SCALE: SUPPORT CALIBRATION WITH ANCHORED LOCAL EXPERTS FOR IMBALANCED TIME-SERIES CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rich frozen time-series representations provide many feature directions, but scarce minority labels may be insufficient to identify those that generalize. Our diagnostics reveal a gap between training and test recall for the minority class under class-balanced readouts. With the encoder held fixed, readouts trained with more labeled examples achieve higher ROC-AUC. Thus, readout performance under sparse supervision can fall short of the discrimination supported by the same frozen representations. Motivated by these observations, we propose Support Calibration with Anchored Local Experts (SCALE) to improve minority generalization without additional labels or encoder updates. SCALE uses shared local references to construct compact residual coordinates and fits complementary experts on observed and modeled support. These experts propose minority-directed corrections to a class-balanced anchor, with training cross-validation selecting the correction rule. Across 690 univariate encoder–task pairs, SCALE achieves the highest aggregate balanced accuracy, Macro-F1, and Min-F1 among the compared methods. It improves on cRT by 2.06 balanced-accuracy and 3.43 Min-F1 percentage points, with larger gains under more severe imbalance in a matched cohort. Multivariate evaluation also shows positive aggregate gains over cRT.

1 INTRODUCTION

An informative pretrained time-series representation does not by itself make minority predictions reliable. With few minority labels, a downstream classifier must distinguish class-relevant directions from task-irrelevant ones. Sparse supervision can leave even a frozen-feature classifier vulnerable to overfitting. Under Gaussian-mixture assumptions, Lyu et al. (2026) characterize disproportionate minority overfitting in high-dimensional linear classification.

To examine how minority fitting translates into generalization, we evaluate class-balanced readouts on frozen pretrained time-series representations (Figure 1). Across 13 binary datasets and four pretrained encoders, minority training recall rises as support shrinks, while test recall falls (c). With the encoder held fixed, reference readouts trained with the original TRAIN labels achieve higher ROC-AUC in 51 of 52 pairs at 20:1 (d). Because changing a decision threshold does not alter ROC-AUC, this difference reflects a gap in ranking performance. The FordB example illustrates the corresponding gap between minority training fit and test recall (a–b). Together, these observations indicate that performance under sparse supervision can fall short of the discrimination supported by the same frozen representations. This motivates us to ask how to improve minority generalization without additional labels or encoder updates.

Class balancing and score correction address the influence of class frequencies on fitting and prediction (Kang et al., 2020; Menon et al., 2021; Bhat et al., 2025). Oversampling models minority variation through synthetic examples (Chawla et al., 2002; Zhao et al., 2022). Building on these approaches, we use shared structure in frozen representations to define local reference coordinates, and combine observed and modeled support to guide minority corrections.

We organize adaptation around three choices. First, all training features can provide local references even when minority labels are few; signed displacements from these references give labels a compact coordinate for learning class distinctions. Second, observed and modeled minority support provide

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

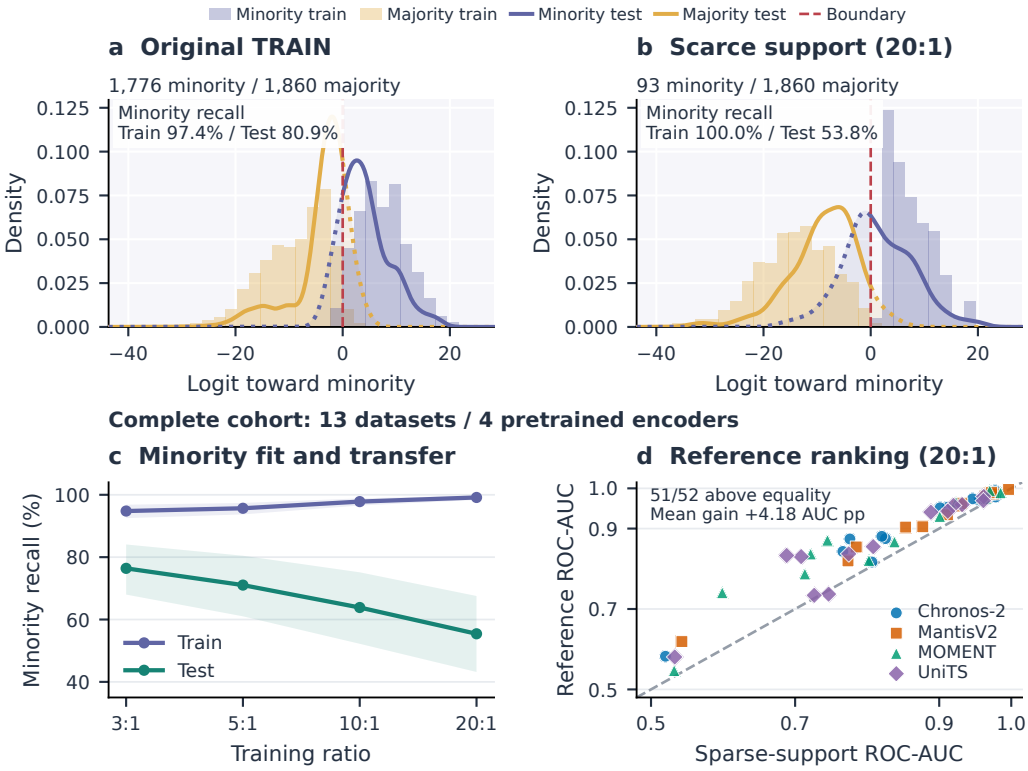


Figure 1: Minority fitting and test generalization in frozen representations. (a–b) An illustrative FordB–MantisV2 comparison: original TRAIN versus 20:1 support, with shared axes. Purple/gold denote minority/majority; histograms/curves show class-conditional train/test densities, and dotted segments lie on the misclassified side. (c) Minority recall across four ratios on 13 binary datasets and four encoders; bands are 95% dataset-bootstrap intervals. (d) All 52 pairs at 20:1: the reference uses additional TRAIN labels. Further details are provided in Appendix I.

alternative views for fitting classifiers. Third, because each view is estimated from limited minority observations, we retain a class-balanced anchor fitted to observed data and use the other classifiers to propose minority-directed corrections. Training cross-validation selects how much expert agreement is required, including an option to retain the anchor.

Support Calibration with Anchored Local Experts (SCALE) implements these choices (Figure 2). Local Reference Geometry (LRG) appends compact supervised residual coordinates. Spectral support expansion (SSE) fits a regularized low-rank generator whose samples enter only one expert. The anchor and remaining experts use observed data. All experts are fitted globally; their votes propose corrections toward the designated minority.

Across five frozen time-series representations, SCALE improves aggregate classification performance, with larger gains under more severe imbalance in a matched cohort. Coordinate comparisons and independently retuned ablations support the use of local residual coordinates together with support expansion. Code, experimental results, and datasets are available in our anonymous repository.[1]

2 Related Work

Learning with imbalanced data. SMOTE interpolates minority observations (Chawla et al., 2002); T-SMOTE incorporates temporal structure (Zhao et al., 2022). Integrated Oversampling (INOS) combines interpolation with eigenspectrum-regularized Gaussian generation (Cao et al., 2013).

[1] https://anonymous.4open.science/r/SCALE-DD85

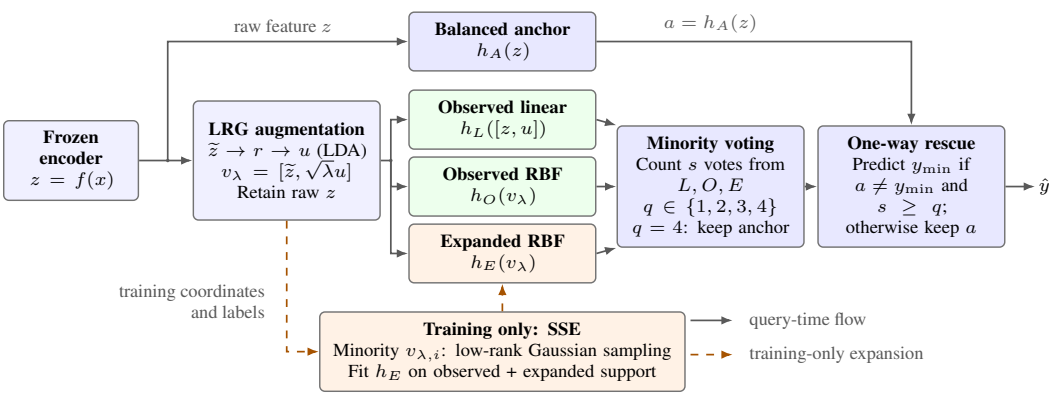


Figure 2: SCALE pipeline. The anchor uses z, the linear expert uses $[z, u]$, and both RBF experts use $v_\lambda = [\widetilde{z}, \sqrt{\lambda}u]$. At least q of the three experts must vote for the minority class for a rest-to-minority correction; $q = 4$ keeps the anchor. Solid arrows show query-time flow; dashed arrows show training-only SSE fitted in v_λ.

Embedding-space oversampling operates after representation learning (Dablain et al., 2023b), while DeepSMOTE couples interpolation with encoder–decoder training (Dablain et al., 2023a). Training objectives include focal and class-balanced losses (Lin et al., 2017; Cui et al., 2019), Balanced Softmax, LDAM (Ren et al., 2020; Cao et al., 2019), and balanced contrastive learning (Zhu et al., 2022). SCALE operates after the encoder is frozen.

Frozen-feature and decision-level correction. Classifier retraining separates representation learning from class balancing (Kang et al., 2020). Threshold tuning changes the minority operating point; logit adjustment corrects the training prior (Menon et al., 2021), and P2P estimates an effective model prior (Bhat et al., 2025). These are our closest comparisons. Our cRT baseline uses class-weighted logistic regression; SCALE supplements its global boundary with local coordinates and alternative support estimates.

Local geometry and classifier ensembles. Prototype classifiers (Snell et al., 2017), radial-basis features (Broomhead & Lowe, 1988), and locality-constrained coding (Wang et al., 2010) use reference points or neighborhoods to structure prediction and representation. Ensemble methods such as HIVE-COTE combine classifiers according to estimated predictive quality (Middlehurst et al., 2021). Our focus is minority adaptation within frozen time-series representations. SCALE uses local residual coordinates and complementary support estimates to guide minority-directed corrections to a class-balanced anchor.

3 Problem Formulation and SCALE

3.1 Frozen-Feature Imbalance Repair

Let f be a fixed encoder, $z = f(x) \in \mathbb{R}^d$ its feature, and $\mathcal{D}_{\text{tr}} = \{(z_i, y_i)\}_{i=1}^n$ the labeled downstream training set with classes $\mathcal{Y} = \{1, \ldots, C\}$. We designate the smallest training class as $y_{\min}$ and retain all remaining classes as distinct labels. The target criterion is class-balanced risk,

$$\mathcal{R}_{\text{bal}}(h) = \frac{1}{C} \sum_{c=1}^{C} \mathbb{P}(h(z) \neq c \mid y = c)\,, \tag{1}$$

which gives every class equal influence irrespective of its observed frequency. With the encoder frozen, SCALE fits its geometry, support models, and classifiers using $\mathcal{D}_{\text{tr}}$ and selects task-specific settings by training CV. Here, support calibration refers to adjusting classification decisions through observed and modeled support.

A class-weighted logistic anchor h_A fits all observed features and supplies the default multiclass decision. SCALE first adds local residual coordinates, then fits experts on observed and expanded

support, and finally restricts their votes to minority-directed corrections. These steps specify the feature representation, support model, and permitted changes to the anchor.

3.2 Local Reference Geometry

Local Reference Geometry (LRG) expresses how a feature departs from a shared local reference. Distances and affinities summarize proximity; a signed residual also retains the direction of departure. LRG projects this residual r into a compact supervised coordinate u for SCALE's experts.

Column standardization first removes coordinate-specific location and scale, after which row-wise ℓ_2 normalization places Euclidean neighborhood construction on feature direction rather than raw feature magnitude. We denote the resulting geometry coordinate by $\widetilde{z}$. Column-standardization parameters are fitted on the training split; row normalization is applied to each sample independently. The same mapping is used at evaluation; the anchor and linear expert still retain the original z. We then fit ordinary k-means with training-selected region count K (Lloyd, 1982), capped by the fitting-set size as specified in Appendix D. Region k stores center c_k and robust radius ρ_k, the median distance of assigned training features to c_k, bounded below by 10^{-3}. The regions use all training features without labels; class information is introduced only after the local coordinate has been formed. These centers are reference locations rather than minority prototypes: a rest-dominated region can still provide a useful frame for a discriminative minority displacement. Centers and scalar radii also avoid estimating a separate high-dimensional covariance from each sparsely populated region.

For a query, the soft affinity and local reference are

$$a_k(\widetilde{z}) = \frac{\exp(-\|\widetilde{z} - c_k\|_2^2/\tau)}{\sum_j \exp(-\|\widetilde{z} - c_j\|_2^2/\tau)}, \tag{2}$$

$$\mu(\widetilde{z}) = \sum_k a_k(\widetilde{z}) c_k, \qquad \sigma(\widetilde{z}) = \sum_k a_k(\widetilde{z}) \rho_k, \tag{3}$$

with fixed temperature $\tau = 1$ in the main experiments. The signed residual is

$$r(\widetilde{z}) = \frac{\widetilde{z} - \mu(\widetilde{z})}{\max(\sigma(\widetilde{z}), 10^{-3})}. \tag{4}$$

We fit shrinkage LDA globally to the training residuals and labels and standardize the selected projections:

$$u(z) = D_s^{-1}\big(W^\top r(\widetilde{z}) - m_{\mathrm{tr}}\big)\,. \tag{5}$$

In both benchmarks, the requested dimension is selected from $\{1, 2\}$ and capped at $C - 1$ and the feature dimension. The mean m_{tr} and diagonal D_s use training projections only; degenerate fits may return an empty block. Thus u appends a compact supervised coordinate without replacing z. For example, a linear expert on $[z, u]$ has class score

$$s_c(z) = w_c^\top z + q_c^\top u(z) + b_c. \tag{6}$$

The first term preserves the global frozen representation, while the second provides a class-specific correction indexed by signed local displacement. This correction retains the direction of local displacement, which scalar density or distance summaries discard.

In the hard-assignment limit, the residual term gives a region-dependent scale and offset to the score (Appendix B.1). Soft affinities smooth this dependence. Augmentation enlarges the score family; Appendix B.4 compares it with simpler local descriptors.

