INTEGRATING LABEL-GUIDED DECISION MODELING WITH CROSS-MODAL STATISTICAL ALIGNMENT FOR MEDICAL MULTIMODAL CLASS-INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical multimodal incremental learning aims to identify emerging diseases or subtypes while retaining knowledge of previously learned conditions to meet evolving diagnostic needs. Existing methods typically map multimodal information into a shared representation space and select classes based on similarity relationships among the extracted features. However, merely maintaining cross-modal alignment fails to model the discriminative space required for downstream classification tasks, making it difficult to effectively distinguish between visually similar categories. To address this, we propose a continual statistical learning framework based on frozen visual-language representations. We first introduce dimensional uncertainty modeling (U^t), which utilizes class-conditional statistical information to dynamically calibrate the reliability of visual features, providing a more reliable basis for subsequent cross-modal statistical organization. Building on this calibrated organization, we then address the more fundamental alignment-classification mismatch through a label-guided decision module, which integrates statistics linking features to class labels to recover missing category information and transforms organized representations into discriminative decision structures. By leveraging accumulated sufficient statistics, the framework achieves replay-free continual adaptation without requiring access to historical samples or further optimization of the pre-trained encoder. Experiments across six medical continual learning protocols demonstrate the dynamic evolution of feature reliability and sustained improvements in classification performance. Notably, decision geometry improvements far exceed gains in mean class recall (MCR), which averages recall equally across classes. This further underscores the necessity of employing label-conditioned decision modeling alongside cross-modal alignment.

1 INTRODUCTION

Medical multimodal class-incremental learning aims to learn new classes from sequential medical tasks while retaining previously acquired knowledge. Conventional methods typically adapt the entire network as new tasks arrive (Rebuffi et al., 2017; Yan et al., 2021). In contrast, CLIP and its medical variants provide transferable multimodal representations, shifting the focus from feature relearning to the continual organization of fixed pretrained representations (Radford et al., 2021; Wang et al., 2022b; Zhang et al., 2025; Li et al., 2026b). Medical foundation models also use routed experts for adaptation and retrieval guidance for continual learning (Nejatimanzari et al., 2026; Chen et al., 2026). However, variation among samples within each class leads to unequal reliability across visual feature dimensions. These reliability estimates evolve as new classes arrive. Moreover, cross-modal correlations learned during pretraining do not necessarily form the discriminative structures required for downstream classification, as illustrated in Fig. 1. Therefore, reliably organizing pretrained representations and transforming cross-modal correlations into class-discriminative relationships remain key challenges in medical multimodal class-incremental learning.

We discuss existing approaches through three components: parameter adaptation, representation organization, and decision modeling. Parameter adaptation methods learn new tasks through fine

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.

tuning or parameter efficient updates (Wang et al., 2022d;c; Liang & Li, 2024). Repeated forward and backward passes still incur considerable computational overhead. In contrast, representation organization methods freeze the pretrained model and adapt to new tasks by directly adjusting features and their relationships (McDonnell et al., 2023; Zhang et al., 2026a). However, these methods typically treat all feature dimensions uniformly, overlooking variations in feature reliability caused by the intra-class heterogeneity of medical data and the continual arrival of new classes. Decision modeling methods enhance class discrimination through classifier calibration, prototype modeling, or adjustments to decision boundaries. These procedures are often separate from upstream representation organization. Approaches based on alignment primarily model correlations between visual and textual features. These correlations alone do not directly encode relations between representations and class labels, leaving a gap in decision modeling. It is therefore essential to jointly model feature reliability and establish an explicit connection between cross-modal representation organization and downstream classification.

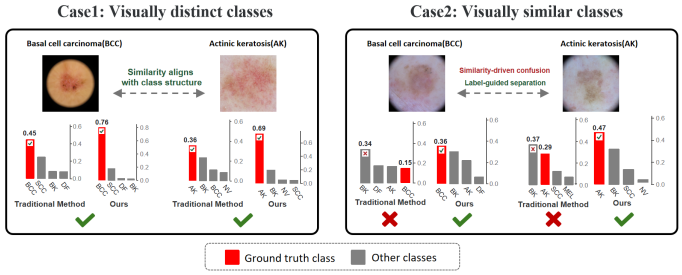

Figure 1: Motivation for label-guided discrimination. Cross-modal similarity establishes correspondence for visually distinct classes (Case 1) but lacks class-discriminative structure for visually similar classes (Case 2), where label information induces the required discriminative space.

To address these issues, we propose a continual statistical learning framework built on frozen vision-language representations. The Visual Uncertainty-Calibrated Alignment Module first estimates the uncertainty matrix U^t from statistics within each class. This matrix tracks evolving visual feature reliability and adaptively calibrates the visual statistical metric. It then organizes cross-modal relationships using second-order statistics between visual and textual representations. Building on this organization, the Label-Guided Decision Module uses statistics linking both modalities to class labels to learn class-discriminative decision relationships, bridging the structural information gap between alignment and classification. These stages share a recursive statistical update process without optimizing the pretrained model.

Our main contributions are as follows:

1. We identify how variation within each class and continual class arrival jointly affect visual feature reliability. We use U^t to update reliability estimates for each dimension as statistical evidence accumulates. These estimates adaptively calibrate the visual statistical metric.
2. We identify a structural information gap and objective mismatch between alignment and classification. Alignment statistics alone do not directly encode relations between representations and class labels. We extend cross-modal representation organization to label-conditioned decision modeling, enabling continual discriminative learning through a unified cumulative statistical update process.

2 Related Work

2.1 Class-Incremental Learning with Pretrained Models

Conventional baselines use rehearsal, classifier correction, or network expansion, as in DarkER++, WA, DynaER, and FOSTER (Buzzega et al., 2020; Zhao et al., 2020; Yan et al., 2021; Wang et al., 2022a). Large-scale pretrained models provide stable, transferable representations for medical class-incremental learning, shifting attention from learning from scratch to continually adapting pretrained knowledge. L2P introduces prompt learning into continual learning, enabling replay-free

adaptation with a frozen pretrained model and learnable prompts (Wang et al., 2022d). DualPrompt uses complementary prompts to capture task-shared and task-specific knowledge, while CODA-Prompt uses decomposable attention-based prompts to improve knowledge reuse and plasticity across tasks (Wang et al., 2022c; Smith et al., 2023). Beyond prompt learning, SLCA and EASE explore lightweight parameter adaptation, classifier adjustment, and task-dependent representation organization (Zhang et al., 2023; Zhou et al., 2024). Recent methods steer adapter updates or align representations across tasks (Zhao et al., 2026; Xu & Krawczyk, 2026). Together, these studies reflect a shift from relearning representations toward efficiently adapting and organizing increasingly capable pretrained features.

2.2 Multimodal Class-Incremental Learning and Cross-Modal Alignment

Vision-language pretraining adds textual semantics and cross-modal structure to continual learning. CLIP-based class-incremental methods have progressed from using text to assist visual adaptation toward continually organizing multimodal representations and their relations. RAPF uses textual semantics to model relations between old and new classes (Huang et al., 2024); C-CLIP extends continual learning to arriving image-text data (Liu et al., 2025); MG-CLIP preserves and compensates for the visual-textual modality gap during continual learning (Huang et al., 2025); and CCA-CL organizes visual-textual relations in frozen pretrained representations through accumulated second-order statistics (Zhang et al., 2026a). Pi-CCA studies certificates based on canonical correlations for continual multimodal learning (Zhang et al., 2026b). Recent work also models modality alignment as it evolves across tasks and network layers (Tan et al., 2026). CLAP4CLIP and PROOF adapt vision-language models for continual learning, while TPGCL uses textual prototypes in medical image classification (Jha et al., 2024; Zhou et al., 2025a;b). Text thus increasingly participates in continual representation organization and cross-modal relation modeling, beyond serving as a prior over class semantics.

2.3 Analytic and Statistical Continual Learning

Strong pretrained representations reduce traditional statistical learning's dependence on input feature quality. Cross-modal alignment further organizes these representations, enabling direct modeling of class knowledge in decision space. ACIL uses accumulated statistics for analytic decision modeling rather than iterative parameter updates (Zhuang et al., 2022), and GKEAL extends analytic learning to few-shot class-incremental settings (Zhuang et al., 2023). DS-AL introduces complementary analytic streams (Zhuang et al., 2024). RanPAC combines frozen pretrained representations with lightweight statistical classifiers (McDonnell et al., 2023), while RA-LDL combines lightweight representation calibration and analytic decision modeling for medical class-incremental learning (Wu et al., 2026). Theoretical work also examines memory and risk under structural regularization for two sequential regression tasks (Li et al., 2026a).

Although decision-space methods improve continual classification, they mainly target efficient classifier updates, forgetting mitigation, or knowledge compatibility (Zhuang et al., 2022; 2023; Wu et al., 2026). In multimodal learning, alignment-based approaches primarily model cross-modal feature relations, while the link between representation organization and label-driven decision modeling remains less explored. We therefore investigate linking cross-modal statistical organization to class-discriminative decision structure for continual classification.

3 Method

3.1 Overview

We consider T sequential medical classification tasks. Task t provides n_t samples $\mathcal{D}_t = \{(I_{t,i}, S_{t,i}, y_{t,i})\}_{i=1}^{n_t}$, each comprising a medical image $I_{t,i}$, its paired textual description $S_{t,i}$, and class label $y_{t,i}$. Task class sets $\mathcal{C}_t$ are pairwise disjoint, with $\mathcal{C}^{(1:t)} = \bigcup_{\tau=1}^{t} \mathcal{C}_\tau$ denoting all observed classes. Training accesses only $\mathcal{D}_t$, without revisiting historical samples. Accumulated statistics may be retained across tasks. At inference, the model uses only an image to predict the highest-scoring class in $\mathcal{C}^{(1:t)}$, without task identity.

