginal overhead relative to feature extraction.

Table 8: Analysis of the five lowest-recall classes on each benchmark under `Diag-Cov`. Conf.% represents the percentage of errors directed toward the main confusion target. D_B and d_E denote the pairwise Bhattacharyya and Euclidean distances to the main confusion target, respectively, while $\overline{D_B}$ and $\overline{d_E}$ represent the macro-averaged pairwise distances across the entire dataset split. Lower localized values relative to the macro-averages indicate greater distributional overlap.

Dataset	**Class**	**Acc%**	**Main Confusion**	**Conf.%**	D_B	$\overline{D_B}$	d_E	$\overline{d_E}$
Split CIFAR-100	*boy*	50.0	*girl*	30.0	3.45	89.43	0.145	0.749
	maple_tree	60.0	*oak_tree*	19.0	19.41	89.43	0.327	0.749
	girl	65.0	*boy*	23.0	3.45	89.43	0.145	0.749
	cloud	65.0	*sea*	8.0	30.54	89.43	0.451	0.749
	shrew	69.0	*mouse*	16.0	16.45	89.43	0.331	0.749
Split ImageNet-R	*cabbage*	35.3	*vase*	17.6	50.50	84.81	0.587	0.727
	saint_bernard	36.4	*basset_hound*	9.1	66.44	84.81	0.643	0.727
	bow_tie	39.3	*scarf*	21.4	16.57	84.81	0.331	0.727
	beaver	40.0	*backpack*	8.6	34.74	84.81	0.500	0.727
	bee	41.6	*fly*	15.6	20.51	84.81	0.378	0.727
Split Caltech-256	*top-hat*	31.2	*cowboy-hat*	31.2	29.75	220.61	0.319	0.978
	paperclip	52.6	*drinking-straw*	21.1	62.94	220.61	0.645	0.978
	canoe	57.1	*people*	28.6	57.87	220.61	0.641	0.978
	drinking-straw	64.7	*people*	11.8	27.33	220.61	0.465	0.978
	traffic-light	65.0	*people*	15.0	66.27	220.61	0.675	0.978



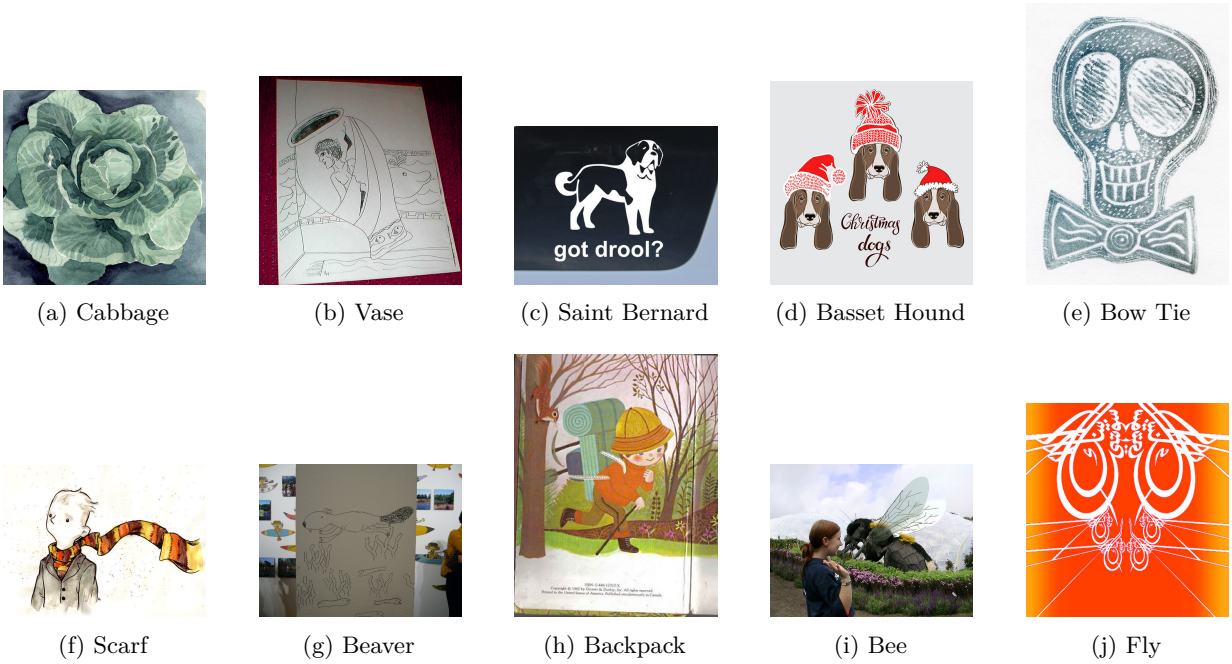
(a) Boy (b) Girl (c) Oak Tree (d) Maple Tree