3.3 Expert Construction and Support Expansion

Alongside the anchor, SCALE fits a class-weighted linear expert $h_L(z, u)$ and an RBF-SVM $h_O(v_\lambda)$. Both use observed training support. The RBF expert uses

$$v_\lambda(z) = [\widetilde{z}, \sqrt{\lambda}\, u(z)], \qquad d_\lambda^2(z, z') = \|\widetilde{z} - \widetilde{z}'\|^2 + \lambda \|u(z) - u(z')\|^2. \tag{7}$$

The training-selected residual weight λ controls its contribution to nonlinear distances. At $\lambda = 0$, the nonlinear branches remove the residual columns; the local linear expert still retains $[z, u]$.

Spectral support expansion (SSE) models minority variation to provide an additional support view. The expanded RBF expert h_E combines synthetic coordinates with observed samples, while h_A, h_L, and h_O use observed samples only. Let the minority coordinates in v_λ have mean $\bar{v}_\lambda$ and retained right-singular basis $V_{\lambda,r}$. The retained rank is capped by $n_{\min} - 1$ and the coordinate dimension. We estimate an Oracle Approximating Shrinkage (OAS) covariance (Chen et al., 2010), $\widehat{\Sigma}_{\lambda,r}$, from the projected minority scores and sample

$$v^+ = \bar{v}_\lambda + V_{\lambda,r}\xi, \qquad \xi \sim \mathcal{N}(0, \widehat{\Sigma}_{\lambda,r}). \tag{8}$$

Gaussian expansion fills the minority count deficit. In both benchmarks, training CV selects the maximum rank from $\{8, 16\}$ together with the residual weight and classifier settings (Appendix D). For each candidate coordinate, SSE refits its basis and covariance to the minority training observations. Generated coordinates sample this shared support model.

The anchor and local linear expert share logistic regularization, and the observed and expanded RBF experts share both RBF regularization and the bandwidth multiplier. The support-head fitting conventions are specified in Appendix D.

3.4 One-Way Minority Correction

The experts can change an anchor prediction only toward the designated minority class. The vote threshold determines how many proposals are needed for this action. Let the minority-vote count be

$$s(z) = \mathbf{1}\{h_L(z) = y_{\min}\} + \mathbf{1}\{h_O(z) = y_{\min}\} + \mathbf{1}\{h_E(z) = y_{\min}\}, \qquad g_q(z) = \mathbf{1}\{s(z) \geq q\}. \tag{9}$$

The final prediction remains a one-way correction:

$$\hat{y}(z) = \begin{cases} y_{\min}, & h_A(z) \neq y_{\min} \ \wedge \ g_q(z) = 1, \\ h_A(z), & \text{otherwise}, \end{cases} \qquad q \in \{1, 2, 3, 4\}. \tag{10}$$

At $q = 1$, any expert can trigger correction; $q = 2$ or 3 requires at least two or three minority votes, respectively. With only three experts, $q = 4$ retains the anchor for every input. Pooled out-of-fold (OOF) training predictions select the threshold. Proposition 1 characterizes the balanced-accuracy change for any such gate.

3.5 Balanced-Risk Target of One-Way Rescue

We analyze how minority-directed corrections change balanced accuracy by separating recovered errors from overturned correct predictions. Let $g : \mathbb{R}^d \to \{0, 1\}$ be any training-fitted rescue gate and define its active region

$$\Gamma_g = \{z : g(z) = 1, \ h_A(z) \neq y_{\min}\}. \tag{11}$$

The corresponding classifier h_g predicts $y_{\min}$ on Γ_g and otherwise retains h_A. Let $\mathrm{BA}(h) = 1 - \mathcal{R}_{\mathrm{bal}}(h)$.

Proposition 1 (Balanced rescue identity)**.** *For any anchor h_A and measurable gate g,*

$$\mathrm{BA}(h_g) - \mathrm{BA}(h_A) = \frac{1}{C}\Bigg[\mathbb{P}(\Gamma_g \mid Y = y_{\min}) - \sum_{c \neq y_{\min}} \mathbb{P}(\Gamma_g, h_A(Z) = c \mid Y = c)\Bigg]. \tag{12}$$

The first term is recovered minority mass. For a rest class c, only overturning an anchor-correct prediction changes accuracy; replacing one wrong prediction with another has zero effect. Summing class-conditional changes proves the identity (Appendix A). Correction improves balanced accuracy exactly when recovered minority mass exceeds class-balanced harm on anchor-correct rest examples.

SCALE constructs a gate within this one-way correction family through expert voting and selects its threshold using training cross-validation. The identity separates the resulting gain into recovery and harm. Section 4.3 evaluates the resulting corrections by measuring minority recovery and harm relative to SCALE's fitted anchor.

4 EXPERIMENTS

We evaluate whether anchored support calibration improves frozen-feature classification, how its gains change as minority support shrinks, and how local coordinates and support estimates contribute. Univariate and multivariate benchmarks, correction diagnostics, and component comparisons address these questions. For statistical comparisons, we average paired performance differences across imbalance ratios and encoders within each source dataset, then use datasets as the units of analysis.

4.1 UNIVARIATE EVALUATION

Datasets and Representations. We consider 112 fixed-length UCR datasets (Dau et al., 2019) and 30 datasets from the Bake Off Redux collection (Middlehurst et al., 2024). We check each ratio (3:1, 5:1, 10:1, 20:1) separately, requiring at least 100 original training examples and 15 retained minority examples. The smallest training class is designated as $y_{\min}$ and downsampled against the largest class; all other classes and the original test split are retained. The evaluated release contains 46, 43, 29 and 20 datasets at the four ratios: 138 tasks from 58 source datasets (46 UCR and 12 Redux). Appendix H gives the construction rule, release coverage and per-dataset counts. We evaluate four shared pretrained encoders—MOMENT (Goswami et al., 2024), MantisV2 (Feofanov et al., 2026), Chronos-2 (Ansari et al., 2025), and UniTS (Gao et al., 2024)—together with a single task-trained H-Inception network (denoted Inception) (Ismail-Fawaz et al., 2022), giving 690 encoder–task pairs. Inception provides a supervised representation learned from scratch, complementing the pretrained encoders. This choice is motivated by the strong benchmark performance of the H-InceptionTime family (Middlehurst et al., 2024). All encoders are frozen before downstream adaptation, and all post-encoder methods share the same feature cache and data split. Appendix C details the architecture and training protocol.

Table 1: Univariate classification performance (%, higher is better). BA denotes balanced accuracy. Bold and underlined values indicate the best and second-best results, respectively, based on unrounded means.

	Balanced accuracy by encoder					Aggregate		
Method	Inception	MOMENT	MantisV2	Chronos-2	UniTS	BA	Macro-F1	Min-F1
Natural	75.87	70.92	77.13	74.30	67.74	73.19	72.73	58.43
cRT	78.48	72.29	78.66	75.88	71.27	75.32	75.03	63.94
Latent-SMOTE	78.50	72.50	78.77	75.82	71.44	75.41	75.21	63.92
T-SMOTE	77.55	72.13	77.62	75.13	63.64	73.22	72.51	58.42
Threshold	77.92	72.91	78.98	76.46	70.26	75.31	74.64	63.40
LogitAdj	78.58	<u>73.43</u>	79.83	<u>77.21</u>	71.14	<u>76.04</u>	<u>75.77</u>	<u>64.99</u>
P2P	78.58	73.43	<u>79.83</u>	77.21	71.13	76.04	75.77	64.99
Raw ensemble	<u>78.67</u>	73.14	79.58	76.78	<u>71.67</u>	75.97	75.62	64.99
SCALE	**79.43**	**74.93**	**80.69**	**78.49**	**73.36**	**77.38**	**76.76**	**67.37**

Baselines and Hyperparameter Tuning. We compare Natural logistic regression, class-weighted cRT (Kang et al., 2020), Latent-SMOTE on frozen features (Chawla et al., 2002), T-SMOTE (Zhao et al., 2022), minority threshold tuning, logit adjustment (Menon et al., 2021), P2P (Bhat et al., 2025), the raw ensemble, and SCALE. Only T-SMOTE synthesizes training series before frozen encoding. Every method uses 100 TPE trials via Optuna (Bergstra et al., 2011; Akiba et al., 2019) per encoder–task pair and seed, with stratified three-fold training CV. Selection maximizes pooled OOF BA, with Macro-F1 breaking ties; the selected configuration is refitted on the full training partition. Results average ten model-fitting seeds on fixed downsampled training sets (Appendix H.1). Appendices C–D specify the search spaces and fitting details.

Metrics and Statistical Unit. Balanced accuracy averages the C class recalls; Macro-F1 averages the C classwise F1 scores; Min-F1 is $\min_{c \in \mathcal{Y}} \mathrm{F1}_c$, the worst-class F1, which need not be the F1 of $y_{\min}$. Sensitivity and specificity use $y_{\min}$ versus rest. We report scores as percentages and absolute differences in percentage points (pp). Reported metrics are first averaged over the ten separately selected and refitted seed runs within each encoder–task pair. Encoder means then give

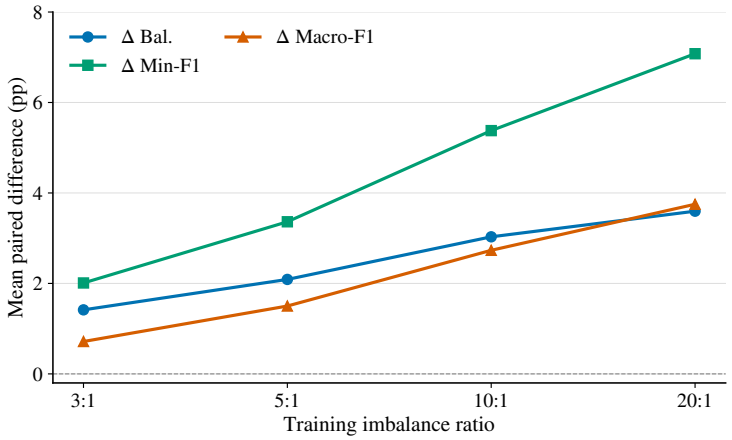


Figure 3: Gains increase with imbalance. SCALE minus cRT on the same 14 datasets and five encoders at each ratio (70 pairs), after averaging ten seeds. Points show mean gains in percentage points.

each task equal weight, and aggregate means weight the five encoders equally. Paired differences are averaged within source datasets and assessed using exact two-sided Wilcoxon tests (58 datasets for univariate evaluation, eight for multivariate evaluation). Reported p-values use Holm correction within three families: 24 univariate comparisons (eight baselines $\times$ three metrics), six component-ablation comparisons (two variants $\times$ three metrics), and three multivariate comparisons against cRT. Deltas are computed before rounding.

Performance Across Representations. SCALE has the highest balanced accuracy on each representation and the highest aggregate value for all three metrics (Table 1). Relative to cRT, it gains 2.06 balanced-accuracy, 1.73 Macro-F1 and 3.43 Min-F1 points ($p < 10^{-5}$ for each). Logit adjustment and P2P are the strongest standard remedies in aggregate BA, both rounding to 76.04%; SCALE exceeds logit adjustment by 1.34 BA, 1.00 Macro-F1 and 2.38 Min-F1 points ($p < 0.001$ for each).

The raw ensemble supplies a nonlinear correction reference. SCALE exceeds it by 1.41 BA points, with gains in all three metrics ($p < 0.001$ for each); Section 4.4 examines the contributions of local coordinates and support expansion. SCALE also has the best average within-task BA rank on every encoder (Appendix G). Its anchor-only fallback ($q = 4$) is selected in 68 of 6,900 seed-level runs (0.99%).

Adaptation cost. With two allocated CPU cores, SCALE averages 187.63 seconds per task–seed run, compared with 62.49 for cRT. Appendix G.1 reports complete search-and-refit costs and accounts for T-SMOTE's GPU-based synthetic-sample encoding.

Gains increase as minority support shrinks. We compare all four ratios on the same 14 datasets and five encoders. Figure 1 uses the 13 binary datasets within this cohort. Figure 3 shows that SCALE's gains over cRT rise from 1.42 to 3.60 BA points and from 2.01 to 7.08 Min-F1 points between 3:1 and 20:1; Macro-F1 gains rise from 0.72 to 3.75 points. Each point averages ten seeds over the same 70 encoder–dataset pairs. Gains are largest where minority supervision is most restricted within this cohort. Appendix H lists the datasets.

4.2 Multivariate Evaluation

Datasets and Evaluation Protocol. We evaluate eight datasets from the Multiverse archive (Middlehurst et al., 2026), using the same five frozen representations and four imbalance ratios as in Section 4.1, giving 32 tasks per encoder and 160 pairs. Appendix H.3 and Table 18 list the datasets, channels and sample counts. The downstream tuning and fitting protocol matches the univariate evaluation (Appendix D).

Table 2: Multivariate classification performance (%, higher is better). Highlighting follows Table 1.