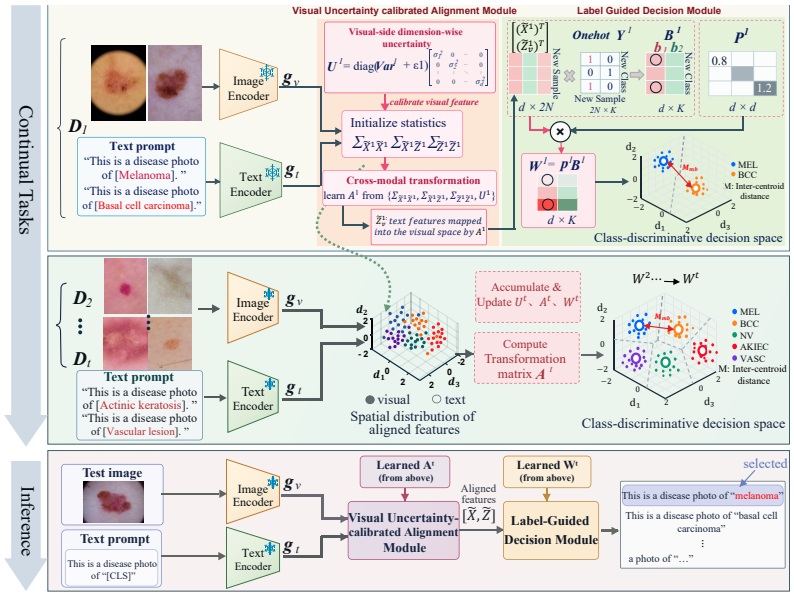

Figure 2: Overview of our framework across initial learning, continual updates, and inference. Visual uncertainty calibrates alignment, while label supervision shapes class decisions. Here, g_v expands visual feature dimensions through Gaussian random projection.

To leverage pretrained vision-language representations and transform cross-modal correlations into class-discriminative decision structures, we propose a cumulative statistical learning framework comprising Visual Uncertainty-Calibrated Alignment Module and Label-Guided Decision Module. Visual Uncertainty-Calibrated Alignment Module calibrates visual feature reliability and organizes cross-modal relationships, while Label-Guided Decision Module learns label-guided class-discriminative decision structures. All statistics are recursively updated without tuning the pretrained model or replaying historical samples. The framework is shown in Fig. 2.

3.2 Frozen Multimodal Feature Extraction and Expansion

We first construct fixed visual and textual representations for incremental classification. For task t, the frozen visual and text encoders, $f_v(\cdot)$ and $f_t(\cdot)$, extract feature matrices from paired images and textual descriptions:

$$V_t = f_v(\{I_{t,i}\}_{i=1}^{n_t}) \in \mathbb{R}^{n_t \times d_v}, \qquad S_t = f_t(\{S_{t,i}\}_{i=1}^{n_t}) \in \mathbb{R}^{n_t \times d_s}, \tag{1}$$

where the encoders are applied independently to each sample and their outputs are stacked row-wise in the same sample order. d_v and d_s denote the visual and textual feature dimensions, respectively. To facilitate the modeling of nonlinear structure, we apply random feature expansion to each modality:

$$X_t = g_v(V_t), \qquad Z_t = g_t(S_t), \tag{2}$$

where $X_t, Z_t \in \mathbb{R}^{n_t \times d_x}$ and d_x denotes the common expanded feature dimension. The mappings g_v and g_t transform the pretrained representations into higher-dimensional feature spaces through random projections followed by element-wise ReLU activations, as in analytic continual learning with random projection (McDonnell et al., 2023). The expanded features serve as the basis for the statistical computations in the following modules. Appendix B.1 specifies the fixed random projections and initialization.

3.3 Visual Uncertainty-Calibrated Alignment Module

Statistical continual learning models second-order relations within and across modalities, while visual feature reliability varies across dimensions, classes, and tasks. We first estimate dimension-wise uncertainty from within-class variability, then incorporate it into visual-side regularization while recursively accumulating cross-modal statistics.

Dimension-wise uncertainty estimation. To account for differences in visual feature reliability arising from intra-class heterogeneity, we model dimension-wise uncertainty using within-class variability across all observed classes. Specifically, the cumulative dimension-wise variance Var^t is defined as the equally weighted average of within-class variances over all classes observed up to task t:

$$\mathrm{Var}^t = \frac{1}{K^t} \sum_{\tau=1}^{t} \sum_{c \in \mathcal{C}_\tau} \frac{1}{n_{\tau,c} - 1} \sum_{i \in \mathcal{I}_{\tau,c}} (x_{\tau,i} - \mu_{\tau,c})^{\odot 2}, \tag{3}$$

where $x_{\tau,i} \in \mathbb{R}^{d_x}$ is the i-th expanded visual feature of task τ. For class c, $\mathcal{I}_{\tau,c} = \{i : y_{\tau,i} = c\}$ denotes its sample index set, $n_{\tau,c} = |\mathcal{I}_{\tau,c}|$ its sample count, and $\mu_{\tau,c}$ its mean feature vector. $K^t = |\mathcal{C}^{(1:t)}|$ denotes the cumulative number of observed classes, while $\odot 2$ indicates element-wise squaring. Since the task-specific class sets are disjoint, the sum of within-class variances over previously observed classes is given by $K^{t-1}\mathrm{Var}^{t-1}$. The cumulative variance is therefore updated by adding the within-class variances of the newly introduced classes and normalizing by the cumulative number of classes:

$$\mathrm{Var}^t = \frac{1}{K^t} \left[K^{t-1}\mathrm{Var}^{t-1} + \sum_{c \in \mathcal{C}_t} \frac{1}{n_{t,c} - 1} \sum_{i \in \mathcal{I}_{t,c}} (x_{t,i} - \mu_{t,c})^{\odot 2} \right], \tag{4}$$

where $K^t = K^{t-1} + |\mathcal{C}_t|$. Each entry Var_j^t represents the within-class variability of the j-th visual feature dimension, averaged equally across all observed classes. Based on this variance vector, we construct the dimension-wise uncertainty matrix:

$$U^t = \mathrm{diag}\big(\mathrm{Var}^t + \epsilon \mathbf{1}\big), \tag{5}$$

where $\epsilon > 0$ ensures numerical stability. A larger diagonal entry U_{jj}^t indicates greater within-class variability in the corresponding visual feature dimension and thus higher estimated uncertainty. This matrix provides dimension-specific regularization weights for the subsequent cross-modal transformation, imposing stronger regularization on dimensions with higher estimated uncertainty. Thus, $\mathrm{Var}^t \in \mathbb{R}^{d_x}$, $U^t \in \mathbb{R}^{d_x \times d_x}$, and $\mathbf{1} \in \mathbb{R}^{d_x}$.

Cumulative cross-modal statistics. The uncertainty matrix captures variability across visual dimensions, while cross-modal mapping also requires within-modality feature dependencies and cross-modal associations. Let $N_t = \sum_{\tau=1}^{t} n_\tau$ denote the accumulated sample count. We describe these relationships using centered scatter matrices, whose visual and cross-modal forms are

$$\begin{aligned} \sum\nolimits_{\widetilde{x}\widetilde{x}}^{t} &= \sum_{i=1}^{N_t} (x_i - \mu_x^t)(x_i - \mu_x^t)^\top, \\ \sum\nolimits_{\widetilde{x}\widetilde{z}}^{t} &= \sum_{i=1}^{N_t} (x_i - \mu_x^t)(z_i - \mu_z^t)^\top, \end{aligned} \tag{6}$$

where x_i and z_i denote the i-th pair of visual and textual features among all samples observed up to task t, while μ_x^t and μ_z^t are their corresponding global means. Dividing these scatter matrices by $N_t - 1$ yields the covariance matrices used for subsequent cross-modal alignment.

However, direct evaluation of these definitions requires access to all observed features. We first expand the centered terms to express the scatter matrices through uncentered second-order statistics and global means, providing the basis for recursive computation:

$$\begin{aligned} \sum\nolimits_{\widetilde{x}\widetilde{x}}^{t} &= \Sigma_{xx}^t - N_t \mu_x^t (\mu_x^t)^\top, \\ \sum\nolimits_{\widetilde{x}\widetilde{z}}^{t} &= \Sigma_{xz}^t - N_t \mu_x^t (\mu_z^t)^\top, \end{aligned} \tag{7}$$

where $\Sigma_{xx}^t = \sum_{i=1}^{N_t} x_i x_i^\top$ and $\Sigma_{xz}^t = \sum_{i=1}^{N_t} x_i z_i^\top$ are the cumulative uncentered visual and cross-modal second-order statistics, respectively. Their additive structure enables recursive updates with current-task increments:

$$\begin{aligned} \Sigma_{xx}^t &= \Sigma_{xx}^{t-1} + \Delta\Sigma_{xx}, \qquad \Delta\Sigma_{xx} = X_t^\top X_t, \\ \Sigma_{xz}^t &= \Sigma_{xz}^{t-1} + \Delta\Sigma_{xz}, \qquad \Delta\Sigma_{xz} = X_t^\top Z_t. \end{aligned} \tag{8}$$

Similarly, the textual statistic Σ_{zz}^t is updated with increment $Z_t^\top Z_t$. Global means are updated by weighting historical and current-task means by their sample counts. The same mean correction then reconstructs centered scatter matrices from these means and accumulated statistics without revisiting historical features. Appendix A gives the complete additive state, while Appendices A.1 and A.2 analyze update invariance and storage.

Transformation based on uncertainty. The second-order statistics $\frac{1}{N_t-1}\sum_{\widetilde{x}\widetilde{x}}^t$, $\frac{1}{N_t-1}\sum_{\widetilde{z}\widetilde{z}}^t$, and $\frac{1}{N_t-1}\sum_{\widetilde{x}\widetilde{z}}^t$ characterize intra- and cross-modal relationships. We incorporate U^t into the visual-side regularization to construct the cross-modal transformation matrix A^t, adaptively weighting visual dimensions according to their reliability

$$A^t = \left(\frac{1}{N_t-1}\sum\nolimits_{\widetilde{x}\widetilde{x}}^t + \lambda U^t\right)^{-1} \frac{1}{N_t-1}\sum\nolimits_{\widetilde{x}\widetilde{z}}^t \times \left(\frac{1}{N_t-1}\sum\nolimits_{\widetilde{z}\widetilde{z}}^t + \lambda I\right)^{-1}, \tag{9}$$

where $\lambda > 0$ is the regularization coefficient and $I = I_{d_x}$ is the identity matrix. The textual modality uses uniform regularization through λI, while λU^t imposes stronger constraints on visual dimensions with higher estimated uncertainty. As new tasks arrive, the transformation matrix is recomputed from the updated uncertainty estimates and cumulative statistics, which captures both the newly observed classes and the reliable alignment from previous tasks. Appendix B.2 derives the equivalent direct form and reports the runtime comparison in Appendix B.2.2.