(e) Cloud (f) Sea (g) Shrew (h) Mouse

Figure 9: Lowest-recall classes on CIFAR-100 benchmark according to Table 8.

5 Conclusion

We formulated class-incremental learning as the recursive estimation of class-conditional Gaussian densities in stationary foundation embeddings, eliminating the need for backpropagation, exemplar storage, or parametric prompt tuning. To handle data scarcity and rank deficiency in high-dimensional feature regimes, we introduced the Hybrid-Cov estimator, which decouples the Mahalanobis shape term from the log-normalization constant to provide a numerically stable probabilistic classifier. Finally, we provided a rigorous statisti-



(a) Cabbage (b) Vase (c) Saint Bernard (d) Basset Hound (e) Bow Tie

(f) Scarf (g) Beaver (h) Backpack (i) Bee (j) Fly

Figure 10: Lowest-recall classes on ImageNet-R benchmark according to Table 8. Classes shown side by side with their main confusion target.

(a) Top Hat (b) Cowboy Hat (c) Paperclip (d) Drinking Straw

(e) Canoe (f) Traffic Light (g) People

Figure 11: Lowest-recall classes on Caltech-256 benchmark according to Table 8.

cal validation of this framework using univariate and multivariate normality tests alongside comprehensive Bhattacharyya distance separability analyses across three distinct vision benchmarks

Empirically, our full-covariance configurations and MOG4-F match the competitive performance of ConvPrompt on CIFAR-100 (88.87%) without utilizing any trainable parameters during incremental tasks, while significantly outperforming all prompt-based baselines on Caltech-256 (94.93%). This competitive classification is achieved at a negligible computational cost below 0.3 GFLOPs—orders of magnitude below the forward pass of the underlying foundation backbone. Furthermore, our Bhattacharyya separability analysis confirms that over 99.8% of all class pairs exceed a distance threshold of $D_B = 50$ on every benchmark,

providing a robust, principled probabilistic explanation for the strong classification performance observed throughout sequential streams.

These results suggest that, in the visual foundation model era, the central challenge of class-incremental learning shifts from protecting network parameters against catastrophic forgetting to properly modeling and exploiting the statistical geometry of stable representation spaces. The remarkable success of classical density estimation within this modern context is not a coincidence: it is made possible by a qualitative change in the topological structure of embedding spaces that large-scale self-supervised pretraining produces. By decoupling representation plasticity from semantic categorization, we demonstrate that stable, well-behaved foundation features eliminate representation drift by construction, allowing lifelong recognition to be reframed as an ongoing probability density modeling problem.

Moving forward, future work will focus on two key areas. First, we aim to develop variants that reduce sensitivity to the rank of the covariance matrices while fully preserving the information provided by the training samples. Second, we will investigate representation alignment methods designed to improve feature-space separability, specifically targeting the small minority of semantically overlapping and visually ambiguous class pairs (e.g., boy/girl, top-hat/cowboy-hat) that our error analysis identified as the primary failure mode of the current framework.

Acknowledgements

Please insert your acknowledgments here.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, pp. 144–161, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01218-2. doi: 10.1007/978-3-030-01219-9_9. URL https://doi.org/10.1007/978-3-030-01219-9_9.