	Balanced accuracy by encoder					Aggregate		
Method	Inception	MOMENT	MantisV2	Chronos-2	UniTS	BA	Macro-F1	Min-F1
Natural	76.18	62.13	71.08	65.06	57.11	66.31	62.34	46.80
cRT	77.42	64.37	72.11	67.96	62.78	68.93	67.86	58.95
Latent-SMOTE	77.30	64.29	71.91	67.87	62.81	68.84	67.74	58.60
T-SMOTE	76.78	63.39	71.48	66.28	54.75	66.54	63.58	49.91
Threshold	77.13	64.58	72.49	67.64	62.59	68.89	67.25	57.63
LogitAdj	77.22	64.41	72.44	67.87	62.84	68.96	67.26	57.30
P2P	77.25	64.43	72.46	67.89	62.65	68.93	67.40	57.71
Raw ensemble	77.44	64.41	72.53	68.01	62.78	69.03	67.92	58.93
SCALE	**77.65**	**65.80**	**73.36**	**68.26**	**63.01**	**69.62**	**68.56**	**60.39**

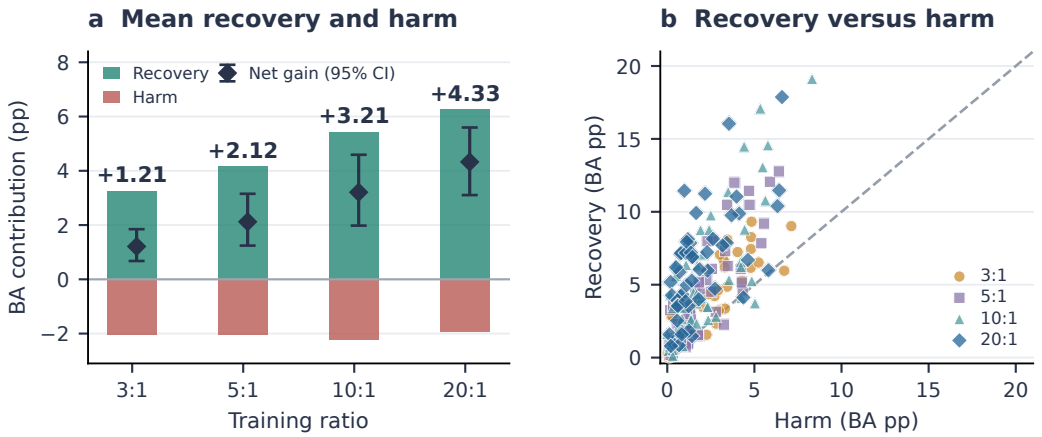


Figure 4: Recovery and harm relative to SCALE's own fitted anchor. (a) Class-balanced contributions and net BA gains with 95% dataset-bootstrap intervals; numbers show mean net gains. (b) All 208 combinations: 13 binary datasets × 4 encoders × 4 imbalance ratios, each averaged over ten seeds. Points above equality indicate positive net BA gains. Count-based definitions are in Appendix I.

Performance Across Representations. SCALE has the highest BA on all five representations and the highest aggregate scores in Table 2: 69.62% BA, 68.56% Macro-F1 and 60.39% Min-F1. Gains over cRT, computed before rounding, are 0.69, 0.70 and 1.44 points, respectively. Two-sided exact Wilcoxon tests on the eight source-dataset mean differences give Holm-adjusted $p = 0.0234$, 0.1094 and 0.1484, respectively.

4.3 Recovery and Harm of Anchor Correction

We analyze minority recovery and prediction harm on the same diagnostic cohort as Figure 1, averaging the ten main-benchmark seeds within each combination. Appendix I provides the experimental protocol and complete results.

Recovery and harm. Figure 4 measures both terms of Equation (12) against each fitted internal anchor. Mean recovery exceeds harm at all four ratios, yielding net BA gains of 1.21, 2.12, 3.21, and 4.33 points. Across 208 seed-averaged combinations, 188 improve and 20 decline.

Recovery under greater imbalance. From 3:1 to 20:1, recovery contributions rise from 3.26 to 6.27 points, while harm remains between 1.94 and 2.23 points. The growing net gain thus comes primarily from increased minority recovery, with a comparatively stable cost in majority errors. At 20:1, 51 of 52 combinations have positive mean gains. Together with the readout diagnostic in Figure 1, this pattern is consistent with useful class information remaining underused when minority supervision is scarce. Complementary support views allow SCALE to recover some of these errors while retaining the frozen encoder.

4.4 Ablation Study

We evaluate LRG and SSE through independently retuned variants. Removing LRG removes the residual coordinate and drops the now-redundant linear expert; removing SSE deletes the expanded-support expert E. Removing both gives the raw ensemble. Each variant is selected separately within training CV on the same 690 encoder–task pairs and ten seeds, with deltas relative to the same selected cRT baseline (Table 3). Both two-expert variants include an anchor-only fallback ($q = 3$).

Table 3: Component ablations. Gains over cRT are reported in percentage points; higher is better.

Variant	Δ BA	Δ Macro-F1	Δ Min-F1
w/o LRG and SSE (raw ensemble)	+0.650	+0.582	+1.044
w/o SSE	+1.597	+1.380	+2.651
w/o LRG	+1.657	+1.391	+2.772
SCALE	**+2.064**	**+1.730**	**+3.426**

Contributions of LRG and SSE. Removing LRG lowers BA by 0.407 and Min-F1 by 0.654 points; removing SSE lowers them by 0.467 and 0.775 points. Both removals also reduce Macro-F1. Full SCALE outperforms both ablations across all three metrics (Holm-adjusted $p < 0.001$ for each comparison). These independently retuned comparisons support using local residual coordinates together with support expansion for minority correction. Appendix B provides complementary evidence for LRG through coordinate comparisons and local-risk analysis.

Expert votes and parameter sensitivity. Appendix E.2 separately holds the fitted anchor and experts fixed and changes only the enabled votes. Adding E to L + O gains 0.643 BA points; adding O to L + E gains 0.444 points. Appendix F reports a post hoc sensitivity analysis of the construction parameters.

5 Discussion and Conclusion

Rich frozen time-series representations can retain useful class information that remains underused when minority labels are scarce. Our diagnostics show that fitting minority training examples well does not ensure reliable generalization, motivating adaptation of how the readout uses the available representation. SCALE combines shared local geometry with observed and modeled support to propose minority-directed corrections to a class-balanced anchor, without additional labels or encoder updates. Across the evaluated benchmarks, this approach improves aggregate classification performance. The correction analysis shows that larger gains under greater imbalance come primarily from increased minority recovery while harm remains relatively stable. Together, these findings support using shared feature structure and complementary support views to improve minority predictions under scarce supervision.

Future work will examine how these benefits depend on representation quality, feature dimensionality, and minority-support variability through controlled comparisons and repeated support sampling. We will also extend correction to multiple rare classes and select support experts adaptively to balance predictive gains and computational cost.

AI USE STATEMENT

Generative AI tools assisted language editing, consistency checks, and code refactoring. The authors have reviewed and verified the AI-assisted material and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the method and evaluation. Appendix A proves Proposition 1; Appendices C–D give the selection and fitting protocols; Appendix H lists dataset coverage and task construction. Appendix I documents the additional frozen-feature diagnostic, reports all 208 net changes, and states the illustrative-example selection rule. Code, experimental results, and datasets are available at `https://anonymous.4open.science/r/SCALE-DD85`.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019. doi: 10.1145/3292500.3330701.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2546–2554, 2011.

S. Divakar Bhat, Amit More, Mudit Soni, and Surbhi Agrawal. Prior2Posterior: Model prior correction for long-tailed learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1289–1298, 2025.

David S. Broomhead and David Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

Hong Cao, Xiao-Li Li, David Yew-Kwong Woon, and See-Kiong Ng. Integrated oversampling for imbalanced time series classification. *IEEE Transactions on Knowledge and Data Engineering*, 25 (12):2809–2822, 2013. doi: 10.1109/TKDE.2013.37.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002.

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. doi: 10.1109/TSP.2010.2053029.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

Damien Dablain, Bartosz Krawczyk, and Nitesh V Chawla. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (9):6390–6404, 2023a. doi: 10.1109/TNNLS.2021.3136503.

Damien A. Dablain, Colin Bellinger, Bartosz Krawczyk, and Nitesh V. Chawla. Efficient augmentation for imbalanced deep learning. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pp. 1433–1446, 2023b. doi: 10.1109/ICDE55515.2023.00114.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019. doi: 10.1109/JAS.2019.1911747.

Angus Dempster, Daniel F Schmidt, and Geoffrey I Webb. Hydra: competing convolutional kernels for fast and accurate time series classification. *Data Mining and Knowledge Discovery*, 37(5): 1779–1805, 2023. doi: 10.1007/s10618-023-00939-3.

Vasilii Feofanov, Songkang Wen, Jianfeng Zhang, Lujia Pan, and Ievgen Redko. MantisV2: Closing the zero-shot gap in time series classification with synthetic data and test-time strategies. *arXiv preprint arXiv:2602.17868*, 2026.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Ali Ismail-Fawaz, Maxime Devanne, Jonathan Weber, and Germain Forestier. Deep learning for time series classification using new hand-crafted convolution filters. In *2022 IEEE International conference on big data (Big Data)*, pp. 972–981. IEEE, 2022.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Jingyang Lyu, Kangjie Zhou, and Yiqiao Zhong. A statistical theory of overfitting for imbalanced classification. In *International Conference on Learning Representations*, 2026.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Matthew Middlehurst, James Large, Michael Flynn, Jason Lines, Aaron Bostrom, and Anthony Bagnall. HIVE-COTE 2.0: a new meta ensemble for time series classification. *Machine Learning*, 110:3211–3243, 2021.

Matthew Middlehurst, Patrick Schäfer, and Anthony Bagnall. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38(4):1958–2031, 2024. doi: 10.1007/s10618-024-01022-1.

Matthew Middlehurst, Aiden Rushbrooke, Ali Ismail-Fawaz, Maxime Devanne, Germain Forestier, Angus Dempster, Geoffrey I. Webb, Christopher Holder, and Anthony Bagnall. The multiverse of time series machine learning: an archive for multivariate time series classification. *arXiv preprint arXiv:2603.20352*, 2026.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chang Wei Tan, Angus Dempster, Christoph Bergmeir, and Geoffrey I. Webb. MultiRocket: Multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, 36(5):1623–1646, 2022.

Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367. IEEE, 2010.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Pu Zhao, Chuan Luo, Bo Qiao, Lu Wang, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. T-SMOTE: Temporal-oriented synthetic minority oversampling technique for imbalanced time series classification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2406–2412, 2022.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6908–6917, 2022.

A Proof of the Balanced Rescue Identity

Let $I_A(z, y) = \mathbb{1}\{h_A(z) = y\}$ and $I_g(z, y) = \mathbb{1}\{h_g(z) = y\}$. Since $h_g = h_A$ outside Γ_g, their class-conditional accuracies differ only on this region. If $Y = y_{\min}$, membership in Γ_g implies $h_A(Z) \neq y_{\min}$ and $h_g(Z) = y_{\min}$, hence

$$\mathbb{E}[I_g - I_A \mid Y = y_{\min}] = \mathbb{P}(\Gamma_g \mid Y = y_{\min}). \tag{13}$$

For a rest class $c \neq y_{\min}$, h_g predicts $y_{\min}$ on Γ_g. The change is -1 exactly when $h_A(Z) = c$ and zero when the anchor was already wrong. Therefore

$$\mathbb{E}[I_g - I_A \mid Y = c] = -\mathbb{P}(\Gamma_g, h_A(Z) = c \mid Y = c). \tag{14}$$

Averaging the C class-conditional differences proves Equation (12).

Pointwise characterization. The same result has a pointwise form. Define a class-balanced distribution Q by $Q(Y = c) = 1/C$ and $Q(Z \mid Y = c) = P(Z \mid Y = c)$, and let $\eta_c^{\mathrm{bal}}(z) = Q(Y = c \mid Z = z)$. If the anchor predicts $a = h_A(z) \neq y_{\min}$, switching from a to $y_{\min}$ changes conditional balanced utility by

$$\delta_A(z) = \eta_{y_{\min}}^{\mathrm{bal}}(z) - \eta_a^{\mathrm{bal}}(z). \tag{15}$$

Corollary 1 (Optimal one-way correction)**.** *Among all measurable gates restricted to rest-to-minority changes, one balanced-risk-optimal gate rescues when $h_A(z) \neq y_{\min}$ and $\delta_A(z) > 0$, and otherwise retains the anchor. Either action is optimal when $\delta_A(z) = 0$.*

For the corollary, evaluate the two admissible actions under the balanced distribution Q. At a point where $h_A(z) = a \neq y_{\min}$, retaining the anchor has conditional correctness $\eta_a^{\mathrm{bal}}(z)$ and rescuing has conditional correctness $\eta_{y_{\min}}^{\mathrm{bal}}(z)$. The latter is preferable exactly when Equation (15) is positive. Pointwise selection of the better admissible action minimizes balanced risk over the restricted gate class.

B Additional Evidence for Local Reference Geometry

Diagnostic protocol. We evaluate residual augmentation through paired comparisons of raw and augmented features, holding the representation, train/test split, and classifier protocol fixed within each pair. The default geometry is $K = 8$, $\tau = 1$, and one LDA direction; the full augmentation is $[z, r, u]$. These fixed-geometry diagnostics assess the coordinate construction independently of the ten-seed search for the complete SCALE pipeline. SCALE selects its local region count within training CV and uses the compact u block. We report balanced accuracy and Macro-F1 for the coordinate comparisons; Table 1 evaluates the complete procedure.

B.1 Geometric Effect of the Residual Coordinate

Let $a_c = W D_s^{-1} q_c$ and $\widetilde{b}_c = b_c - q_c^\top D_s^{-1} m_{\mathrm{tr}}$. In the hard-assignment limit, a query assigned to region k has $r(\widetilde{z}) \approx (\widetilde{z} - c_k)/\rho_k$, giving

$$s_c(z) \approx w_c^\top z + \frac{a_c^\top \widetilde{z}}{\rho_k} + \widetilde{b}_c - \frac{a_c^\top c_k}{\rho_k}. \tag{16}$$

Thus a shared linear head can express a correction whose scale and offset depend on the local reference region. Soft affinities smooth this dependence. The correction is locally affine in the normalized geometry coordinate; row normalization means it need not be affine in the original feature z. Setting $q_c = 0$ recovers a raw-feature linear score, while the residual term expands the available score family. Appendix B.4 compares these coordinates with simpler local descriptors.

B.2 Breadth Across Representations and Classifier Corrections

The paired improvements span learned, pretrained, and randomized features, including TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), and MultiRocket–Hydra (Tan et al., 2022; Dempster et al., 2023). They show that the coordinate can help without updating the feature extractor, including

Table 4: Paired raw-versus-LRG representation evaluation, with 138 tasks per row. Values are raw → augmented scores in percent. Learned encoders use balanced mini-batch training. Raw and augmented features use the same classifier within each row. MultiRocket-Hydra uses projected-residual augmentation and threshold tuning.

Representation	Type	Bal. acc. (%)	Macro-F1 (%)
Inception	learned	76.9 → 78.6	75.7 → 78.2
TimesNet	learned	69.4 → 73.7	67.8 → 73.0
ResNet1D	learned	74.8 → 77.4	73.3 → 76.7
TCN	learned	71.0 → 75.1	69.0 → 74.2
PatchTST	learned	74.2 → 76.0	73.1 → 75.4
MOMENT	frozen pretrained	65.1 → 70.1	61.9 → 69.6
MultiRocket-Hydra	randomized transform	76.5 → 77.8	74.9 → 77.4

Table 5: Compatibility of LRG with post-encoder correction on 138 tasks with fixed Inception CE features. Values are base → full-LRG scores in percent; deltas are percentage points computed before rounding. Calibration and prior estimation use training-only OOF predictions, with geometry refitted inside each OOF training fold.