3.4 Label-Guided Decision Module

The preceding module organizes cross-modal statistical relationships through uncertainty calibration. However, such alignment does not directly impose class-discriminative constraints. Appendix C formalizes this limitation. To bridge this structural information gap, we introduce the Label-Guided Decision Module to learn a shared label-guided discriminative mapping. Specifically, A^t maps centered textual features into the visual space, yielding $\widetilde{Z}_v^t$, which are jointly used with centered visual features to learn the classifier.

Let $Y^t \in \mathbb{R}^{N_t \times K^t}$ denote the one-hot label matrix of all observed samples. We formulate the following shared ridge regression objective (Hoerl & Kennard, 1970):

$$\min_W \frac{\|\widetilde{X}^t W - Y^t\|_F^2 + \alpha\|\widetilde{Z}_v^t W - Y^t\|_F^2}{2N_t - 1} + \beta\|W\|_F^2, \tag{10}$$

where $W \in \mathbb{R}^{d_x \times K^t}$ is the optimization variable, with stage t solution W^t, $\alpha \geq 0$ balances textual representations, and $\beta > 0$ controls regularization. The objective jointly learns the class mapping from visual and textual representations through label-conditioned statistics.

The preceding module already maintains the required feature statistics. We additionally accumulate the products of features and labels. For the visual modality, the update is

$$\Sigma_{xy}^t = \Sigma_{xy}^{t-1} + X_t^\top Y_t, \tag{11}$$

where Y_t is the one-hot label matrix of the current task, with columns corresponding to all observed classes. The textual statistic Σ_{zy}^t is updated using the same procedure, with contribution $Z_t^\top Y_t$. The cumulative class counts p_y^t are maintained alongside these statistics, with zero padding for newly introduced classes. Together with the cumulative feature means, these counts support the same mean correction principle used above, yielding the centered products $\sum_{\widetilde{x}y}^t$ and $\sum_{\widetilde{z}y}^t$.

Setting the gradient of the objective with respect to W to zero and expressing the transformed textual terms through A^t and the corresponding statistics yields the closed-form solution:

$$W^t = \left[\sum\nolimits_{\widetilde{x}\widetilde{x}}^t + \alpha A^t \sum\nolimits_{\widetilde{z}\widetilde{z}}^t (A^t)^\top + \beta(2N_t - 1)I\right]^{-1} \times \left[\sum\nolimits_{\widetilde{x}y}^t + \alpha A^t \sum\nolimits_{\widetilde{z}y}^t\right]. \tag{12}$$

The full derivation is provided in Appendix B.3; Appendix D analyzes the resulting class directions and boundaries. We write this solution as $W^t = P^t B^t$. Here, $P^t \in \mathbb{R}^{d_x \times d_x}$ is the inverse of the first bracketed matrix. The second bracketed matrix, $B^t \in \mathbb{R}^{d_x \times K^t}$, contains statistics linking features to class labels. This matrix is unrelated to the scalar geometry metric B defined in Appendix E. As new tasks arrive, W^t is updated in closed form using accumulated statistics and A^t. At inference, W^t maps centered visual representations to class scores.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets, protocols, and metrics. We evaluate on six medical image classification datasets: Skin8, Skin40, Path16, MedMNIST (Yang et al., 2023), COVID, and Blood. The detailed continual-learning protocols are provided in Appendix G.1.

Here, T is the total number of tasks in the class-incremental sequence. At each stage, evaluation covers all classes observed so far. We use MCR to give equal weight to each observed class. Recall is the fraction of a class's test samples classified correctly. Last denotes MCR after the final task. Avg denotes the arithmetic mean of MCR across all task stages. Mechanistic analyses and ablations primarily report Last, while overall comparisons report both metrics.

Decision geometry metrics. Beyond MCR, we evaluate the learned decision structure through class separation and boundary geometry. Projected center separation M measures separation between class centers along classifier-defined directions. Center safety margin C measures each center's smallest signed margin against competing classes. Class-balanced sample margin B averages the smallest signed sample margins with equal class weights. Boundary violation rate V measures the proportion of samples on or beyond competing decision boundaries. The detailed definitions are provided in Appendix E.

Implementation details. Our method and its ablations use frozen **OpenCLIP ViT-B/16** image and text encoders (Cherti et al., 2023). Unless otherwise specified, the expanded dimension is $d_x = 5120$, and the regularization coefficients for the Visual Uncertainty-Calibrated Alignment Module and the Label-Guided Decision Module are $\lambda = \beta = 1$. The modality fusion weight is selected from $\alpha \in \{0.25, 0.5\}$. Hyperparameters are selected using only a training-validation split within the current task's training data. Test data are not used for model selection or tuning. Our continual updates follow the replay-free setting defined in Section 3.1; baseline results use their stated protocols.

4.2 FROM ACCESSIBLE CLASS INFORMATION TO DECISION GEOMETRY

We examine three increasingly specific properties of frozen vision-language representations in continual medical classification: whether the features extracted by the frozen model contain rich class information, whether this information has uniform and stable reliability, and whether feature alignment suffices to determine downstream class decisions after reliability modeling and cross-modal organization.

4.2.1 ACCESSIBLE CLASS INFORMATION IN FROZEN REPRESENTATIONS

To disentangle representation quality from the zero-shot matching rule, we apply Nearest Class Center (NCC) (Mensink et al., 2013) as a simple readout of the same frozen visual features, and then examine whether cross-modal statistical organization can further exploit the retained class information.

Mean Last MCR rises from 12.92% with zero-shot prediction to 37.77% with NCC, showing that frozen visual features already retain accessible class information. Cross-modal alignment further raises it to 58.28% by organizing visual-textual relations, indicating that this information is present but not yet effectively organized.

4.2.2 HETEROGENEOUS AND EVOLVING FEATURE DIMENSION RELIABILITY

Frozen visual features retain class information, but their within-class variance differs across dimensions. Figure 3(a) shows this variation for the Path16 breast-cancer class. For this diagnostic, we

use pre-expansion visual features $v \in \mathbb{R}^{d_v}$ only to inspect the variance profile. For class c, let $\mathrm{Var}_{v,c} \in \mathbb{R}^{d_v}$ denote its dimension-wise within-class variance. At task t, we average these vectors over $\mathcal{C}_t$ and over $\mathcal{C}^{(1:t)}$ to obtain task-specific and cumulative profiles. Figure 3(b) reports the 298th component on the COVID dataset. The model matrix U^t is computed from expanded features $x = g_v(v)$ in Section 3. Further cross-dataset and temporal analyses appear in Appendices F.3 and F.4.

These observations motivate dynamic dimension-wise reliability calibration rather than a fixed, uniform reliability assumption.

4.2.3 Alignment-Classification Objective Mismatch

Our theory shows that even perfect cross-modal alignment cannot determine Bayes-optimal classification performance: image-text correspondence leaves class-discriminative directions underdetermined. However, MCR cannot reflect geometric changes among already separated samples. We therefore compare models with and without the Label-Guided Decision Module using the decision-geometry metrics in Appendix E, keeping all representation-side components fixed.

Figure 4 shows that adding label-conditioned modeling substantially increases M, C, and B, reduces V from 0.2757 to 0.2189, and improves mean Last MCR by 5.89 percentage points. These changes indicate that label-conditioned information reshapes relations among classes, boundaries, and samples beyond the discrete corrections captured by MCR. The results support using label-conditioned statistics to transform cross-modal correspondence into class-discriminative decision structure, rather than merely strengthening image-text alignment. Appendices F.1 and F.2 provide cross-dataset geometry analyses and representative decision corrections. Additional robustness, order, and capacity analyses are reported in Appendices G.2, G.3, and G.4.

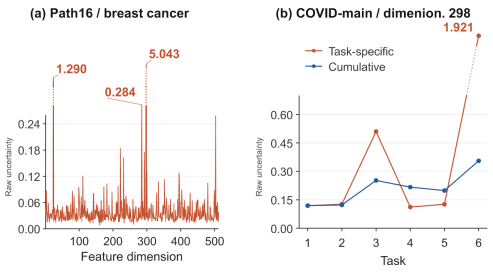

Figure 3: Dimension-wise uncertainty within a class (left) and across tasks (right).

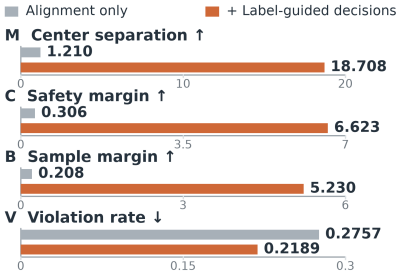

Figure 4: Decision geometry with and without label guidance. Each metric uses its own scale.

4.3 Overall Comparison with Existing Methods

We next compare the full method with existing approaches across all six medical continual-learning datasets.

Our method achieves the highest Last MCR on all six datasets. It also obtains the best mean Last and Avg MCR among fully evaluated methods. These averages are 80.27% and 85.87%, compared with 76.10% and 83.55% for CCA-CL.

4.4 Component Ablations and Contributions

Table 3 reports cumulative ablations across six datasets.

Table 1: Comparison on six datasets: Last and Avg MCR (%).