Eden Belouadah and Adrian Daniel Popescu. Il2m: Class incremental learning with dual memory. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 583–592, 2019. URL https://api.semanticscholar.org/CorpusID:204923710.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 532–547, 2018.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

Haoran Chen, Zuxuan Wu, Xintong Han, Menglin Jia, and Yu-Gang Jiang. Promptfusion: Decoupling stability and plasticity for continual learning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVII*, pp. 196–212, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73020-7. doi: 10.1007/978-3-031-73021-4_12. URL https://doi.org/10.1007/978-3-031-73021-4_12.

Haoran Chen, Ping Wang, Zihan Zhou, Xu Zhang, Zuxuan Wu, and Yu-Gang Jiang. Achieving more with less: Additive prompt tuning for rehearsal-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 340–349, October 2025.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Ralph B. D'Agostino. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2): 341–348, 1971.

Ralph B. D'Agostino and E. S. Pearson. Tests for departure from normality. empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika*, 60(3):613–622, 1973.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28463–28473, 2024. URL `https://api.semanticscholar.org/CorpusID:268379362`.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost Van De Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36:6582–6595, 2023.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech 256, 2022. doi: 10.22002/D1.20087.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00823. URL `https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00823`.

Norbert Henze and Bettina Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics – Theory and Methods*, 19(10):3595–3617, 1990. doi: 10.1080/03610929008830400.

Kiseong Hong, Gyeong-hyeon Kim, and Eunwoo Kim. Rainbowprompt: Diversity-enhanced prompt-evolving for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140, October 2025.

Steven C. Y. Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. *Compacting, picking and growing for unforgetting continual learning*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Paul Janson, Wenxuan Zhang, Rahaf Aljundi, and Mohamed Elhoseiny. A simple baseline that questions the use of pretrained models in continual learning. In *Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems*, 2022.

Hyundong Jin and Eunwoo Kim. Helpful or harmful: Inter-task association in continual learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pp. 519–535, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20082-3. doi: 10.1007/978-3-031-20083-0_31. URL `https://doi.org/10.1007/978-3-031-20083-0_31`.

Sangwon Jung, Hongjoon Ahn, Sungmin Cha, and Taesup Moon. Continual learning with node-importance based adaptive group sparse regularization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, December 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2773081. URL https://doi.org/10.1109/TPAMI.2017.2773081.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6470–6479, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Mark D. McDonnell, Dong Gong, Amin Parveneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10107–10123, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification. Code available at: https://github.com/facebookresearch/dinov2.

Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jähnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11313–11321, 2019. doi: 10.1109/CVPR.2019.01158.

Karl Pearson. Contributions to the mathematical theory of evolution. II. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London A*, 186:343–414, 1895.

Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3911–3920, 2023.

Jingyang Qiao, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yong Peng, and Yuan Xie. Prompt gradient projection for continual learning. In *International Conference on Learning Representations*, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, G. Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2016. URL https://api.semanticscholar.org/CorpusID:206596260.

Anurag Roy, Riddhiman Moulick, Vinay K. Verma, Saptarshi Ghosh, and Abir Das. Convolutional prompting meets language models for continual learning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23616–23626, 2024. URL https://api.semanticscholar.org/CorpusID:268793650.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 2994–3003, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

R.C. Van Sluyters, J. Atkinson, M.S. Banks, R.M. Held, K.-P. Hoffmann, and C.J. Shatz. *Visual Perception: The Neurophysiological Foundations*, chapter The Development of Vision and Visual Perception. Academic Press, San Diego, CA, 1990.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11909–11919, 2023.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 139–149, 2021. URL https://api.semanticscholar.org/CorpusID:245218925.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pp. 631–648, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19808-3. doi: 10.1007/978-3-031-19809-0_36. URL https://doi.org/10.1007/978-3-031-19809-0_36.

Ross Wightman. Pytorch image models, 2019. URL https://github.com/huggingface/pytorch-image-models.

Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Memory replay gans: learning to generate images from new categories without forgetting. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 5966–5976, Red Hook, NY, USA, 2018. Curran Associates Inc.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3013–3022, 2021. URL https://api.semanticscholar.org/CorpusID:232427923.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.