Correction	Bal. acc. (%)	Δ Bal. (pp)	Macro-F1 (%)	Δ Macro-F1 (pp)
CW-logistic	76.2 → 79.0	+2.80	74.9 → 78.4	+3.50
Threshold	76.4 → 78.5	+2.00	75.7 → 78.1	+2.47
LogitAdj	76.6 → 78.8	+2.17	75.7 → 78.2	+2.50
P2P	76.5 → 78.8	+2.22	75.6 → 78.2	+2.62
VectorScale	76.8 → 79.0	+2.16	75.8 → 78.6	+2.77
TauNorm	68.8 → 75.2	+6.36	65.7 → 74.8	+9.04
cRT	77.0 → 78.9	+1.82	75.9 → 78.5	+2.59

on representations outside SCALE's main evaluation. The relevant comparison is within each row; different encoder-training and head protocols preclude a direct ranking against Table 1.

Positive paired gains persist after threshold, prior, calibration (vector scaling; Guo et al., 2017), and classifier corrections. In this study, geometry and decision-level correction are complementary. The large TauNorm gain partly reflects its weaker starting point; the cRT and P2P comparisons are more relevant to SCALE's balanced anchor. This compatibility diagnostic includes CW-logistic and classifier-retrained cRT as separate settings. The ten-seed benchmark uses the baseline implementation and selection protocol in Appendix C; absolute scores should be compared within their respective protocols.

B.3 Training-Local Diagnostics and Error Concentration

For each minority training observation i, let e_i be the fraction of its five nearest other training neighbors belonging to the rest classes. We define relative rest exposure as

$$\mathrm{RRE} = \frac{\frac{1}{n_{\min}} \sum_{i:y_i=y_{\min}} e_i}{n_{\mathrm{rest}}/(n_{\mathrm{tr}}-1)}. \tag{17}$$

Neighborhoods use training-standardized, row-normalized features. The denominator is the expected rest fraction for a random training neighbor; RRE is computed separately within each task before averaging.

RRE rises across the evaluated ratio-specific task sets, including for fixed and pretrained extractors. The values remain below one: minority neighbors retain more same-class structure than random neighbors, but the observed support becomes less separated from the rest-class prior. Dataset eligibility varies with the ratio, so these means summarize different task sets.

A complementary diagnostic assigns queries a training-derived hybrid risk using local minority/rest proportions, their ambiguity, and multiclass impurity, aggregated through reference affinities. We split queries into low- and high-risk bins within each task and evaluate errors afterward; test labels

Table 6: Relative rest exposure across feature families and imbalance ratios. Fixed features summarize Hydra and MultiRocket; learned features summarize weighted-CE Inception and PatchTST; pretrained features use frozen MOMENT. RRE is dimensionless.

Feature family	3:1	5:1	10:1	20:1
Fixed	0.399	0.414	0.484	0.590
Learned	0.192	0.258	0.350	0.421
Pretrained	0.578	0.656	0.694	0.741

do not construct the risk score. Table 7 reports this diagnostic on 92 dataset–ratio tasks. These are task-averaged bin error rates, rather than pooled sample errors or balanced accuracy.

Table 7: Risk-bin analysis of raw and LRG-augmented features under class-weighted logistic regression on 92 tasks. Error rates are percentages and reductions are percentage points, computed before rounding.

Encoder	Risk bin	Raw error (%)	LRG error (%)	Reduction (pp)
Inception	Low	19.30	17.32	1.97
Inception	High	28.63	24.09	4.55
ResNet1D	Low	21.92	19.37	2.55
ResNet1D	High	31.65	26.03	5.62
TCN	Low	24.71	21.12	3.58
TCN	High	34.67	27.20	7.47

Larger reductions in high-risk bins link the residual's benefit to regions where the original head is less reliable. In this diagnostic, LRG adapts the full classifier and can change both minority and rest decisions. SCALE uses the coordinate in its experts and restricts their corrections to the minority class. The bin analysis supports using local residuals to inform these correction proposals.

B.4 Residual Construction and Sensitivity

The controlled append-only comparisons use $K = 8$, $\tau = 1$, and one LDA direction on 138 Inception CE dataset–ratio tasks. Each delta compares augmented and raw features under the same head, averaging class-weighted logistic and threshold-tuned results within a task before averaging tasks. These fixed-geometry diagnostics are separate from the ten-seed SCALE benchmark. Signed residuals outperform the tested alternatives. The comparison assesses the complete coordinate constructions, including their dimension and supervised projection.

Table 8: Controlled local-coordinate comparisons on 138 tasks under fixed geometry. Values are percentage-point gains over the matched raw-feature head. All listed signals are appended to the raw feature. The full-LRG row uses both residual blocks; SCALE uses the compact projected block. These comparisons isolate feature augmentation from the full ensemble.

Added coordinate	Δ Bal.	Δ Macro-F1
Class-prototype distances	+0.14	+0.40
k-means-center distances	+0.35	+0.63
RBF affinities	+0.06	+0.31
Scalar local statistics	+0.10	+0.40
RBF affinities + statistics	+0.01	+0.30
Full residual r	+2.11	+2.87
Projected residual u	+1.88	+2.59
Full LRG: $r + u$	+2.27	+3.07

The compact projected residual used by SCALE gains 1.88 BA points over matched raw-feature heads, compared with 2.27 for the full augmentation. These comparisons support the compact residual coordinate. The local-feature controls in Table 8 retain the raw feature and append the indicated signal. A further comparison with a VLAD-style replacement code yields changes of -8.83

balanced-accuracy and -8.17 Macro-F1 points. This replacement changes both the encoding and access to raw features, whereas the controls above retain the raw-feature path.

Table 9: One-factor sensitivity of full LRG, using the same 138-task, two-head aggregation as Table 8. Percentage-point gains over matched raw features. The default is $K = 8$, $d_{\mathrm{LDA}} \leq 1$, $\tau = 1$.

Setting	Δ Bal.	Δ Macro-F1
$K = 4$	+2.25	+3.02
$K = 8$	+2.27	+3.07
$K = 16$	+2.33	+3.12
$d_{\mathrm{LDA}} \leq 1$	+2.27	+3.07
$d_{\mathrm{LDA}} \leq 2$	+2.35	+3.16
$d_{\mathrm{LDA}} \leq 3$	+2.41	+3.21
$\tau = 0.5$	+1.85	+2.64
$\tau = 1$	+2.27	+3.07
$\tau = 2$	+2.10	+2.88

All tested settings retain positive average gains in this fixed-geometry robustness check. For binary tasks the LDA caps coincide because there is at most one discriminant direction.

C Baseline and Selection Details

Task-trained H-Inception encoder. We follow the single-network H-Inception architecture (Ismail-Fawaz et al., 2022). The network contains six Inception modules arranged in two residual blocks, with standard multiscale convolutional branches. Fixed handcrafted filters for increasing trends, decreasing trends, and peaks operate in parallel with the first Inception module. We extract the global average pooling output after the final residual block as one latent vector per series. We use a single network for this representation-based evaluation; the standard H-InceptionTime classifier ensembles predictions from five independently trained H-Inception networks.

Encoder training. For each dataset–imbalance-ratio task, we train the network from scratch with seed 2025 using the official training partition after imbalance construction. We create an internal 80/20 training–validation split that preserves every class in both subsets. Training uses AdamW, balanced mini-batches of size 64, unweighted cross-entropy, and no data augmentation. The initial learning rate is 10^{-3}. Training runs for at most 1,500 epochs, with a minimum of 100 epochs and early-stopping patience of 50 epochs based on validation balanced accuracy. Learning-rate reduction uses a patience of 50 epochs, a reduction factor of 0.5, and a minimum learning rate of 10^{-4}. We retain the checkpoint with the highest validation balanced accuracy.

Frozen features and downstream selection. Channel normalization statistics and checkpoint selection use only task-training data. The selected encoder is frozen before downstream cross-validation, and its feature cache is shared across adaptation methods and downstream seeds. Downstream cross-validation therefore selects adapters conditional on this fixed, task-trained encoder. Subsequent feature preprocessing is fitted within each downstream training fold and refitted on the full training partition after selection. The official test partition is used only for evaluation.

Common selection protocol. All methods use the same original train/test partitions, designated minority class, and frozen representation for each encoder–task pair. Hyperparameters are selected independently for each of ten seeds (2025–2034) using shuffled stratified three-fold cross-validation on the training partition. Each seed defines its own downstream CV split; within that seed, methods and candidate configurations share the same folds. All scalers, prior estimators, geometry transforms, samplers, bandwidth references, and classifiers are fitted using only the training portion of each fold. Validation examples are not resampled or used for fitting. T-SMOTE synthesizes series only from the fold's training portion and encodes them with the corresponding frozen encoder.

Every method uses 100 Optuna TPE trials per seed and encoder–task pair, with every trial evaluated by the same training-only three-fold CV protocol. Table 10 gives the baseline search spaces; Appendix D

specifies the SCALE search and sampler. Within each seed, configurations are ranked by balanced accuracy computed after pooling the held-out predictions across folds, with pooled Macro-F1 breaking ties. For methods with post-processing or anchored correction, selection evaluates the final multiclass prediction rule. For each method, encoder–task pair, and seed, the selected configuration is refitted on the full training partition and evaluated on the original test partition. Test labels are excluded from fold fitting and per-task CV selection. Reported results average the test metrics over the ten seeds.

Search spaces. Natural, cRT, Latent-SMOTE, T-SMOTE, Threshold, logit adjustment, and P2P search logistic regularization $C \in [0.1, 10]$ using `suggest_float(0.1, 10, log=True)`. Their other continuous parameters use linear scales; the Latent-SMOTE neighbor count is an integer. The intervals in Table 10 specify parameter domains and scales, with TPE proposing values adaptively.

For the raw ensemble, let $\mathcal{C}_{\log} = \{0.1, 1, 10\}$, $\mathcal{B} = \{0.25, 1, 4\}$, and $\mathcal{W} = \{1, \frac{n_{\text{fit}}}{C_{\text{classes}}\, n_{y_i,\text{fit}}}\}$, where n_{fit} is the number of fitting examples, C_{classes} is the number of classes, and $n_{y_i,\text{fit}}$ is the fitting count of class y_i. The effective RBF regularization is searched directly over

$$\mathcal{C}_{\text{RBF}} = \{0.025, 0.1, 1, 10, 100, 400\}.$$

The raw ensemble retains this discrete search space of 108 nominal combinations and uses the same 100-trial TPE budget.

Table 10: Baseline hyperparameter search spaces. Each method uses 100 TPE trials with training-only three-fold CV.

Method	Search space
Natural	Logistic C
cRT	Logistic C
Latent-SMOTE	C; integer neighbors $k \in \{3, 4, 5, 6, 7\}$; deficit fraction $\rho \in [0.5, 1]$
T-SMOTE	C; spy fraction $\in [0.10, 0.20]$; window ratio $\in [0.10, 0.30]$
Threshold	C; rule $\in$ {argmax, threshold}; $t \in [0.05, 0.95]$ in the threshold branch
Logit adjustment	C; strength $\alpha \in [0, 1]$
P2P	C; strength $\alpha \in [0, 1]$
Raw ensemble	Anchor $C \in \mathcal{C}_{\log}$; RBF $C \in \mathcal{C}_{\text{RBF}}$; bandwidth multiplier $\mathcal{B}$; loss weights $\mathcal{W}$

Baseline definitions. Natural uses unweighted logistic regression and cRT uses inverse-class-frequency weights computed from the fitting labels. Latent-SMOTE applies SMOTE (Chawla et al., 2002) to frozen embeddings; it does not reproduce the encoder–decoder training or loss of DeepSMOTE (Dablain et al., 2023a). It interpolates only the designated minority in features standardized using the original fitting examples. Its target count is $n_{\min} + \lceil \rho(n_{\max} - n_{\min}) \rceil$, where ρ is the selected deficit fraction. Neighborhood sizes are capped by the available minority count; singleton minorities use repeat sampling. The benchmark T-SMOTE implementation retains temporal border scoring and full-length interpolation, augments only the designated minority, and balances it to the largest fitting class. Its minimum window size, maximum leading time, and internal scoring-classifier C remain fixed at 8, 64, and 1, respectively, with windows capped by sequence length. Both samplers fit an unweighted logistic head, with standardization learned from original fitting features and applied to synthetic features.

Threshold, logit adjustment, and P2P jointly select the Natural head regularization and their adjustment parameter. Logit adjustment uses fitting-label class frequencies. P2P estimates its effective prior from mean inner-cross-fitted probabilities within each fitting fold, using up to three inner folds. If fewer than two inner folds are feasible, it uses the fitting head's in-sample probability mean. Final fitting uses the selected head's full-training out-of-fold probability mean. No outer validation examples enter a fold's prior estimate. The raw ensemble combines a class-balanced logistic anchor with one raw-feature RBF expert under the same one-way minority-correction rule. Its bandwidth reference uses observed standardized fitting features. Selection maximizes the final corrected-prediction BA.

SCALE and component ablations. SCALE uses one 100-trial search per seed and encoder–task pair over the eight-parameter space in Appendix D, with the remaining geometry and SSE settings fixed across tasks. Every trial evaluates $q \in \{1, 2, 3, 4\}$ using the same fitted expert predictions. The

variants in Table 3 are independently retuned within training CV. Separately, Table 12 reuses fixed expert fits and changes only the enabled votes, without Optuna searches. Both studies use the selected cRT baseline as their reporting reference (Appendix E).

Evaluation scope. The comparison covers 690 univariate encoder–task pairs. CV on the task-trained Inception cache conditions on an encoder fitted before the downstream folds. Results are first averaged across seeds, and statistical comparisons are made at the source-dataset level, keeping its ratios and encoder views together. Balanced accuracy is the primary metric and Macro-F1 and Min-F1 are secondary metrics. Multivariate tasks use the same search spaces, CV selection and fitting protocol, and are aggregated separately over the five encoders.

D Hyperparameter Tuning and Fitting Details

Per-seed search and selection. For each encoder–task pair, SCALE runs an independent Optuna study within each of ten seeds (2025–2034). Each study evaluates 100 trials on the training partition using shuffled stratified three-fold CV, reducing the fold count only when the smallest class requires it. All trials within a seed use the same folds. Each fold refits all downstream standardizers, local references, LDA projections, support models, bandwidth references and classifiers on its fitting portion alone.