Method	Skin8		Skin40		Path16		MedMNIST		COVID		Blood		Mean	
	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg	Last	Avg
iCaRL	$24.68_{\pm 0.21}$	$50.97_{\pm 0.17}$	$31.83_{\pm 2.74}$	$66.90_{\pm 0.54}$	$52.58_{\pm 1.08}$	$68.12_{\pm 1.77}$	$20.21_{\pm 0.99}$	$51.27_{\pm 0.16}$	$45.27_{\pm 1.16}$	$59.52_{\pm 3.49}$	$59.93_{\pm 1.61}$	$76.42_{\pm 0.76}$	39.08	62.20
DarkER++	$25.05_{\pm 0.19}$	$49.86_{\pm 0.71}$	$18.42_{\pm 1.66}$	$53.12_{\pm 2.89}$	$54.55_{\pm 1.24}$	$62.48_{\pm 2.37}$	$27.44_{\pm 2.84}$	$53.55_{\pm 0.40}$	$36.01_{\pm 4.61}$	$55.96_{\pm 0.86}$	$59.72_{\pm 4.00}$	$77.33_{\pm 1.36}$	36.87	58.72
WA	$24.49_{\pm 0.52}$	$48.86_{\pm 0.13}$	$15.58_{\pm 0.88}$	$47.90_{\pm 3.08}$	$44.10_{\pm 2.33}$	$66.38_{\pm 1.57}$	$11.63_{\pm 0.81}$	$33.01_{\pm 2.22}$	$42.37_{\pm 3.36}$	$60.08_{\pm 5.11}$	$58.42_{\pm 2.49}$	$77.15_{\pm 2.88}$	32.77	55.56
DynaER	$26.60_{\pm 1.83}$	$48.91_{\pm 1.89}$	$21.08_{\pm 0.52}$	$48.19_{\pm 0.95}$	$52.11_{\pm 5.07}$	$59.48_{\pm 3.33}$	$31.46_{\pm 2.25}$	$59.43_{\pm 1.19}$	$49.05_{\pm 1.50}$	$67.23_{\pm 1.97}$	$74.17_{\pm 0.56}$	$83.27_{\pm 1.72}$	42.41	61.09
FOSTER	$24.92_{\pm 0.58}$	$51.40_{\pm 0.99}$	$26.83_{\pm 1.81}$	$65.22_{\pm 2.44}$	$53.56_{\pm 2.43}$	$66.55_{\pm 4.24}$	$30.19_{\pm 4.47}$	$53.13_{\pm 2.20}$	$16.34_{\pm 1.79}$	$43.68_{\pm 2.17}$	$33.95_{\pm 6.76}$	$65.15_{\pm 2.87}$	30.97	57.52
CLAP4CLIP	$32.22_{\pm 1.69}$	$53.94_{\pm 1.41}$	$48.67_{\pm 0.29}$	$64.75_{\pm 0.94}$	$58.67_{\pm 1.19}$	$62.96_{\pm 1.07}$	$8.33_{\pm 3.54}$	$27.03_{\pm 4.72}$	$22.45_{\pm 2.83}$	$25.36_{\pm 4.54}$	$12.73_{\pm 2.92}$	$28.17_{\pm 5.14}$	30.51	43.70
PROOF	$28.89_{\pm 2.91}$	$55.18_{\pm 1.32}$	$47.67_{\pm 1.01}$	$66.38_{\pm 0.97}$	$55.68_{\pm 0.58}$	$68.70_{\pm 0.20}$	$43.40_{\pm 1.13}$	$62.44_{\pm 1.32}$	$79.60_{\pm 1.58}$	$88.68_{\pm 0.51}$	$74.75_{\pm 0.63}$	$83.59_{\pm 0.49}$	55.00	70.83
TPGCL	$45.38_{\pm 1.67}$	$61.30_{\pm 0.59}$	$55.67_{\pm 0.63}$	$75.38_{\pm 0.37}$	$64.72_{\pm 0.85}$	$72.09_{\pm 0.28}$	-	-	-	-	-	-	-	-
CCA-CL	$\underline{65.19}^{*}$	$\underline{75.08}^{*}$	$\underline{73.50}^{*}$	$\underline{84.08}^{*}$	$\underline{71.41}^{*}$	$\mathbf{73.47}^{*}$	$67.89_{\pm 0.03}$	$\underline{80.06}_{\pm 0.04}$	$\underline{88.73}_{\pm 0.18}$	$\underline{95.24}_{\pm 0.03}$	$\underline{89.90}_{\pm 0.20}$	$\underline{93.34}_{\pm 0.13}$	$\underline{76.10}$	$\underline{83.55}$
Ours	$\mathbf{67.83}^{*}$	$\mathbf{75.96}^{*}$	$\mathbf{79.25}^{*}$	$\mathbf{88.02}^{*}$	$\mathbf{73.35}^{*}$	$\underline{73.40}^{*}$	$\mathbf{73.53}_{\pm 0.19}$	$\mathbf{84.31}_{\pm 0.08}$	$\mathbf{93.96}_{\pm 0.23}$	$\mathbf{97.23}_{\pm 0.14}$	$\mathbf{93.72}_{\pm 0.23}$	$\mathbf{96.31}_{\pm 0.18}$	**80.27**	**85.87**

Values after ± report standard deviations. Bold and underline mark the best and second-best means in each column. * CCA-CL and Ours report point estimates without standard deviations for Skin8, Skin40, and Path16. For MedMNIST, COVID, and Blood, the reported values summarize three rerun seeds. Mean averages the six dataset means equally and has no standard deviation.

Table 2: Accessible class information retained in frozen pretrained representations. Mean Last MCR is averaged across six datasets.

Method	Mean Last MCR (%)
Zero-shot	12.92
NCC	37.77
Cross-modal alignment	**58.28**

Table 3: Component ablation with cumulative additions.

Configuration	Mean Last MCR (%)
Cross-modal alignment	58.28
+ Uncertainty calibration	71.75
+ Random Anchor	74.42
+ Label-guided decisions	**80.31**

Adding U^t improves mean Last MCR by 13.47 percentage points, supporting dimension-wise reliability calibration in cross-modal statistical organization. Random Anchor adds 2.67 points through feature expansion.

After both components, the Label-Guided Decision Module provides a further 5.89 percentage point gain. This improvement suggests that reliability calibration and feature expansion alone leave room for label-conditioned decision modeling, which introduces class-discriminative constraints through statistics linking features to class labels. Additional robustness, order, and capacity analyses are reported in Appendices G.2, G.3, and G.4.

5 Conclusion

This work identifies two remaining challenges in replay-free continual learning with frozen medical vision-language models: feature reliability evolves with class arrival, while cross-modal alignment leaves label-conditioned discrimination underdetermined, reflecting a **structural information gap and objective mismatch**. We address them through dynamic reliability calibration and label-guided decision modeling within a cumulative statistical framework.

Across six medical datasets, experiments confirm evolving reliability and show that decision-geometry gains are markedly more pronounced than the corresponding MCR improvement would suggest. These findings suggest that continual multimodal learning should not end with cross-modal alignment alone, but further connect organized representations with label-conditioned decision structure.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting, translating, and editing the manuscript; LaTeX formatting and reference organization; the organization and interpretation of existing experimental results; and the checking and revision of code and figures. All AI-assisted material was reviewed and verified by the authors. The authors take full responsibility for the final text, theoretical statements, experimental results, citations, and figures in this work.

ETHICS STATEMENT

We evaluate the method on existing medical image classification datasets. These benchmark results do not establish clinical safety or performance across patient groups. Any clinical use would require external validation, privacy safeguards, subgroup assessment, and clinician oversight.

REPRODUCIBILITY STATEMENT

Section 3 defines the task and analytic method. Appendices A and B give the cumulative updates and derivations. Section 4 reports the datasets, metrics, and implementation settings. Appendix G.1 lists the task protocols.

REFERENCES

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930. Curran Associates, Inc., 2020.

Zizhi Chen, Yizhen Gao, Minghao Han, Yizhou Liu, Zhaoyu Chen, Dingkang Yang, and Lihua Zhang. Forging a dynamic memory: Retrieval-guided continual learning for generalist medical foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32309–32321, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. doi: 10.1162/0899766042321814.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in Statistics: Methodology and Distribution*, pp. 162–190. Springer, 1992. doi: 10.1007/978-1-4612-4380-9_14.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with CLIP: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pp. 214–231, 2024. doi: 10.1007/978-3-031-72949-2_13.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in CLIP-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3777–3786, 2025.

Saurav Jha, Dong Gong, and Lina Yao. CLAP4CLIP: Continual learning with probabilistic finetuning for vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 129146–129186. Curran Associates, Inc., 2024. doi: 10.52202/079017-4102.

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Memory-statistics tradeoff in continual learning with structural regularization. In *International Conference on Learning Representations*, 2026a.

Zihan Li, Yiqing Wang, Sina Farsiu, and Paul Kinahan. Boosting medical visual understanding from multi-granular language learning. In *International Conference on Learning Representations*, 2026b.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23638–23647, 2024.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-CLIP: Multimodal continual learning for vision-language model. In *International Conference on Learning Representations*, 2025.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 12022–12053. Curran Associates, Inc., 2023. doi: 10.52202/075280-0526.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2624–2637, 2013. doi: 10.1109/tpami.2013.83.

Omid Nejatimanzari, Hojat Asgariandehkordi, Taha Koleilat, Yiming Xiao, and Hassan Rivaz. Sparse spectral LoRA: Routed experts for medical VLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35351–35362, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, 2017.

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

Jiayao Tan, Fan Lyu, Tianle Liu, Fuyuan Hu, and Wei Feng. Towards dynamic modality alignment in multimodal continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39911–39921, 2026.

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. FOSTER: Feature boosting and compression for class-incremental learning. In *European Conference on Computer Vision*, pp. 398–414, 2022a. doi: 10.1007/978-3-031-19806-9_23.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887. Association for Computational Linguistics, 2022b. doi: 10.18653/v1/2022.emnlp-main.256.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648, 2022c. doi: 10.1007/978-3-031-19809-0_36.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 139–149, 2022d.

Xinyao Wu, Zhe Xu, and Raymond Kai-yu Tong. Random anchors with low-rank decorrelated learning: A minimalist pipeline for class-incremental medical image classification. In *International Conference on Learning Representations*, 2026.

Hongye Xu and Bartosz Krawczyk. Two-way is better than one: Bidirectional alignment with cycle consistency for exemplar-free class-incremental learning. In *International Conference on Learning Representations*, 2026.

Shipeng Yan, Jiangwei Xie, and Xuming He. DER: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3014–3023, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19148–19158, 2023.

Huan Zhang, Shuyu Dong, Yujin Zheng, Dingwen Wang, Shenghua Fan, and Fan Lyu. Subspace alignment for CLIP-based continual learning via canonical correlation analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32257–32266, 2026a.

Jiayu Zhang, Chuangxin Zhao, Canran Xiao, Ruibo Duan, Wenyi Mo, Haoyu Gao, and Wenshuo Wang. Pi-CCA: Prompt-invariant CCA certificates for replay-free continual multimodal learning. In *International Conference on Learning Representations*, 2026b.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image-text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. doi: 10.1056/aioa2400640.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13208–13217, 2020.

Jiarui Zhao, Libo Huang, Xiangqi Li, Zhulin An, Chuanguang Yang, Yu Wang, Boyu Diao, and Yongjun Xu. Representation-steered incremental adapter-tuning for class-incremental learning with pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18010–18020, 2026.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23554–23564, 2024.

Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4489–4504, 2025a. doi: 10.1109/tpami.2025.3540889.

Zhiping Zhou, Tong Yu, Yizhe Zhang, Wei-Shi Zheng, and Ruixuan Wang. Textual prototype-guided continual learning for medical image classification. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1978–1983, 2025b. doi: 10.1109/bibm66473.2025.11356104.

Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. ACIL: Analytic class-incremental learning with absolute memorization and privacy protection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11602–11614. Curran Associates, Inc., 2022. doi: 10.52202/068431-0843.

Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, and Ziqian Zeng. GKEAL: Gaussian kernel embedded analytic learning for few-shot class incremental task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7746–7755, 2023.

Huiping Zhuang, Run He, Kai Tong, Ziqian Zeng, Cen Chen, and Zhiping Lin. DS-AL: A dual-stream analytic learning for exemplar-free class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17237–17244, 2024. doi: 10.1609/aaai.v38i15.29670.

Appendix Contents

A Continual Statistical Learning without Replay 15

A.1 Order and Partition Invariance of the Statistical State 15

A.2 Storage Complexity 16

B Detailed Method Formulation 16

B.1 Random Feature Expansion 16

B.2 Visual Uncertainty-Calibrated Alignment Module 16

B.2.2 Runtime Efficiency 19

B.3 Label-Guided Decision Module 19

C Theoretical Limits of Across Modalities Alignment for Classification 20

D Theoretical Analysis of Label-Guided Decision Module 22

D.1 Pairwise Discriminative Directions from Labels 22

D.2 Covariance Scale Adjustment from Joint Statistics 23

D.3 Connection to the Final Decision Boundary 23

E Decision Geometry Metrics 24

F Extended Analysis of Decision Geometry and Uncertainty 24

F.1 Decision Geometry Changes across Datasets 24

F.2 Representative Decision Correction Cases 26

F.3 Temporal Variation across Dimensions 28

F.4 Uncertainty under Continual Class Arrival 28

G Additional Experimental Analysis 30

G.1 Dataset and Continual Learning Protocols 30

G.2 Robustness to Random Anchors Dimension and Initialization 30

G.3 Consistency across Sample Orders and Data Partitions 31

G.4 Data Accumulation and Representation Capacity 32

A CONTINUAL STATISTICAL LEARNING WITHOUT REPLAY

Let Γ^t collect the additive statistics accumulated through task t. Let $\Delta\Gamma_t$ denote the contribution from task t. We set $\Gamma^0 = 0$ and update the state as follows:

$$\Gamma^t = \Gamma^{t-1} + \Delta\Gamma_t. \tag{A.1}$$

Using the quantities defined in the Method section, the state is

$$\Gamma^t = \Big(\mathbf{m}_x^t, \mathbf{m}_z^t, \Sigma_{xx}^t, \Sigma_{zz}^t, \Sigma_{xz}^t, K^t\mathrm{Var}^t, \Sigma_{xy}^t, \Sigma_{zy}^t, p_y^t\Big). \tag{A.2}$$

Here, $\mathbf{m}_x^t$ and $\mathbf{m}_z^t$ are the cumulative feature sums. The vector $p_y^t \in \mathbb{R}^{K^t}$ contains the class counts. The term $K^t\mathrm{Var}^t$ is the additive sum underlying the variance update in the Method section. All state components are initialized elementwise to zero, including $K^0\mathrm{Var}^0 = 0$. We use a common class order for Y_t and zero pad historical arrays when new classes arrive.

Let $\mathbf{1}_{n_t}$ denote an all ones vector of length n_t. The feature sums and class counts update as follows:

$$\begin{aligned} \mathbf{m}_x^t &= \mathbf{m}_x^{t-1} + X_t^\top \mathbf{1}_{n_t}, \\ \mathbf{m}_z^t &= \mathbf{m}_z^{t-1} + Z_t^\top \mathbf{1}_{n_t}, \\ p_y^t &= p_y^{t-1} + Y_t^\top \mathbf{1}_{n_t}. \end{aligned} \tag{A.3}$$

Here $Y_t \in \mathbb{R}^{n_t \times K^t}$ is embedded in the current cumulative class space. Its absent-class columns are zero, and new columns are appended when new classes arrive. The remaining additive components update as

$$\begin{aligned} &\Sigma_{xx}^t = \Sigma_{xx}^{t-1} + X_t^\top X_t, \qquad \Sigma_{zz}^t = \Sigma_{zz}^{t-1} + Z_t^\top Z_t, \\ &\Sigma_{xz}^t = \Sigma_{xz}^{t-1} + X_t^\top Z_t, \qquad \Sigma_{xy}^t = \Sigma_{xy}^{t-1} + X_t^\top Y_t, \\ &\Sigma_{zy}^t = \Sigma_{zy}^{t-1} + Z_t^\top Y_t, \quad K^t\mathrm{Var}^t = K^{t-1}\mathrm{Var}^{t-1} + \sum_{c\in\mathcal{C}_t} \frac{1}{n_{t,c}-1} \sum_{i\in\mathcal{I}_{t,c}} (x_{t,i} - \mu_{t,c})^{\odot 2}. \end{aligned} \tag{A.4}$$

We recover the sample count and global means from the state:

$$N_t = \mathbf{1}_{K^t}^\top p_y^t, \qquad \mu_x^t = \frac{\mathbf{m}_x^t}{N_t}, \qquad \mu_z^t = \frac{\mathbf{m}_z^t}{N_t}. \tag{A.5}$$

These quantities need no separate storage.

The centered scatter formulas in the main text follow by expanding their products. For the visual scatter,

$$\begin{aligned} \sum\nolimits_{\widetilde{x}\widetilde{x}}^t &= \sum_{i=1}^{N_t} (x_i - \mu_x^t)(x_i - \mu_x^t)^\top \\ &= \sum_{i=1}^{N_t} x_i x_i^\top - \mathbf{m}_x^t(\mu_x^t)^\top \\ &\quad - \mu_x^t(\mathbf{m}_x^t)^\top + N_t\mu_x^t(\mu_x^t)^\top \\ &= \Sigma_{xx}^t - N_t\mu_x^t(\mu_x^t)^\top. \end{aligned} \tag{A.6}$$

The cross modality scatter follows in the same way:

$$\begin{aligned} \sum\nolimits_{\widetilde{x}\widetilde{z}}^t &= \sum_{i=1}^{N_t} (x_i - \mu_x^t)(z_i - \mu_z^t)^\top \\ &= \sum_{i=1}^{N_t} x_i z_i^\top - \mathbf{m}_x^t(\mu_z^t)^\top \\ &\quad - \mu_x^t(\mathbf{m}_z^t)^\top + N_t\mu_x^t(\mu_z^t)^\top \\ &= \Sigma_{xz}^t - N_t\mu_x^t(\mu_z^t)^\top. \end{aligned} \tag{A.7}$$

The final equalities use $\mathbf{m}_x^t = N_t \mu_x^t$ and $\mathbf{m}_z^t = N_t \mu_z^t$.

Mean correction also recovers the centered products of features and labels used by W^t:

$$\begin{aligned} \sum\nolimits_{\widetilde{x}y}^t &= \Sigma_{xy}^t - \mu_x^t (p_y^t)^\top, \\ \sum\nolimits_{\widetilde{z}y}^t &= \Sigma_{zy}^t - \mu_z^t (p_y^t)^\top. \end{aligned} \tag{A.8}$$

A.1 Order and Partition Invariance of the Statistical State

Let $\mathcal{D}^t = \bigcup_{\tau=1}^t \mathcal{D}_\tau$, where $\mathcal{D}_\tau$ is defined in Section 3.1. Consider any order within each task and any partition of each task into task-local disjoint blocks.

Proposition 1 (Order and partition invariance). For fixed encoders, maps, and observations, order and partition do not change the final state in exact arithmetic.

Proof. Feature sums, second order statistics, products of features and labels, and class counts are sums over observations. Addition is associative and commutative. Thus, these statistics do not depend on the order or task-local partition.

For the variance update, we merge each class's count, feature sum, and elementwise squared feature sum across blocks. We compute the class variance after this merge. Each class therefore contributes the same variance under every partition. Hence, Γ^t is unchanged.

With fixed hyperparameters, μ_x^t, μ_z^t, U^t, A^t, and W^t are deterministic functions of this state. Their values are also unchanged.

A.2 Storage Complexity

The persistent state excludes the fixed encoders and feature expansion parameters. It stores three $d_x \times d_x$ matrices, two $d_x \times K^t$ matrices, three vectors of length d_x, and one vector of length K^t. Thus, its storage is

$$\begin{aligned} &O\big(3d_x^2 + 2d_x K^t + 3d_x + K^t\big) \\ &\quad = O\big(d_x^2 + K^t d_x\big)\,. \end{aligned} \tag{A.9}$$

Storage is independent of N_t. We discard temporary class statistics and sample features after each task. The means, Var^t, U^t, A^t, and W^t can be reconstructed from the stored state.

B Detailed Method Formulation

This appendix provides supporting derivations for the method. It defines the fixed feature maps. It derives the SVD form of Visual Uncertainty-Calibrated Alignment Module and the closed form of Label-Guided Decision Module.

B.1 Random Feature Expansion

The main text defines the expanded features as $X_t = g_v(V_t)$ and $Z_t = g_t(S_t)$. Here, we specify the fixed projections used by these mappings. We draw each matrix entry independently from a zero mean Gaussian distribution:

$$W_{\mathrm{RA}}^v[a,b] \sim \mathcal{N}(0, \sigma_{\mathrm{RA}}^2), \qquad W_{\mathrm{RA}}^{\mathrm{txt}}[a,b] \sim \mathcal{N}(0, \sigma_{\mathrm{RA}}^2). \tag{B.1}$$

Here, $W_{\mathrm{RA}}^v \in \mathbb{R}^{d_v \times d_x}$ and $W_{\mathrm{RA}}^{\mathrm{txt}} \in \mathbb{R}^{d_s \times d_x}$. The parameter σ_{RA} sets the fixed projection scale. The superscripts identify the visual and text modalities.

The mappings in the main text are

$$X_t = \mathrm{ReLU}(V_t W_{\mathrm{RA}}^v), \qquad Z_t = \mathrm{ReLU}(S_t W_{\mathrm{RA}}^{\mathrm{txt}}). \tag{B.2}$$

We sample each matrix once with a fixed seed and keep it unchanged across tasks. ReLU acts elementwise on each row. Thus, every task uses the same feature space. These mappings add no trainable parameters.