Each trial selects the eight search parameters in Table 11; the three fixed settings are shared across tasks and trials. After fitting the anchor and three experts in each fold, we evaluate $q \in \{1, 2, 3, 4\}$ on their pooled out-of-fold predictions. These thresholds reuse the fitted expert predictions. Selection ranks configuration–threshold pairs by pooled OOF balanced accuracy, then Macro-F1, with remaining ties favoring the earlier trial and then the smaller q. A trial's objective is its best threshold's balanced accuracy. All reported metrics use the same selected configuration and threshold. Test labels and test performance enter neither per-task CV selection nor tie breaking. The compact search space and shared fixed settings were specified before the sensitivity analysis in Appendix F.

Hyperparameter search. We use TPE (Bergstra et al., 2011) through Optuna (Akiba et al., 2019), with `TPESampler` seeded by the declared downstream seed and 30 startup trials. The prespecified canonical configuration is enqueued. Each study completes 100 trials with pruning disabled. Each trial evaluates four vote thresholds using the same fitted components, giving 400 trial–threshold evaluations without additional fitting. The budget counts completed trials, which may include repeated configurations.

Table 11: SCALE hyperparameter search space.

Component	Parameter	Candidate values
LRG	Requested local regions K	$\{4, 8\}$
SSE	Maximum spectral rank $r_{\max}$	$\{8, 16\}$
Classifier	Shared logistic $C_{\log}$	$\{0.1, 1, 10\}$
LRG	LDA dimension cap d_{LDA}	$\{1, 2\}$
Classifier	Shared effective RBF C_{RBF}	$\{1, 10, 100\}$
Classifier	Shared bandwidth multiplier β	$\{0.25, 1, 4\}$
Residual	RBF residual weight λ	$\{0, 0.5, 1, 2\}$
Classifier	RBF fitting mode	`original`, `balanced_reference`

The hyperparameter search space in Table 11 is shared across all tasks. The anchor and LRG linear expert share the selected logistic regularization $C_{\log}$; the observed and expanded RBF experts share the selected C_{RBF} and bandwidth multiplier β.

Across all tasks and trials, we fix the soft-reference temperature $\tau = 1$, Gaussian share $\eta_G = 1$, and deficit-filling fraction $\phi_{\text{fill}} = 1$. Thus, support expansion uses Gaussian sampling to fill the entire minority deficit. Both univariate and multivariate evaluations use this search space and these fixed settings, with $q \in \{1, 2, 3, 4\}$ selected within each seed. The separate fixed expert ablations use the settings in Table 13 and $q = 1$.

Geometry and support fitting. The reference frame uses ordinary, unweighted K-means with ten initializations. For a fitting partition containing n observations, the effective region count is $\max\{1, \min(K, n, \max(2, \text{round}\sqrt{n}))\}$. Empty regions, if any, use radius one. Discriminant directions are fitted to unstandardized residuals using shrinkage LDA. For multiple classes, an SVD of centered class coefficients supplies the requested directions, capped at $C-1$ and the feature dimension, where C is the number of classes. Degenerate LDA fits fall back to the SVD of class means, or to an empty projected block if no nonzero direction is available. Projection means and scales are estimated from the fitting partition. The anchor and local linear expert each fit a separate feature standardizer before class-balanced logistic regression.

Residual scaling precedes SSE: both RBF experts use $v_\lambda = [\widetilde{z}, \sqrt{\lambda}u]$. At $\lambda = 0$, their residual columns are removed, while the linear expert retains $[z, u]$. The SSE subspace and covariance are fitted in the selected coordinate, rather than generating support in an unweighted coordinate and rescaling it afterward. The retained rank is capped by $n_{\min} - 1$ and the coordinate dimension; the OAS covariance eigenvalues have a 10^{-8} floor.

For a deficit-filling fraction ϕ_{fill}, the number of generated minority coordinates is $G = \lceil \phi_{\text{fill}}(n_{\max} - n_{\min}) \rceil$. The main configuration uses $\phi_{\text{fill}} = 1$ and generates all G points from the Gaussian model. All counts, references and covariance estimates use only the current fold's fitting partition. Shared coordinates, supports and identical fitting operations are cached within a split.

Joint RBF fitting modes. In `original` mode, both RBF experts use unit loss weights, and each computes its own reference bandwidth from its fitting matrix, observed for O and observed plus synthetic for E. The bandwidth is $\gamma_s = \beta/[d_s \operatorname{Var}_{64}(X_s)]$ for a nonconstant dense fitting matrix X_s of expert s. Here d_s is its column count, and Var_b denotes the variance over all matrix entries evaluated at b-bit floating-point precision. Its constant-matrix fallback is $\gamma_s = \beta$. A multiplier of one retains scikit-learn's literal `gamma="scale"` call.

In `balanced_reference` mode, both RBF experts use class-balanced loss weights and share the reference bandwidth computed from observed support: $\gamma_O = \gamma_E = \beta/[D_\lambda \operatorname{Var}_{32}(X_O)]$ when this variance exceeds 10^{-12}, with fallback β/D_λ, where $D_\lambda = \dim(v_\lambda)$. This variance is computed from the float32 observed coordinate matrix and is not class-weighted. Original-sample-size class-mass weights are passed as SVM sample weights:

$$w_i^{(s)} = \frac{n}{C\, n_{y_i}^{(s)}}, \qquad \sum_{i:y_i=c} w_i^{(s)} = \frac{n}{C}, \quad s \in \{O, E\}. \tag{18}$$

Here n is the original fitting-partition size, and $n_c^{(s)}$ is the number of class-c samples in expert s's training support. Both experts therefore have the same class loss mass despite their different support sizes. Each mode jointly sets the loss weights and bandwidth reference.

Computational scope. For I iterations, constructing K local references from n features of dimension d costs $O(InKd)$. This excludes residual projection, support generation, classifier fitting, and selection. Let D_λ be the dimension of v_λ, and let n_s be the fitting-support size for RBF expert s. A fully materialized RBF kernel matrix requires $O(n_s^2 D_\lambda)$ work to form and $O(n_s^2)$ storage; actual SVM costs depend on the solver and its cache. Expansion increases n_E relative to n_O. The 100-trial budget counts candidate evaluations, not equal numbers of model fits or equal wall-clock budgets across methods.

Refitting and reporting. Within each seed, the selected configuration is refitted on the full training partition. Its selected q is then applied to the official test predictions without further selection. The same predictor supplies all test metrics for that seed. Reported results average these metrics over ten seeds. All transformations are fitted after the encoder has been frozen. For task-trained Inception, upstream encoder training uses task-training labels before downstream CV folds are created, so CV is conditional on that already fitted encoder. Univariate and multivariate results are aggregated separately.

E Component Ablation Details

E.1 Independently Retuned LRG and SSE Ablations

Table 3 uses the same 690 univariate encoder–task pairs, ten seeds (2025–2034), and selected cRT reference as Table 1. Each variant independently selects its hyperparameters within training CV and is refitted on the training set. Aggregation follows the main evaluation.

The w/o LRG variant removes residual coordinates from every expert branch. Without these coordinates, L and the anchor use the same raw features, class-balanced logistic regression, regularization, fitting data, and random seed, yielding identical predictions. The implementation therefore drops L and retains only O and E as voting experts; the anchor remains the default predictor and does not cast an additional vote. This variant selects $q \in \{1, 2, 3\}$. This is not the $\lambda = 0$ setting, which retains LRG in L. The w/o SSE variant deletes the entire expanded-support expert E and its generated support. It retains L and O as voting experts and also selects $q \in \{1, 2, 3\}$. For both variants, $q = 3$ retains the fitted anchor for every input. Joint removal is the raw ensemble with one raw-feature RBF correction expert defined in Appendix C.

These comparisons assess each component within a separately selected pipeline. Deleting E changes the available experts, and removing LRG changes their input coordinates; the variants therefore do not hold model capacity fixed. They complement the fixed-fit vote comparisons below.

E.2 Fixed-Configuration Expert Ablations

Table 12: Fixed-configuration component ablations on 690 univariate encoder–task pairs, averaged over ten seeds. L/E/O denote the linear, expanded-support and observed-support experts. Component variants reuse the same fixed expert fits with $q = 1$ and do not use Optuna. Each variant includes the fixed internal anchor. Deltas are percentage points relative to the separately CV-selected cRT baseline in Table 1.

Anchor correction	L	E	O	Δ BA	Δ Macro-F1	Δ Min-F1
cRT (reference)	–	–	–	0.000	0.000	0.000
Anchor + L	✓	–	–	+0.518	+0.464	+0.967
Anchor + E	–	✓	–	+1.113	+0.940	+1.947
Anchor + O	–	–	✓	+0.840	+0.787	+1.484
Anchor + L + O	✓	–	✓	+0.978	+0.868	+1.752
Anchor + L + E	✓	✓	–	+1.177	+0.953	+2.071
Anchor + L + E + O	✓	✓	✓	+1.621	+1.447	+2.554

Reference and protocol. The component results in Table 12 average ten seeds (2025–2034), with equal weight given to each encoder–task pair. Each variant uses the fixed configuration in Table 13 and $q = 1$; no Optuna search is performed for these variants. For each task and seed, the same anchor and expert fits are reused, and only the enabled expert votes change. The reported deltas are relative to the independently CV-selected cRT baseline (BA 75.32%, Macro-F1 75.03%, Min-F1 63.94%; rounded), rather than the fixed internal anchor. Consequently, pairwise component contrasts share both fixed fits and a common reporting reference.

Expert inputs and fitting. The internal anchor uses raw features; L uses raw features and the LRG residual coordinate; O and E use $[\widetilde{z}, u]$ at the fixed $\lambda = 1$. SSE supplies synthetic minority coordinates only to E. In these fixed experiments, $\eta_G = 0.8$: the Gaussian count is $\eta_G G$ rounded to the nearest integer, with half-integer ties rounded to the nearest even integer. The remaining points interpolate minority neighbors as $v^+ = v_i + t(v_j - v_i)$, $t \sim \mathcal{U}(0, 1)$, with the neighbor count capped by available minority observations. The anchor and L use class-balanced logistic regression; O and E use unit loss weights and separate `gamma="scale"` bandwidths computed from their own fitting samples. Preprocessing and dimension caps follow Appendix D.

Fixed-configuration component contrasts. L + O improves the selected cRT baseline by 0.978 BA points. Replacing its observed RBF vote with the expanded-support vote raises BA by 0.199

Table 13: Shared fixed settings for the separate component ablations with $q = 1$. Parameters for an absent component are inactive; requested dimensions and neighbor counts are capped by available data.

Parameter	Role	Fixed value
K	Requested local regions	4
τ	Soft-reference temperature	1
d_{LDA}	Requested residual directions	2
$C_{\log}$	Internal anchor and linear regularization	1
C_{RBF}	Observed and expanded RBF regularization	10
β	RBF bandwidth multiplier	1
$r_{\max}$	Maximum SSE spectral rank	32
η_G	Gaussian share of generated support	0.8
k_{SSE}	Interpolation neighbors	5
ϕ_{fill}	Fraction of minority count deficit filled	1
λ	Residual weight in nonlinear coordinates	1
Mode	Joint RBF fitting configuration	`original`
q	Required minority votes	1

points and Min-F1 by 0.319 points. Enabling both support views in L + E + O adds a further 0.444 BA and 0.483 Min-F1 points relative to L + E. These comparisons change the enabled votes while retaining the same fitted models; the full Optuna-selected SCALE procedure is evaluated separately in Table 1. In this fixed `original` mode, expansion changes both class counts and the fitted RBF bandwidth. These contrasts therefore measure the effect of including each support expert, not support geometry alone.

F Construction-Parameter Sensitivity

We examine nine construction parameters with $q = 1$: local region count K, temperature τ, residual dimension d_{LDA}, logistic and RBF regularization $(C_{\log}, C_{\mathrm{RBF}})$, spectral rank $r_{\max}$, Gaussian share η_G, interpolation-neighbor count k_{SSE}, and deficit-filling fraction ϕ_{fill}. A three-level L27 orthogonal design and the shared default are evaluated on all 690 encoder–task pairs using a single seed (2025). The balanced accuracies in this analysis are measured on the official test partitions of the same benchmark tasks. This post hoc analysis characterizes parameter sensitivity; the compact search space and shared fixed settings were specified before these scores were obtained. The ten-seed benchmark uses training-only CV within that space to select each task's configuration (Appendix D). Each factor level appears in nine design configurations, and each pair of levels for two factors appears three times.

Table 14: L27 sensitivity analysis (seed 2025, $q = 1$). Parentheses show mean BA (%); range is the largest difference between factor-level means (pp).

Factor	Level 1 (BA)	Level 2 (BA)	Level 3 (BA)	Range (pp)
K	2 (76.85)	4 (76.91)	8 (76.86)	0.057
τ	0.5 (76.82)	1 (76.91)	2 (76.89)	0.093
d_{LDA}	1 (76.85)	2 (76.89)	3 (76.89)	0.042
$C_{\log}$	0.1 (76.84)	1 (76.91)	10 (76.87)	0.065
C_{RBF}	1 (76.95)	10 (76.87)	100 (76.80)	0.149
$r_{\max}$	8 (76.88)	16 (76.88)	32 (76.86)	0.020
η_G	0.5 (76.87)	0.8 (76.87)	1 (76.88)	0.010
k_{SSE}	3 (76.88)	5 (76.87)	7 (76.87)	0.008
ϕ_{fill}	0.5 (76.84)	0.75 (76.88)	1 (76.90)	0.056

Table 14 shows modest average sensitivity over the tested levels. RBF regularization has the largest range (0.149 BA points), followed by temperature (0.093 points); every other range is at most 0.065 points. These factor-level means summarize the joint design; the performance of a complete configuration also depends on interactions among its settings.

G Ranks, Class Structure, and Adaptation Cost

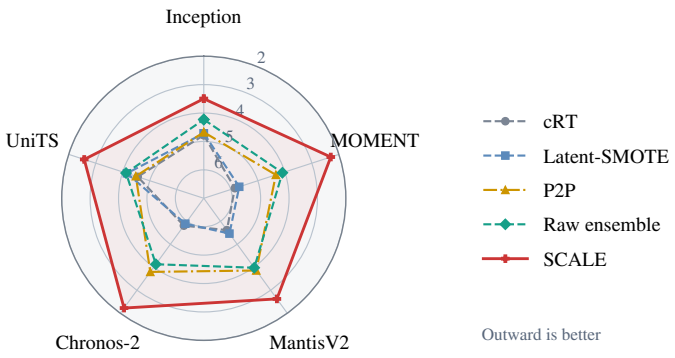


Figure 5: Mean task-wise balanced-accuracy ranks computed over all nine methods, with five selected methods shown for readability. Lower rank is better; the radial axis is reversed, so outward is better. Scores rounded to 12 decimal places receive midranks for ties.

Table 15: Optuna-selected SCALE by task structure. Performance deltas are percentage points relative to the independently selected cRT baseline in Table 1. Results average ten seeds within each encoder–task pair.

Scope	N	Δ Bal.	Δ Macro-F1	Δ Min-F1
Binary	405	+2.44	+2.07	+4.24
Multiclass	285	+1.53	+1.25	+2.27
All	690	+2.06	+1.73	+3.43

SCALE improves balanced accuracy over the selected cRT baseline by 2.44 points on 405 binary encoder–task pairs and 1.53 points on 285 multiclass pairs. The corresponding Min-F1 gains are 4.24 and 2.27 points.

G.1 Measured Adaptation Cost

Table 16 reports adaptation cost for eight methods on the same 690 univariate encoder–task pairs and ten seeds (2025–2034), giving 6,900 runs per method. Each run completes 100 trials with stratified three-fold training CV, followed by full-training refitting and test prediction. Selection uses pooled OOF balanced accuracy, then Macro-F1 and the smaller trial index. The initial encoding of the original train/test series is excluded from these adaptation costs.

Experiments use AMD EPYC 7452 processors (2.35 GHz base frequency). Each run uses two allocated CPU cores, except T-SMOTE, which uses four allocated CPU cores and one NVIDIA L4 GPU. Means and medians are over the 6,900 task–seed runs. Total task-wall time sums their individual wall times; CPU-core and GPU hours multiply wall time by the allocated core and device counts, respectively.

The feature-space methods reuse the original frozen-feature cache across folds and trials, so their adaptation runs use CPU resources only. T-SMOTE generates series in the input space. These samples depend on the fitting fold and the trial's spy-fraction and temporal-window settings, requiring repeated forward encoding before classifier fitting. The frozen encoder is loaded once per run; synthetic-sample encoding recurs across folds and trials. Its reported cost therefore includes both CPU computation and GPU encoding.

SCALE's mean run time is approximately 3.00× cRT's under the same two-core allocation. T-SMOTE averages 350.82 seconds under its four-core, one-GPU allocation. These measurements quantify the search-and-refit cost of the reported implementations with their respective resource allocations.

Table 16: Adaptation cost over 6,900 univariate task–seed runs per method. Times include the complete 100-trial search, refitting and test prediction. Resource hours are allocation-hours. Entries are rounded to the displayed precision.

Method	Mean/run (s)	Median/run (s)	Task-wall (h)	CPU-core (h)	GPU (h)
Natural	62.71	8.75	120.2	240.4	0
Threshold	60.48	8.38	115.9	231.8	0
Logit Adjustment	57.95	7.96	111.1	222.2	0
P2P	58.12	8.16	111.4	222.8	0
Latent-SMOTE	70.13	21.45	134.4	268.8	0
cRT	62.49	9.37	119.8	239.5	0
SCALE	187.63	88.08	359.6	719.3	0
T-SMOTE	350.82	109.43	672.4	2689.6	672.4

H Dataset Descriptions and Task Construction

H.1 Archive Splits and Imbalance Construction

Each task is a source dataset paired with one target imbalance ratio. The archive training and test partitions remain separate. Let N_0 and N_{te} denote their original sizes, C the number of classes, and n_c^0 the original training count of class c. Labels are encoded in lexicographic order of their archive strings. If all training class counts are equal, the last encoded class is designated as $y_{\min}$; otherwise, the first encoded class with the smallest count is used. Thus partial ties and fully balanced datasets have distinct, deterministic tie rules.

The configuration generator scans all supplied datasets and checks each target $r \in \{3, 5, 10, 20\}$ independently. Writing $m_0 = n^0_{y_{\min}}$ and $n^0_{\max} = \max_c n_c^0$, the count requirements are

$$N_0 \geq 100, \qquad 15 \leq \frac{n^0_{\max}}{r} \leq m_0. \tag{19}$$

For an eligible ratio, uniform sampling without replacement retains

$$m_r = \left\lfloor \frac{n^0_{\max}}{r} \right\rfloor, \qquad N_r = N_0 - m_0 + m_r. \tag{20}$$

Only the designated minority class is reduced; every observation of every other class is retained. Consequently, the realized largest-to-minority ratio is $n^0_{\max}/m_r$, which can slightly exceed the nominal r because of integer rounding. The class vocabulary and all test observations remain unchanged. For multiclass tasks, this constructs one rare class rather than a geometric progression of class sizes. A dataset need only support one ratio to enter the task pool; eligibility does not require all four ratios. The construction code also supports 50:1, which is outside this study.

The univariate materialization uses random downsampling with generator seed 2025. The multivariate materialization uses a separate seed for each dataset–ratio pair: the first four bytes of the SHA-256 digest of `2025:dataset:r:random`, interpreted as an unsigned big-endian integer, where `r` is the integer ratio. The saved train/test files fix the sample identities and are shared across methods and encoders. These data materialization seeds are separate from the model-fitting and support-expansion and CV seeds in Appendix C; the Optuna sampler seed is specified separately in Appendix D. The counts below refer to observed series before feature extraction or SSE, so synthetic support points do not increase N_r. No balancing or resampling is applied to the test set.

H.2 Univariate Datasets and Evaluated Coverage

The candidate inventory contains the 112 fixed-length UCR datasets and all 30 datasets in the Bake Off Redux collection (Dau et al., 2019; Middlehurst et al., 2024). The evaluated release contains 58 source datasets: 46 from the former and 12 from the latter. These include sensor and device signals (for example, FordA/B, ElectricDevices and the Aconity printer datasets), physiological recordings (ECG200 and NonInvasiveFetalECGThorax1/2), motion and gesture signals (GunPoint and UWaveGestureLibrary variants), image-outline sequences (the phalanx and HandOutlines datasets), spectra

(EthanolLevel and Ham), environmental and economic series, and synthetic classification problems. Their lengths range from 15 to 4201, with 2–42 classes, 100–8926 original training examples and 61–8236 test examples. The `_disc` names identify the supplied discretized classification versions. We retain their provided labels and fixed splits; the counts in Table 17 refer to the versions actually used, including the corrected Redux test files.

The ratio-wise task counts are 46, 43, 29 and 20, for 138 tasks in total. Of these, 81 tasks are binary and 57 are multiclass. Five encoders give 690 encoder–task pairs. Fourteen source datasets support all four ratios in the evaluated release; the remaining datasets contribute only their available ratios. The tables enumerate the release used by the reported experiments, rather than treating the entire candidate inventory as evaluated. In particular, the supplied materialization configuration routes MelbournePedestrian to the variable-length group. Thus count eligibility alone does not fully specify the released task set; Table 17 defines the exact scope of our comparisons.

The matched benchmark cohort in Figure 3 consists of AconityMINIPrinterLarge, AconityMINIPrinterSmall, AsphaltPavementTypeUni, AsphaltRegularityUni, DistalPhalanxOutlineCorrect, FloodModeling1_disc, FordA, FordB, HandOutlines, MiddlePhalanxOutlineCorrect, PhalangesOutlinesCorrect, ProximalPhalanxOutlineCorrect, SharePriceIncrease, and Strawberry.

Reading the dataset tables. Each dataset occupies two rows. The first reports C, temporal length T (or channels D and length T), the original training and test sizes N_0, N_{te}, the zero-based encoded minority label $y_{\min}$, and the counts m_0, m_{te} of that same class in the original training and test partitions. The second reports N_r/m_r, the total observed training size and retained minority count, separately for each ratio. A dash denotes an absent task; it is not a zero count. Test counts are identical across ratios. Superscripts U and R denote the fixed-length UCR subset and the Redux collection, respectively.

Table 17: Univariate dataset inventory and ratio-specific training counts. Each second row lists total/minority training counts (N_r/m_r). All datasets have one channel; sample counts exclude synthetic support.

Dataset	C	T	N_0	N_{te}	$y_{\min}$	m_0	m_{te}
AconityMINIPrinterLarge$^{\mathrm{R}}$	2	300	2403	1184	1	818	403
N_r/m_r 3:1: 2113/528	5:1: 1902/317		10:1: 1743/158		20:1: 1664/79		
AconityMINIPrinterSmall$^{\mathrm{R}}$	2	300	589	292	1	166	83
N_r/m_r 3:1: 564/141	5:1: 507/84		10:1: 465/42		20:1: 444/21		
AsphaltObstaclesUni$^{\mathrm{R}}$	4	736	390	391	0	80	80
N_r/m_r 3:1: 347/37	5:1: 332/22		10:1: –		20:1: –		
AsphaltPavementTypeUni$^{\mathrm{R}}$	3	2371	1055	1056	0	263	264
N_r/m_r 3:1: 928/136	5:1: 873/81		10:1: 832/40		20:1: 812/20		
AsphaltRegularityUni$^{\mathrm{R}}$	2	4201	751	751	0	370	370
N_r/m_r 3:1: 508/127	5:1: 457/76		10:1: 419/38		20:1: 400/19		
ChlorineConcentration$^{\mathrm{U}}$	3	166	467	3840	1	91	909
N_r/m_r 3:1: 463/87	5:1: 428/52		10:1: 402/26		20:1: –		
Computers$^{\mathrm{U}}$	2	720	250	250	1	125	125
N_r/m_r 3:1: 166/41	5:1: 150/25		10:1: –		20:1: –		
Covid3Month_disc$^{\mathrm{R}}$	3	84	140	61	0	29	12
N_r/m_r 3:1: 136/25	5:1: 126/15		10:1: –		20:1: –		
DistalPhalanxOutlineAgeGroup$^{\mathrm{U}}$	3	80	400	139	0	30	15
N_r/m_r 3:1: –	5:1: –		10:1: 395/25		20:1: –		
DistalPhalanxOutlineCorrect$^{\mathrm{U}}$	2	80	600	276	0	222	115
N_r/m_r 3:1: 504/126	5:1: 453/75		10:1: 415/37		20:1: 396/18		
Earthquakes$^{\mathrm{U}}$	2	512	322	139	1	58	35
N_r/m_r 3:1: –	5:1: 316/52		10:1: 290/26		20:1: –		
ECG200$^{\mathrm{U}}$	2	96	100	100	0	31	36
N_r/m_r 3:1: 92/23	5:1: –		10:1: –		20:1: –		

Continued on the next page.

Table 17 continued. Second rows: total/minority training counts (N_r/m_r).

Dataset		C	T	N_0	N_{te}	$y_{\min}$	m_0	m_{te}
ElectricDeviceDetection[R]		2	256	624	3768	1	81	530
N_r/m_r	3:1: –	5:1: –		10:1: 597/54		20:1: 570/27		
ElectricDevices[U]		7	96	8926	7711	5	509	743
N_r/m_r	3:1: –	5:1: 8898/481		10:1: 8657/240		20:1: 8537/120		
EthanolLevel[U]		4	1751	504	500	3	126	124
N_r/m_r	3:1: 420/42	5:1: 403/25		10:1: –		20:1: –		
FloodModeling1_disc[R]		2	266	471	202	1	127	51
N_r/m_r	3:1: 458/114	5:1: 412/68		10:1: 378/34		20:1: 361/17		
FloodModeling2_disc[R]		2	266	466	201	1	41	15
N_r/m_r	3:1: –	5:1: –		10:1: –		20:1: 446/21		
FloodModeling3_disc[R]		2	266	429	184	1	100	58
N_r/m_r	3:1: –	5:1: 394/65		10:1: 361/32		20:1: 345/16		
FordA[U]		2	500	3601	1320	1	1755	639
N_r/m_r	3:1: 2461/615	5:1: 2215/369		10:1: 2030/184		20:1: 1938/92		
FordB[U]		2	500	3636	810	1	1776	409
N_r/m_r	3:1: 2480/620	5:1: 2232/372		10:1: 2046/186		20:1: 1953/93		
FreezerRegularTrain[U]		2	301	150	2850	1	75	1425
N_r/m_r	3:1: 100/25	5:1: 90/15		10:1: –		20:1: –		
GunPointAgeSpan[U]		2	150	135	316	1	67	156
N_r/m_r	3:1: 90/22	5:1: –		10:1: –		20:1: –		
GunPointMaleVersusFemale[U]		2	150	135	316	1	64	150
N_r/m_r	3:1: 94/23	5:1: –		10:1: –		20:1: –		
GunPointOldVersusYoung[U]		2	150	136	315	0	65	150
N_r/m_r	3:1: 94/23	5:1: –		10:1: –		20:1: –		
Ham[U]		2	431	109	105	0	52	51
N_r/m_r	3:1: 76/19	5:1: –		10:1: –		20:1: –		
HandOutlines[U]		2	2709	1000	370	0	362	133
N_r/m_r	3:1: 850/212	5:1: 765/127		10:1: 701/63		20:1: 669/31		
LargeKitchenAppliances[U]		3	720	375	375	2	125	125
N_r/m_r	3:1: 291/41	5:1: 275/25		10:1: –		20:1: –		
MiddlePhalanxOutlineAgeGroup[U]		3	80	400	154	0	55	37
N_r/m_r	3:1: –	5:1: 392/47		10:1: 368/23		20:1: –		
MiddlePhalanxOutlineCorrect[U]		2	80	600	291	0	212	125
N_r/m_r	3:1: 517/129	5:1: 465/77		10:1: 426/38		20:1: 407/19		
MiddlePhalanxTW[U]		6	80	399	154	4	25	9
N_r/m_r	3:1: –	5:1: –		10:1: 390/16		20:1: –		
MixedShapesRegularTrain[U]		5	1024	500	2425	4	100	314
N_r/m_r	3:1: 433/33	5:1: 420/20		10:1: –		20:1: –		
NonInvasiveFetalECGThorax1[U]		42	750	1800	1965	15	35	52
N_r/m_r	3:1: 1783/18	5:1: –		10:1: –		20:1: –		
NonInvasiveFetalECGThorax2[U]		42	750	1800	1965	15	35	52
N_r/m_r	3:1: 1783/18	5:1: –		10:1: –		20:1: –		
PhalangesOutlinesCorrect[U]		2	80	1800	858	0	628	332
N_r/m_r	3:1: 1562/390	5:1: 1406/234		10:1: 1289/117		20:1: 1230/58		
PhoneHeartbeatSound[R]		5	3053	424	182	1	27	13
N_r/m_r	3:1: –	5:1: –		10:1: 422/25		20:1: –		
PowerCons[U]		2	144	180	180	1	90	90
N_r/m_r	3:1: 120/30	5:1: 108/18		10:1: –		20:1: –		
ProximalPhalanxOutlineAgeGroup[U]		3	80	400	205	0	72	17
N_r/m_r	3:1: 391/63	5:1: 365/37		10:1: 346/18		20:1: –		

Continued on the next page.