B.2 VISUAL UNCERTAINTY-CALIBRATED ALIGNMENT MODULE

The main text defines the centered scatter matrices and the mapping A^t. For this derivation, define the regularized metrics

$$\begin{aligned} G_x^t &= \frac{1}{N_t - 1} \sum\nolimits_{\widetilde{x}\widetilde{x}}^{t} + \lambda U^t, \\ G_z^t &= \frac{1}{N_t - 1} \sum\nolimits_{\widetilde{z}\widetilde{z}}^{t} + \lambda I. \end{aligned} \tag{B.3}$$

Both metrics are positive definite because $\lambda > 0$ and U^t has positive diagonal entries. The mapping becomes

$$A^t = (G_x^t)^{-1} \left(\frac{1}{N_t - 1} \sum\nolimits_{\widetilde{x}\widetilde{z}}^{t} \right) (G_z^t)^{-1}. \tag{B.4}$$

This mapping has a regularized canonical correlation interpretation (Hotelling, 1992; Hardoon et al., 2004). Let $p, q \in \mathbb{R}^{d_x}$ be directions in the expanded visual and text feature spaces. We first maximize the Pearson correlation between their projected sample values. When both projected variances are nonzero, it is

$$\mathrm{Corr}_t(p, q) = \frac{p^\top \sum_{\widetilde{x}\widetilde{z}}^{t} q}{\sqrt{p^\top \sum_{\widetilde{x}\widetilde{x}}^{t} p} \sqrt{q^\top \sum_{\widetilde{z}\widetilde{z}}^{t} q}}.$$

The common covariance factor $1/(N_t - 1)$ cancels in this ratio. We then replace the two marginal covariances with G_x^t and G_z^t. This includes visual uncertainty and stabilizes both metrics. The regularized correlation objective is

$$\frac{p^\top \left(\frac{1}{N_t - 1} \sum_{\widetilde{x}\widetilde{z}}^{t} \right) q}{\sqrt{p^\top G_x^t p} \sqrt{q^\top G_z^t q}}.$$

Positive rescaling of either direction does not change this ratio. We therefore impose unit metric norms and solve the bilinear problem

$$\max_{p,q} \quad p^\top \left(\frac{1}{N_t - 1} \sum\nolimits_{\widetilde{x}\widetilde{z}}^{t} \right) q, \qquad p^\top G_x^t p = q^\top G_z^t q = 1. \tag{B.5}$$

Canonical correlation analysis whitens each modality and seeks paired directions with maximal correlation. A common solution obtains these directions by singular value decomposition of the whitened relation. Our method instead computes A^t directly in the expanded feature coordinates and maps text into the visual space. Its visual metric includes the class balanced uncertainty U^t, which reflects within class variation. As we show below, retaining every nonzero singular component and mapping it back gives exactly A^t. Thus, the direct computation captures the complete regularized relation without explicitly computing the singular vectors. Section B.2.2 evaluates the resulting runtime benefit.

Set $p_{\mathrm{w}} = (G_x^t)^{1/2} p$ and $q_{\mathrm{w}} = (G_z^t)^{1/2} q$. These are directions in the whitened visual and text spaces. Conversely, $p = (G_x^t)^{-1/2} p_{\mathrm{w}}$ and $q = (G_z^t)^{-1/2} q_{\mathrm{w}}$ recover directions in the expanded feature spaces. The constraints then become

$$\|p_{\mathrm{w}}\|_2 = \|q_{\mathrm{w}}\|_2 = 1. \tag{B.6}$$

The whitened objective uses the matrix

$$\Omega^t = (G_x^t)^{-1/2} \left(\frac{1}{N_t - 1} \sum\nolimits_{\widetilde{x}\widetilde{z}}^{t} \right) (G_z^t)^{-1/2}. \tag{B.7}$$

Therefore, the problem reduces to maximizing $p_{\mathrm{w}}^\top \Omega^t q_{\mathrm{w}}$ over unit vectors.

We solve this problem with the compact singular value decomposition of Ω^t:

$$\Omega^t = L_{\mathrm{svd}} \Lambda_{\mathrm{svd}} R_{\mathrm{svd}}^\top, \tag{B.8}$$

where $\Lambda_{\mathrm{svd}} = \mathrm{diag}(\omega_1, \ldots, \omega_{r_\Omega})$ and $r_\Omega = \mathrm{rank}(\Omega^t)$. The columns of $L_{\mathrm{svd}}, R_{\mathrm{svd}} \in \mathbb{R}^{d_x \times r_\Omega}$ are orthonormal. The singular values are ordered from largest to smallest. If $r_\Omega = 0$, then $A^t = 0$ and the reconstruction sum is empty. We assume $r_\Omega > 0$ when discussing individual singular vectors.

The SVD also solves the whitened optimization problem. For each k, let $\ell_{\mathrm{svd},k}$ and $r_{\mathrm{svd},k}$ denote the left and right singular vectors. They satisfy

$$\Omega^t r_{\mathrm{svd},k} = \omega_k \ell_{\mathrm{svd},k}, \qquad (\Omega^t)^\top \ell_{\mathrm{svd},k} = \omega_k r_{\mathrm{svd},k}. \tag{B.9}$$

For unit vectors p_{w} and q_{w}, the SVD gives

$$p_{\mathrm{w}}^\top \Omega^t q_{\mathrm{w}} = p_{\mathrm{w}}^\top L_{\mathrm{svd}} \Lambda_{\mathrm{svd}} R_{\mathrm{svd}}^\top q_{\mathrm{w}}. \tag{B.10}$$

Set $a = L_{\mathrm{svd}}^\top p_{\mathrm{w}}$ and $b = R_{\mathrm{svd}}^\top q_{\mathrm{w}}$. Orthonormality gives $\|a\|_2 \le 1$ and $\|b\|_2 \le 1$. Therefore,

$$\begin{aligned} p_{\mathrm{w}}^\top \Omega^t q_{\mathrm{w}} &= a^\top \Lambda_{\mathrm{svd}} b \\ &\le \omega_1 \|a\|_2 \|b\|_2 \\ &\le \omega_1. \end{aligned} \tag{B.11}$$

The leading singular vectors attain this bound. For each singular pair, choose

$$p_{\mathrm{w},k} = \ell_{\mathrm{svd},k}, \qquad q_{\mathrm{w},k} = r_{\mathrm{svd},k}, \qquad a = b = \mathbf{e}_k,$$

where $\mathbf{e}_k$ is the kth standard basis vector. For $k = 1$, both a and b have unit norm, so the score reaches ω_1. For later pairs, we restrict each direction to the orthogonal complement of the preceding singular vectors. For directions collinear with a selected singular pair, unit length fixes their scale up to sign. We choose matching signs to obtain the positive score ω_k.

We map these whitened directions back to the expanded feature coordinates:

$$\begin{aligned} p_{\mathrm{cca},k} &= (G_x^t)^{-1/2} p_{\mathrm{w},k} = (G_x^t)^{-1/2} \ell_{\mathrm{svd},k}, \\ q_{\mathrm{cca},k} &= (G_z^t)^{-1/2} q_{\mathrm{w},k} = (G_z^t)^{-1/2} r_{\mathrm{svd},k}. \end{aligned} \tag{B.12}$$

Substituting these directions verifies the regularized unit constraints:

$$\begin{aligned} p_{\mathrm{cca},k}^\top G_x^t p_{\mathrm{cca},k} &= \ell_{\mathrm{svd},k}^\top (G_x^t)^{-1/2} G_x^t (G_x^t)^{-1/2} \ell_{\mathrm{svd},k} \\ &= \ell_{\mathrm{svd},k}^\top \ell_{\mathrm{svd},k} = 1, \\ q_{\mathrm{cca},k}^\top G_z^t q_{\mathrm{cca},k} &= r_{\mathrm{svd},k}^\top (G_z^t)^{-1/2} G_z^t (G_z^t)^{-1/2} r_{\mathrm{svd},k} \\ &= r_{\mathrm{svd},k}^\top r_{\mathrm{svd},k} = 1. \end{aligned} \tag{B.13}$$

Their bilinear score equals the corresponding singular value:

$$\begin{aligned} p_{\mathrm{cca},k}^\top \left(\frac{1}{N_t - 1} \sum\nolimits_{\widetilde{x}\widetilde{z}}^t \right) q_{\mathrm{cca},k} &= \ell_{\mathrm{svd},k}^\top \Omega^t r_{\mathrm{svd},k} \\ &= \omega_k. \end{aligned} \tag{B.14}$$

We now reconstruct the full mapping from all nonzero singular components. The inverse square roots are symmetric because G_x^t and G_z^t are positive definite.

$$\begin{aligned} A_{\mathrm{SVD}}^t &= \sum\nolimits_{k=1}^{r_\Omega} \omega_k p_{\mathrm{cca},k} q_{\mathrm{cca},k}^\top \\ &= \sum\nolimits_{k=1}^{r_\Omega} \omega_k \big((G_x^t)^{-1/2} p_{\mathrm{w},k} \big) \big((G_z^t)^{-1/2} q_{\mathrm{w},k} \big)^\top \\ &= \sum\nolimits_{k=1}^{r_\Omega} \omega_k \big((G_x^t)^{-1/2} \ell_{\mathrm{svd},k} \big) \big((G_z^t)^{-1/2} r_{\mathrm{svd},k} \big)^\top \\ &= (G_x^t)^{-1/2} \left(\sum\nolimits_{k=1}^{r_\Omega} \omega_k \ell_{\mathrm{svd},k} r_{\mathrm{svd},k}^\top \right) (G_z^t)^{-1/2} \\ &= (G_x^t)^{-1/2} L_{\mathrm{svd}} \Lambda_{\mathrm{svd}} R_{\mathrm{svd}}^\top (G_z^t)^{-1/2} \\ &= (G_x^t)^{-1/2} \Omega^t (G_z^t)^{-1/2} \\ &= (G_x^t)^{-1/2} \left[(G_x^t)^{-1/2} \left(\frac{1}{N_t - 1} \sum\nolimits_{\widetilde{x}\widetilde{z}}^t \right) (G_z^t)^{-1/2} \right] (G_z^t)^{-1/2} \\ &= (G_x^t)^{-1} \left(\frac{1}{N_t - 1} \sum\nolimits_{\widetilde{x}\widetilde{z}}^t \right) (G_z^t)^{-1} \\ &= A^t. \end{aligned} \tag{B.15}$$

The first substitutions replace each canonical direction with its whitened counterpart, then with its singular vector. We factor out the inverse square roots, using the symmetry of $(G_z^t)^{-1/2}$ when transposing the text direction. The compact SVD reconstructs Ω^t. Substituting its definition and combining the inverse square roots gives the mapping A^t from the main text. Thus, the direct mapping equals the full SVD reconstruction in exact arithmetic. It does not require the SVD factors.