Table 17 continued. Second rows: total/minority training counts (N_r/m_r).

Dataset	C	T	N_0	N_{te}	$y_{\min}$	m_0	m_{te}
ProximalPhalanxOutlineCorrect[U]	2	80	600	291	0	194	92
N_r/m_r 3:1: 541/135	5:1: 487/81		10:1: 446/40		20:1: 426/20		
RefrigerationDevices[U]	3	720	375	375	2	125	125
N_r/m_r 3:1: 291/41	5:1: 275/25		10:1: –		20:1: –		
ScreenType[U]	3	720	375	375	2	125	125
N_r/m_r 3:1: 291/41	5:1: 275/25		10:1: –		20:1: –		
SemgHandGenderCh2[U]	2	1500	300	600	1	150	210
N_r/m_r 3:1: 200/50	5:1: 180/30		10:1: 165/15		20:1: –		
SemgHandMovementCh2[U]	6	1500	450	450	5	75	75
N_r/m_r 3:1: 400/25	5:1: 390/15		10:1: –		20:1: –		
SemgHandSubjectCh2[U]	5	1500	450	450	4	90	90
N_r/m_r 3:1: 390/30	5:1: 378/18		10:1: –		20:1: –		
SharePriceIncrease[R]	2	60	965	966	1	302	303
N_r/m_r 3:1: 884/221	5:1: 795/132		10:1: 729/66		20:1: 696/33		
SmallKitchenAppliances[U]	3	720	375	375	2	125	125
N_r/m_r 3:1: 291/41	5:1: 275/25		10:1: –		20:1: –		
SmoothSubspace[U]	3	15	150	150	2	50	50
N_r/m_r 3:1: 116/16	5:1: –		10:1: –		20:1: –		
StarLightCurves[U]	3	1024	1000	8236	0	152	1177
N_r/m_r 3:1: –	5:1: 962/114		10:1: 905/57		20:1: 876/28		
Strawberry[U]	2	235	613	370	0	219	132
N_r/m_r 3:1: 525/131	5:1: 472/78		10:1: 433/39		20:1: 413/19		
SyntheticControl[U]	6	60	300	300	5	50	50
N_r/m_r 3:1: 266/16	5:1: –		10:1: –		20:1: –		
TwoPatterns[U]	4	128	1000	4000	1	237	1011
N_r/m_r 3:1: 853/90	5:1: 817/54		10:1: 790/27		20:1: –		
UWaveGestureLibraryAll[U]	8	945	896	3582	7	100	460
N_r/m_r 3:1: 838/42	5:1: 821/25		10:1: –		20:1: –		
UWaveGestureLibraryX[U]	8	315	896	3582	7	100	460
N_r/m_r 3:1: 838/42	5:1: 821/25		10:1: –		20:1: –		
UWaveGestureLibraryY[U]	8	315	896	3582	7	100	460
N_r/m_r 3:1: 838/42	5:1: 821/25		10:1: –		20:1: –		
UWaveGestureLibraryZ[U]	8	315	896	3582	7	100	460
N_r/m_r 3:1: 838/42	5:1: 821/25		10:1: –		20:1: –		
Wafer[U]	2	152	1000	6164	0	97	665
N_r/m_r 3:1: –	5:1: –		10:1: 993/90		20:1: 948/45		
Worms[U]	5	900	181	77	4	17	8
N_r/m_r 3:1: –	5:1: 179/15		10:1: –		20:1: –		
WormsTwoClass[U]	2	900	181	77	0	76	33
N_r/m_r 3:1: 140/35	5:1: 126/21		10:1: –		20:1: –		
Yoga[U]	2	426	300	3000	0	137	1393
N_r/m_r 3:1: 217/54	5:1: 195/32		10:1: 179/16		20:1: –		

H.3 Multivariate Datasets

The multivariate study uses the eight datasets in the common evaluation manifest, with fixed, equal-length, nonmissing archive versions. These datasets are included in the Multiverse collection (Middlehurst et al., 2026). CrowdSourced classifies eyes-open versus eyes-closed EEG; STEW classifies low versus high cognitive workload from EEG; LowCost distinguishes idle states from imagined hand movement. FaceDetection uses MEG recordings to distinguish faces from scrambled images. FordChallenge predicts driver alertness from physiological, environmental and vehicle measurements. PenDigits classifies ten handwritten digits from two-coordinate pen trajectories. UCDHE-Rowing-MC

uses multichannel body-motion recordings for five-class rowing-execution classification. UCIActivity recognizes six activities from smartphone inertial measurements.

All eight support the four ratios under Eq. (19), giving 32 tasks and 160 encoder–task pairs across Inception, MOMENT, MantisV2, Chronos-2 and UniTS. Five datasets are binary and three are multiclass. Channels range from 2 to 144 and temporal lengths from 8 to 375. Table 18 lists the original split sizes and the observed training counts after the same single-class downsampling rule. These eight datasets define the reported multivariate scope; they are not the entire Multiverse archive.

Table 18: Multivariate dataset inventory and ratio-specific training counts. $D \times T$ gives channels and length; each second row lists total/minority training counts (N_r/m_r). The test split is unchanged.

Dataset	C	$D \times T$	N_0	N_{te}	$y_{\min}$	m_0	m_{te}
CrowdSourced	2	14×256	9456	2833	1	4727	1416
N_r/m_r 3:1: 6305/1576	5:1: 5674/945		10:1: 5201/472		20:1: 4965/236		
FaceDetection	2	144×62	5890	3524	1	2945	1762
N_r/m_r 3:1: 3926/981	5:1: 3534/589		10:1: 3239/294		20:1: 3092/147		
FordChallenge	2	30×40	29003	7254	0	10500	2733
N_r/m_r 3:1: 24670/6167	5:1: 22203/3700		10:1: 20353/1850		20:1: 19428/925		
LowCost	2	15×375	600	600	1	300	300
N_r/m_r 3:1: 400/100	5:1: 360/60		10:1: 330/30		20:1: 315/15		
PenDigits	10	2×8	7494	3498	3	719	336
N_r/m_r 3:1: 7035/260	5:1: 6931/156		10:1: 6853/78		20:1: 6814/39		
STEW	2	14×256	21384	7128	1	10692	3564
N_r/m_r 3:1: 14256/3564	5:1: 12830/2138		10:1: 11761/1069		20:1: 11226/534		
UCDHE-Rowing-MC	5	16×161	1992	440	1	380	70
N_r/m_r 3:1: 1748/136	5:1: 1694/82		10:1: 1653/41		20:1: 1632/20		
UCIActivity	6	9×128	7898	2401	2	1083	323
N_r/m_r 3:1: 7309/494	5:1: 7111/296		10:1: 6963/148		20:1: 6889/74		

I FROZEN-FEATURE DIAGNOSTICS AND ANCHOR CORRECTION

I.1 SCOPE AND FITTING PROTOCOL

Cohort and relation to the main benchmark. The aggregate panels of Figure 1 and Figure 4 use all 13 binary datasets that support all four ratios in the evaluated univariate release: AconityMINIPrinterLarge, AconityMINIPrinterSmall, AsphaltRegularityUni, DistalPhalanxOutlineCorrect, FloodModeling1_disc, FordA, FordB, HandOutlines, MiddlePhalanxOutlineCorrect, PhalangesOutlinesCorrect, ProximalPhalanxOutlineCorrect, SharePriceIncrease, and Strawberry. Four pretrained encoders (Chronos-2, MantisV2, MOMENT, UniTS) and four ratios (3:1, 5:1, 10:1, 20:1) give 208 conditions, with 52 at each ratio. No condition is excluded according to its outcome. The cohort differs from the ten-seed ratio analysis in Figure 3: it excludes the three-class AsphaltPavementTypeUni dataset and task-trained Inception. Minority support is fixed by the seed-2025 materialization. The fixed-$C = 1$ readout diagnostic uses this support once; the SCALE correction analysis uses the corresponding ten main-benchmark fits (seeds 2025–2034), which vary model and search randomness on the same support.

Sparse-support readouts. For each encoder–dataset pair, frozen features, all majority training examples, and the official test split are shared across ratios. Only the retained minority training support changes. Feature standardization is fitted on the current training support, followed by L2-regularized, class-balanced logistic regression with $C = 1$, the LBFGS solver, and a maximum of 5,000 iterations. Class balancing gives the two training classes equal total loss weight. Logits are oriented toward the designated minority class. Class-conditional recall is evaluated separately on the fitting support and unchanged test set. The minority gap is training minus test recall; the majority gap is defined analogously. No test labels enter standardization or fitting. Holding C fixed does not hold the regularizer relative to the average loss fixed as support size changes; preprocessing also depends on the current support. Thus the diagnostic measures the behavior of this fitting protocol, not an intervention on direction estimation alone.

Reference readout with additional training labels. We fit the same class-balanced $C = 1$ readout, including its standardizer, on the original TRAIN partition before minority downsampling. This partition is not necessarily class balanced. We retain its score direction and fit only an additive intercept on the scarce support using a class-balanced logistic loss. No TEST samples or labels enter the fit. This reference assesses discrimination in the frozen features with an expanded minority-label budget. Its score differs from the original TRAIN readout by a constant, so their ROC-AUCs are identical. The ROC-AUC comparison therefore distinguishes a ranking difference from an intercept-only effect.

SCALE and its internal anchor. The correction analysis uses the corresponding main-benchmark SCALE fits, selected by training-only three-fold CV with 100 TPE trials and $q \in \{1, 2, 3, 4\}$, including the anchor-only fallback (Appendix D). Each predictor is paired with its own fitted internal anchor, including the selected logistic regularization. This anchor can differ from both the fixed-$C = 1$ diagnostic readout and independently selected cRT. Recovery and harm are calculated from the paired test predictions. Across the 2,080 fits, $q = 1$, $q = 2$, $q = 3$, and $q = 4$ occur 1,585, 438, 44, and 13 times, respectively; the 13 fallback fits have zero recovery and harm.

Aggregation and uncertainty. For SCALE, we first average the ten seeds within each encoder–dataset–ratio combination. For each ratio, we then average the four encoders within each source dataset and weight the 13 datasets equally. The fixed-$C = 1$ readouts are deterministic on the fixed support and use the same encoder and dataset weighting. The pointwise 95% percentile intervals resample source datasets with replacement, carrying all encoders and all four ratios together in each draw (10,000 draws; random seed 20260920 for the readout diagnostic and 20260924 for the correction analysis). Paired changes use the same draws. Each correction scatter point represents one seed-averaged combination. The intervals summarize variation across datasets conditional on the fixed support; they do not measure uncertainty from resampling minority support.

I.2 Generalization and Recovery Results

Table 19 gives the binary diagnostic cohort trends. The minority train–test gap increases from 18.38 to 43.72 pp, while the majority gap falls from 6.81 to 3.10 pp. The increase in the difference between these two gaps is 29.06 pp (95% interval $[21.64, 36.38]$). At 20:1, the reference readout improves ROC-AUC in 51/52 pairs, with a dataset-equal mean gain of 4.18 pp ($[2.94, 5.54]$). Its mean BA advantage is 7.59 pp ($[5.88, 9.42]$), also positive in 51/52 pairs. These results show that stronger discrimination is accessible with more minority labels at fixed feature dimension.

Table 19: Binary diagnostic cohort: 13 datasets and four pretrained encoders. SCALE contributions average ten seeds per combination. Recall values are percentages; gaps and changes are in pp. The last row counts positive/zero/negative seed-averaged changes among 52 pairs per ratio, using unrounded means.

Quantity	3:1	5:1	10:1	20:1
Minority train recall	94.79	95.68	97.82	99.15
Minority test recall	76.41	71.04	63.85	55.42
Minority train–test gap	18.38	24.64	33.97	43.72
Majority train–test gap	6.81	5.49	4.24	3.10
Recovery contribution	3.26	4.17	5.44	6.27
Harm contribution (magnitude)	2.05	2.05	2.23	1.94
Net BA change	+1.21	+2.12	+3.21	+4.33
Net change: 95% interval	[0.68,1.85]	[1.24,3.15]	[1.98,4.59]	[3.10,5.60]
Positive / zero / negative	45/0/7	44/0/8	48/0/4	51/0/1

Count-based recovery and harm. For a binary test set with $n_{\min}$ minority and n_{maj} majority examples, let n_{rec} count incorrect anchor predictions corrected to the minority class and let n_{harm} count correct majority predictions overturned by SCALE. The BA contributions in pp are

$$R = 50\frac{n_{\mathrm{rec}}}{n_{\min}}, \qquad H = 50\frac{n_{\mathrm{harm}}}{n_{\mathrm{maj}}}, \qquad \Delta\,\mathrm{BA}_{\mathrm{pp}} = R - H. \tag{21}$$

Figure 4a plots R, $-H$, and their sum. The identity holds for each of the 2,080 fits and is preserved by seed averaging. The mean net change increases by 3.12 pp from 3:1 to 20:1 ($[2.29, 3.93]$); all 13 dataset means have positive endpoint changes, although not every dataset increases monotonically at intermediate ratios. Absolute accuracy still falls as support becomes scarcer: mean anchor/SCALE BA is $82.50/83.71\%$ at 3:1 and $74.90/79.23\%$ at 20:1.

Complete condition-level reporting. Table 20 reports all 208 net BA changes and all 52 reference ROC-AUC differences at 20:1. The 20 negative seed-averaged changes occur at 3:1, 5:1, 10:1, and 20:1 in counts 7, 8, 4, and 1. At 20:1, the only negative change is MOMENT–SharePriceIncrease (-0.24 pp).