B.2.1 Numerical Agreement between Full-SVD and No-SVD

The algebra above shows that full SVD reconstruction and direct No-SVD are equal in exact arithmetic. We next compare their numerical results.

Both variants use identical OpenCLIP ViT-B/16 features and RA dimension 5120. They use the same centered scatter matrices. They differ only in the SVD step.

Table B.1: Classification results of Full-SVD and No-SVD Visual Uncertainty-Calibrated Alignment Module.

Dataset	No-SVD Visual Uncertainty-Calibrated Alignment Module	Full-SVD Visual Uncertainty-Calibrated Alignment Module	Δ
Skin8-T4	61.74	61.74	0.00
Skin40-T20	69.75	69.75	0.00
Path16-T7	71.62	71.62	0.00
MedMNIST-T4	61.50	61.51	+0.01
COVID-T6	90.06	90.06	0.00
Blood-T4	91.85	91.85	0.00
Mean	74.42	74.42	+0.0017

At displayed precision, MCR agrees on five datasets. MedMNIST-T4 differs by 0.01 percentage points. The mean difference is 0.0017 percentage points. These results show nearly identical Visual Uncertainty-Calibrated Alignment Module performance across the six datasets.

B.2.2 Runtime Efficiency

After the accuracy comparison, we measured runtime on the same six datasets. Each run used cached features, OpenCLIP ViT-B/16, RA dimension 5120, and one GPU. Each configuration had two warm up runs and five timed runs.

Total time covers cache loading, RA projection, statistic accumulation, matrix construction, and final mapping. Full-SVD also includes the SVD and full reconstruction.

Table B.2: Total runtime of Full-SVD and No-SVD Visual Uncertainty-Calibrated Alignment Module.

Dataset	No-SVD	Full-SVD	Speedup
Skin8-T4	0.421 ± 0.000 s	82.065 ± 0.169 s	$195.1\times$
Skin40-T20	0.286 ± 0.000 s	81.082 ± 0.710 s	$283.1\times$
Path16-T7	0.580 ± 0.000 s	74.100 ± 0.788 s	$127.9\times$
MedMNIST-T4	10.552 ± 0.000 s	94.553 ± 1.992 s	$9.0\times$
COVID-T6	0.300 ± 0.000 s	78.039 ± 1.319 s	$260.3\times$
Blood-T4	1.135 ± 0.000 s	77.635 ± 0.128 s	$68.4\times$
Mean	2.212 s	81.246 s	$36.7\times$

Here, $\pm$ reports the sample standard deviation. The speedup is the ratio of mean runtimes across the six datasets. Direct No-SVD lowers mean runtime from 81.246 s to 2.212 s. This gives a $36.7\times$ speedup.

B.3 Label-Guided Decision Module

The main text defines the Label-Guided Decision Module objective and its closed form. We derive that solution here. In the derivation, W is the optimization variable and W^t is its stage t solution. For the derivation, let $\widetilde{X}^t, \widetilde{Z}^t \in \mathbb{R}^{N_t \times d_x}$ denote the centered visual and text feature arrays. Let $Y^t \in \mathbb{R}^{N_t \times K^t}$ denote the corresponding one hot labels. These arrays are algebraic notation. We compute their products from the accumulated statistics. We do not retain historical sample features.

The mapped text features satisfy

$$\widetilde{Z}_v^t = \widetilde{Z}^t (A^t)^\top. \tag{B.16}$$

Starting from the objective in the main text, the gradient condition is

$$\frac{2}{2N_t-1}\left[(\widetilde{X}^t)^\top(\widetilde{X}^tW-Y^t)+\alpha(\widetilde{Z}_v^t)^\top(\widetilde{Z}_v^tW-Y^t)\right]+2\beta W=0. \tag{B.17}$$

Rearranging gives

$$\left[(\widetilde{X}^t)^\top\widetilde{X}^t+\alpha(\widetilde{Z}_v^t)^\top\widetilde{Z}_v^t+\beta(2N_t-1)I\right]W=(\widetilde{X}^t)^\top Y^t+\alpha(\widetilde{Z}_v^t)^\top Y^t. \tag{B.18}$$

Substituting $\widetilde{Z}_v^t=\widetilde{Z}^t(A^t)^\top$ yields

$$(\widetilde{Z}_v^t)^\top\widetilde{Z}_v^t=A^t((\widetilde{Z}^t)^\top\widetilde{Z}^t)(A^t)^\top,\qquad(\widetilde{Z}_v^t)^\top Y^t=A^t((\widetilde{Z}^t)^\top Y^t). \tag{B.19}$$

Writing these products with the centered statistics from the main text gives the closed form

$$\begin{aligned}W^t=&\left[\sum\nolimits_{\widetilde{x}\widetilde{x}}^t+\alpha A^t\sum\nolimits_{\widetilde{z}\widetilde{z}}^t(A^t)^\top+\beta(2N_t-1)I\right]^{-1}\\&\times\left[\sum\nolimits_{\widetilde{x}y}^t+\alpha A^t\sum\nolimits_{\widetilde{z}y}^t\right].\end{aligned} \tag{B.20}$$

This matches the solution in the main text.

C Theoretical Limits of Across Modalities Alignment for Classification

We construct data distributions with identical joint distributions across modalities but different classification performance. This shows that perfect representation alignment does not determine downstream performance.

Theorem C.1 (Underdetermination of classification under perfect alignment). For every $K\geq 2$, we construct balanced distributions with K classes. They share the same joint distribution P_{XZ} across modalities. Their representations are perfectly aligned. Bayes optimal MCR can take any value in

$$\left[\frac{1}{K},1\right] \tag{C.1}$$

throughout this family.

Proof. Let

$$\xi\sim\mathcal{N}(0,I_{d_x}), \tag{C.2}$$

and define

$$X=\xi,\qquad Z=\xi. \tag{C.3}$$

Because X and Z are zero mean, define their within-modality and cross-modal second-order matrices as $C_{xx}=\mathbb{E}[XX^\top]$, $C_{zz}=\mathbb{E}[ZZ^\top]$, and $C_{xz}=\mathbb{E}[XZ^\top]$. Here, I_{d_x} denotes the d_x-dimensional identity matrix. The two modalities then have identical representations satisfying

$$C_{xx}=C_{zz}=C_{xz}=I_{d_x}. \tag{C.4}$$

All canonical correlations across modalities are 1. Thus, the alignment is perfect.

We keep this joint distribution fixed and vary the relation between labels and representations. Let $F_{\mathcal{N}}$ be the standard normal cumulative distribution function. Set the endpoints and interior quantiles as

$$\vartheta_0=-\infty,\qquad\vartheta_K=+\infty,\qquad\vartheta_c=F_{\mathcal{N}}^{-1}\left(\frac{c}{K}\right),\quad c=1,\dots,K-1. \tag{C.5}$$

Define the class partition function

$$\kappa(\xi)=c,\qquad\vartheta_{c-1}<\xi_1\leq\vartheta_c. \tag{C.6}$$

Since

$$\xi_1 \sim \mathcal{N}(0,1), \tag{C.7}$$

we have

$$\Pr(\kappa(\xi) = c) = \frac{1}{K}, \qquad c = 1, \ldots, K, \tag{C.8}$$

so this partition is balanced across classes.

For any $\gamma \in [0,1]$, let

$$\zeta \sim \text{Bernoulli}(\gamma), \qquad Y_{\text{unif}} \sim \text{Uniform}\{1, \ldots, K\}, \tag{C.9}$$

where ζ, Y_{unif}, and ξ are mutually independent, and define

$$Y_{\text{cls}} = \begin{cases} \kappa(\xi), & \zeta = 1, \\ Y_{\text{unif}}, & \zeta = 0. \end{cases} \tag{C.10}$$

The parameter γ changes only the relation between labels and representations. For every γ,

$$X = Z = \xi. \tag{C.11}$$

The family therefore shares the same joint distribution P_{XZ} and the same alignment structure across modalities.

Given visual representation $X = x$, the posterior probability of class c is

$$\Pr(Y_{\text{cls}} = c \mid X = x) = \gamma \mathbf{1}\{\kappa(x) = c\} + \frac{1-\gamma}{K}. \tag{C.12}$$

When $\gamma > 0$, $\kappa(x)$ has the highest posterior probability. Thus, the Bayes optimal classifier is

$$h^\star(x) = \kappa(x). \tag{C.13}$$

For any class c, we also have

$$\begin{aligned} \Pr(Y_{\text{cls}} = c) &= \gamma \Pr(\kappa(\xi) = c) + (1-\gamma) \Pr(Y_{\text{unif}} = c) \\ &= \gamma \frac{1}{K} + (1-\gamma) \frac{1}{K} \\ &= \frac{1}{K}. \end{aligned} \tag{C.14}$$

All classes retain equal prior probability. Therefore, overall accuracy equals MCR. The Bayes optimal MCR is

$$\text{MCR}^\star(\gamma) = \gamma + \frac{1-\gamma}{K}. \tag{C.15}$$

When $\gamma = 0$, labels and representations are independent. Thus,

$$\text{MCR}^\star = \frac{1}{K}. \tag{C.16}$$

When $\gamma = 1$, the representation determines each label. Thus,

$$\text{MCR}^\star = 1. \tag{C.17}$$

More generally, for any

$$\nu \in \left[\frac{1}{K}, 1\right], \tag{C.18}$$

choosing

$$\gamma = \frac{K\nu - 1}{K - 1}, \tag{C.19}$$

gives

$$\mathrm{MCR}^{\star}(\gamma) = \nu. \tag{C.20}$$

Thus, under identical and perfect alignment across modalities, Bayes optimal classification performance can take any value in

$$\left[\frac{1}{K}, 1\right] \tag{C.21}$$

This completes the proof.

Alignment across modalities describes statistical relations between visual and text representations. Classification also needs structure conditioned on labels. Therefore, ideal representation alignment alone cannot guarantee classification performance. Appendix D shows how Label-Guided Decision Module uses label statistics to form discriminative structure.

D Theoretical Analysis of Label-Guided Decision Module

Appendix C shows that feature relations alone do not determine classification performance, even under perfect alignment across modalities. Here, we start from the closed form Label-Guided Decision Module solution. We analyze how statistics conditioned on labels form class specific directions. We then show how joint second order statistics adjust these directions and affect the decision boundaries. Appendix B.3 gives the objective and its derivation.