Table 20: Complete matched diagnostic. Four columns report SCALE minus its own internal anchor in BA pp, averaged over ten seeds; the last reports reference minus scarce-support ROC-AUC at 20:1 in pp for the fixed-$C = 1$ readouts. All 208 combinations are retained, including negative changes. Source datasets are the bootstrap units.

Dataset	Encoder	3:1	5:1	10:1	20:1	Δ AUC
AcunityMINIPrinterLarge	Chronos-2	+0.67	+0.95	+0.99	+3.44	+1.51
	MantisV2	+0.71	+0.82	+3.09	+6.35	+1.61
	MOMENT	-0.04	+0.73	+2.89	+6.55	+4.23
	UniTS	+0.10	-0.16	+0.23	+2.18	+2.77
AcunityMINIPrinterSmall	Chronos-2	+0.45	+0.21	+1.63	+4.99	+0.02
	MantisV2	-0.10	+0.52	+4.17	+4.01	+1.62
	MOMENT	+2.62	+3.09	+5.93	+6.38	+1.34
	UniTS	+2.54	+3.22	+5.10	+3.48	+1.79

Continued on next page

Table 20 continued

Dataset	Encoder	3:1	5:1	10:1	20:1	Δ AUC
AsphaltRegularityUni	Chronos-2	+0.16	+1.71	+0.81	+3.33	+1.18
	MantisV2	-0.05	+0.29	+0.39	+0.71	+0.10
	MOMENT	+0.87	+0.79	+1.34	+0.86	+0.39
	UniTS	+0.03	-0.09	+1.26	+6.94	+5.24
DistalPhalanxOutlineCorrect	Chronos-2	+4.64	+5.76	+8.79	+7.07	+7.69
	MantisV2	+3.27	+3.98	+4.34	+4.45	+4.63
	MOMENT	+3.47	+6.75	+10.07	+11.29	+11.53
	UniTS	+3.45	+8.15	+11.75	+12.50	+12.19
FloodModeling1_disc	Chronos-2	+2.65	+6.16	+10.81	+4.07	+1.01
	MantisV2	+1.50	-0.32	+2.04	+2.30	+3.13
	MOMENT	+1.90	+2.45	+5.16	+5.05	+7.36
	UniTS	-0.68	+0.36	-1.31	+0.05	+14.49
FordA	Chronos-2	+0.81	+1.05	+2.94	+5.48	+2.77
	MantisV2	+0.86	+0.99	+1.63	+2.99	+1.06
	MOMENT	+0.37	+0.77	+1.23	+3.86	+0.97
	UniTS	+0.50	-0.19	+0.65	+1.99	+3.77
FordB	Chronos-2	+0.67	+1.97	+2.35	+4.52	+4.96
	MantisV2	+0.20	-0.24	+1.14	+3.65	+2.74
	MOMENT	+0.24	+1.49	+2.28	+3.87	+2.91
	UniTS	+0.26	-0.30	+1.18	+2.07	+6.26
HandOutlines	Chronos-2	+1.24	+2.56	+6.57	+6.10	+4.16
	MantisV2	+1.02	+3.21	+2.41	+6.81	+3.06
	MOMENT	+1.41	+2.84	+1.83	+2.92	+14.17
	UniTS	+0.94	+0.57	+0.64	+4.20	+3.14
MiddlePhalanxOutlineCorrect	Chronos-2	+1.37	+1.99	+4.60	+5.84	+6.03
	MantisV2	+0.05	+2.82	+7.87	+8.26	+5.03
	MOMENT	+3.20	+7.05	+7.26	+9.07	+1.82
	UniTS	+1.28	+1.86	+2.34	+5.75	-0.94
PhalangesOutlinesCorrect	Chronos-2	+1.42	+3.07	+5.05	+5.45	+9.81
	MantisV2	+4.04	+5.41	+6.84	+5.52	+6.94
	MOMENT	+2.89	+5.67	+6.36	+10.47	+12.61
	UniTS	+4.48	+6.36	+7.58	+6.07	+0.76
ProximalPhalanxOutlineCorrect	Chronos-2	+0.46	+2.06	+2.14	+5.17	+5.05
	MantisV2	+1.99	+3.35	+2.66	+2.80	+2.30
	MOMENT	+2.10	+1.86	+5.28	+5.69	+2.93
	UniTS	+1.70	+3.52	+1.78	+4.94	+4.66
SharePriceIncrease	Chronos-2	+0.28	+0.28	+0.40	+0.60	+6.24
	MantisV2	+0.97	+3.64	+2.00	+0.17	+7.62
	MOMENT	-0.48	-0.99	-0.14	-0.24	+1.47
	UniTS	-0.75	+0.87	+0.01	+0.18	+4.77
Strawberry	Chronos-2	-0.17	-0.01	+0.61	+1.97	+1.64
	MantisV2	+0.51	+0.25	+0.22	+1.46	+1.69
	MOMENT	+0.28	+0.00	-0.20	+0.59	+2.40
	UniTS	+0.69	+1.17	-0.09	+0.75	+0.81

I.3 Illustrative Examples and Their Selection

Figure 1a–b uses the FordB–MantisV2 example; Figures 6 and 7 retain all three examples to show the shape of train–test separation and the locations of anchor corrections. They were selected post hoc from an earlier exploratory inventory of 29 datasets and four encoders (116 pairs). The selection first required minority train recall of at least 90%, a minority train–test gap of at least 10 pp, and a full-TRAIN BA advantage of at least 2 pp at the scarce-support condition. Among eligible cases with at least 30 test examples per class, the procedure selected three distinct datasets and encoders nearest the median gap/gain vector within that phenomenon-defined subset. Correction displays then prioritized positive net gains among these examples.

For these three examples, the ten-seed mean net BA gains at $20{:}1$ are 5.84, 3.65, and 4.94 pp, compared with a median of 4.13 pp over all 52 seed-averaged combinations. Our aggregate conclusions use the complete binary diagnostic cohort in Figures 1 and 4; Table 20 reports every condition's net change.

The logit diagnostic always uses $C = 1$, whereas the SCALE internal anchor uses the regularization selected in each run, which can vary across seeds. In Figure 7, the left panels show the seed-2025 anchor distributions and corrections; the right panels summarize all ten seeds. Recovery ticks are shown at the *unchanged anchor logits*; they are not logits of SCALE, whose final prediction is determined by the voting rule.

In Figure 6, each original-TRAIN versus scarce-support logit comparison uses two panels that share axis limits covering all plotted scores and 33 equal-width histogram bins. Densities are normalized within each class. Test curves use Gaussian kernel density estimates with Scott's bandwidth; dotted segments mark the misclassified side of the zero-logit boundary.

Diagnostic interpretation. Under the specified readout protocol, decreasing minority support is associated with a larger train–test recall gap, while additional TRAIN labels improve average test ranking with the encoder held fixed. The saved SCALE fits quantify the recovery–harm trade-off of anchored correction. Controlled dimension changes and repeated minority-support sampling, with shared preprocessing and controlled regularization, offer ways to study how these patterns depend on representation size and support stability.

The complete figure collection accompanying our anonymous code release (`https://anonymous.4open.science/r/SCALE-DD85`) includes logit distributions and correction summaries for the complete 13-dataset binary cohort and four encoders (52 pairs, 208 ratio combinations), retaining all outcomes. The correction atlas provides one page per dataset–encoder pair, and the example collection includes all ten seeds for each of the three examples.

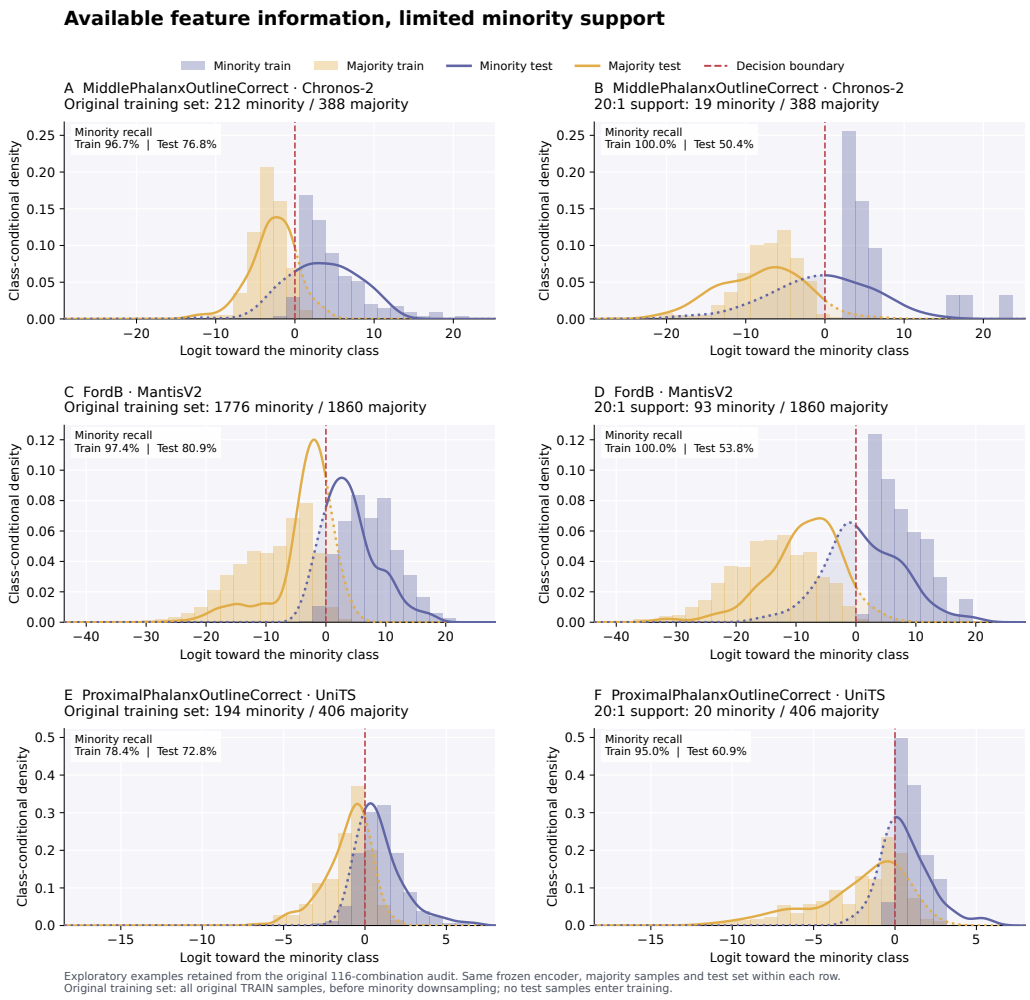


Figure 6: Post-hoc illustrations of minority train–test separation under class-balanced logistic readouts with $C = 1$. Each row uses the same frozen encoder, majority training examples, and test split; the left column uses the original TRAIN partition and the right column its 20:1 minority subset. Histograms show training scores and curves show test score densities. Original TRAIN is not a common imbalance ratio. The selection protocol is given in Appendix I.3.

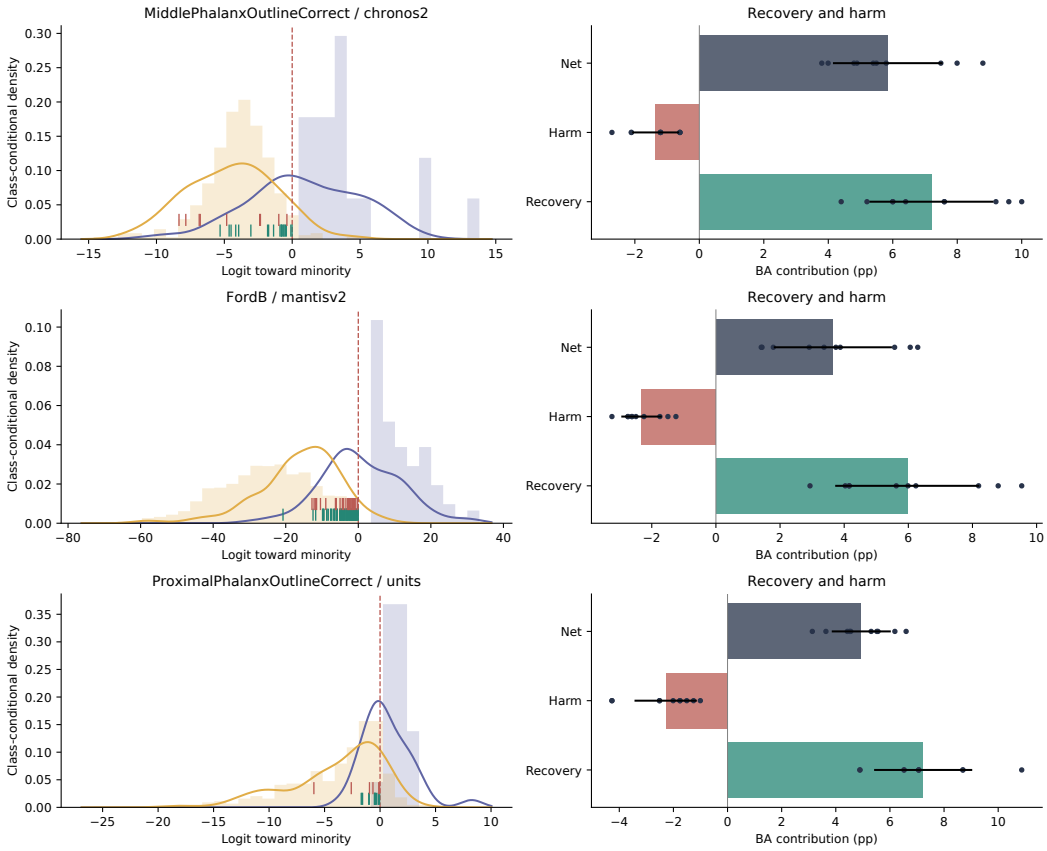


Figure 7: Illustrative anchor corrections at 20:1. Left: internal-anchor scores, with minority/majority distributions in purple/gold and recovered/harmed test examples marked in green/red at their unchanged logits. Histograms/curves show training/test score densities. Right: recovery, harm, and net BA change; bars show means, error bars show one standard deviation, and dots show individual runs. Table 20 reports all combinations; Appendix I.3 gives the example selection protocol.