D.1 Pairwise Discriminative Directions from Labels

Write the Label-Guided Decision Module solution as

$$W^t = (G_\beta^t)^{-1} R^t, \tag{D.1}$$

where

$$G_\beta^t = \frac{\sum_{\widetilde{x}\widetilde{x}}^t + \alpha \sum_{\widetilde{z}\widetilde{z}}^t}{2N_t - 1} + \beta I, \qquad R^t = \frac{\sum_{\widetilde{x}y}^t + \alpha \sum_{\widetilde{z}y}^t}{2N_t - 1}. \tag{D.2}$$

Here $\widetilde{X}^t, \widetilde{Z}_v^t \in \mathbb{R}^{N_t \times d_x}$ are centered visual and mapped textual features, and $Y^t \in \mathbb{R}^{N_t \times K^t}$ is the cumulative one-hot label matrix. For each observed class c, let $\mathcal{I}_c^t = \{i : y_i = c\}$ and $n_c^t = |\mathcal{I}_c^t|$.

Let

$$W^t = [w_1^t, \ldots, w_{K^t}^t], \qquad R^t = [r_1^t, \ldots, r_{K^t}^t], \tag{D.3}$$

Then, for any class c,

$$r_c^t = \frac{(\widetilde{X}^t)^\top (Y^t)_{:,c} + \alpha (\widetilde{Z}_v^t)^\top (Y^t)_{:,c}}{2N_t - 1}, \qquad w_c^t = (G_\beta^t)^{-1} r_c^t. \tag{D.4}$$

For class pair (c, j), let $\mathbf{e}_c$ and $\mathbf{e}_j$ be the corresponding standard basis vectors, and define the label contrast vector

$$d_{cj}^t = Y^t(\mathbf{e}_c - \mathbf{e}_j). \tag{D.5}$$

Then

$$r_{cj}^t = r_c^t - r_j^t = \frac{(\widetilde{X}^t)^\top d_{cj}^t + \alpha (\widetilde{Z}_v^t)^\top d_{cj}^t}{2N_t - 1}, \tag{D.6}$$

and

$$w_c^t - w_j^t = (G_\beta^t)^{-1} r_{cj}^t. \tag{D.7}$$

The vector d_{cj}^t takes value $+1$ on samples from class c, -1 on samples from class j, and 0 otherwise. Denote the class c means of the accumulated centered visual and mapped textual representations by the column vectors

$$\widetilde{\mu}_{x,c}^t = \frac{1}{n_c^t} \sum\nolimits_{i \in \mathcal{I}_c^t} (\widetilde{X}^t)_{i,:}^\top, \qquad \widetilde{\mu}_{z_v,c}^t = \frac{1}{n_c^t} \sum\nolimits_{i \in \mathcal{I}_c^t} (\widetilde{Z}_v^t)_{i,:}^\top, \tag{D.8}$$

The tilde marks representations centered before alignment. The subscripts x and z_v identify visual features and mapped text features. The index c denotes a class.

Then

$$r_{cj}^t = \frac{n_c^t(\widetilde{\mu}_{x,c}^t + \alpha\widetilde{\mu}_{z_v,c}^t) - n_j^t(\widetilde{\mu}_{x,j}^t + \alpha\widetilde{\mu}_{z_v,j}^t)}{2N_t - 1}. \tag{D.9}$$

In particular, if $n_c^t = n_j^t$,

$$r_{cj}^t \propto (\widetilde{\mu}_{x,c}^t - \widetilde{\mu}_{x,j}^t) + \alpha(\widetilde{\mu}_{z_v,c}^t - \widetilde{\mu}_{z_v,j}^t). \tag{D.10}$$

Thus, r_{cj}^t encodes the feature contrast conditioned on labels for classes c and j. Their decision weight difference is

$$w_c^t - w_j^t = (G_\beta^t)^{-1} r_{cj}^t \tag{D.11}$$

Unlike the feature relations modeled by Visual Uncertainty-Calibrated Alignment Module, these class directions depend on statistics conditioned on labels. The matrix $(G_\beta^t)^{-1}$ adjusts their contributions using the joint second order structure of both modalities.

D.2 Covariance Scale Adjustment from Joint Statistics

To analyze the effect of $(G_\beta^t)^{-1}$ on class contrasts, define

$$G^t = \frac{(\widetilde{X}^t)^\top \widetilde{X}^t + \alpha(\widetilde{Z}_v^t)^\top \widetilde{Z}_v^t}{2N_t - 1}, \qquad G_\beta^t = G^t + \beta I. \tag{D.12}$$

Here, G^t combines feature relations from visual and mapped text representations. Its eigendecomposition is

$$G^t = Q_{\text{eig}}^t \operatorname{diag}(\eta_1^t, \ldots, \eta_{d_x}^t)(Q_{\text{eig}}^t)^\top, \tag{D.13}$$

The k-th column $q_{\text{eig},k}^t$ is a statistical direction. The value η_k^t is its second order scale.

Expand the class contrast r_{cj}^t in the same eigenbasis:

$$r_{cj}^t = \sum\nolimits_{k=1}^{d_x} a_{cj,k}^t q_{\text{eig},k}^t. \tag{D.14}$$

Since

$$(G_\beta^t)^{-1} = Q_{\text{eig}}^t \operatorname{diag}\left(\frac{1}{\eta_1^t + \beta}, \ldots, \frac{1}{\eta_{d_x}^t + \beta}\right)(Q_{\text{eig}}^t)^\top, \tag{D.15}$$

we obtain

$$w_c^t - w_j^t = (G_\beta^t)^{-1} r_{cj}^t = \sum\nolimits_{k=1}^{d_x} \frac{a_{cj,k}^t}{\eta_k^t + \beta} q_{\text{eig},k}^t. \tag{D.16}$$

Here, $a_{cj,k}^t$ is the class contrast conditioned on labels along direction k. The factor $(\eta_k^t + \beta)^{-1}$ adjusts this contrast by its joint second order scale. Feature and label statistics supply class specific information. Feature relations adjust it by covariance scale.

Thus, Label-Guided Decision Module class directions depend on both label conditioned contrasts and current second order statistics across modalities.

D.3 Connection to the Final Decision Boundary

Let w_c^t denote the class c column of W^t used for scoring.

$$s_c(x) = x^\top w_c^t, \qquad s_c(x) - s_j(x) = x^\top \Delta_{cj}^t, \qquad \Delta_{cj}^t = w_c^t - w_j^t. \tag{D.17}$$

Here x is a centered visual feature used for evaluation. The pairwise decision boundary is $x^\top \Delta_{cj}^t = 0$. The signed margin is

$$g_{cj}(x) = \frac{x^\top \Delta_{cj}^t}{\max\{\|\Delta_{cj}^t\|_2, \varepsilon\}}, \qquad \varepsilon = 10^{-12}. \tag{D.18}$$

For the evaluation set class centers μ_c and μ_j defined in Appendix E, their projected separation is

$$g_{cj}(\mu_c) - g_{cj}(\mu_j) = \frac{(\mu_c - \mu_j)^\top \Delta_{cj}^t}{\max\{\|\Delta_{cj}^t\|_2, \varepsilon\}}. \tag{D.19}$$

When $\|\Delta_{cj}^t\|_2 \geq \varepsilon$ and both vectors are nonzero, let θ_{cj} be the angle between $\mu_c - \mu_j$ and Δ_{cj}^t. Then the expression equals $\|\mu_c - \mu_j\|_2 \cos\theta_{cj}$. Thus, separation depends on the relation between class centers and the actual classifier direction. The conditioned on labels information and covariance calibration in Appendices D.1 and D.2 affect that direction through the fitted columns w_c^t. Appendix F examines the corresponding changes in class separation, boundary distances, and sample level margins.

E Decision Geometry Metrics

For an evaluation set $\{(x_i, y_i)\}_{i=1}^N$ with K classes, let $\mathcal{I}_c = \{i : y_i = c\}$, $n_c = |\mathcal{I}_c|$, and $\mu_c = n_c^{-1} \sum_{i \in \mathcal{I}_c} x_i$. The bias-free linear classifier assigns score $s_c(x) = x^\top w_c$. For two distinct classes, define

$$\Delta_{cj} = w_c - w_j, \qquad g_{cj}(x) = \frac{x^\top \Delta_{cj}}{\max\{\|\Delta_{cj}\|_2, \varepsilon\}}, \quad \varepsilon = 10^{-12}. \tag{E.1}$$

Here, $g_{cj}(x)$ is a numerically stabilized signed margin: when $\|\Delta_{cj}\|_2 \geq \varepsilon$, it equals the signed distance to the pairwise hyperplane $s_c(x) = s_j(x)$, with positive values favoring class c.

At the class level, we measure *projected center separation* M, averaged over unordered class pairs, and *center safety margin* C, averaged over classes:

$$M = \frac{2}{K(K-1)} \sum_{c<j} \big[g_{cj}(\mu_c) - g_{cj}(\mu_j)\big], \qquad C = \frac{1}{K} \sum_{c=1}^{K} \min_{j \neq c} g_{cj}(\mu_c). \tag{E.2}$$

Thus, M measures separation along classifier-defined directions, while C measures each center's smallest signed margin against a competing class.

At the sample level, let $b_i = \min_{j \neq y_i} g_{y_i j}(x_i)$. We report the *class-balanced sample margin* B and the *boundary violation rate* V:

$$B = \frac{1}{K} \sum_{c=1}^{K} \frac{1}{n_c} \sum_{i \in \mathcal{I}_c} b_i, \qquad V = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{b_i \leq 0\}. \tag{E.3}$$

The indicator $\mathbb{I}\{\cdot\}$ counts samples on or beyond at least one competing boundary, including score ties. Higher M, C, and B, and lower V, indicate improved decision geometry. Unlike MCR, V weights samples equally rather than balancing classes.

F Extended Analysis of Decision Geometry and Uncertainty

This appendix extends the main text in two ways. Sections F.1 and F.2 examine decision geometry across datasets and in representative samples. Sections F.3 and F.4 describe variation in each dimension across stages.

F.1 Decision Geometry Changes across Datasets

The main text shows that adding Label-Guided Decision Module to Visual Uncertainty-Calibrated Alignment Module increases mean class separation M, nearest competing boundary margin C, and sample margin B. It also reduces boundary violation rate V. Here, we report results for each dataset to test whether these changes are consistent.

We use the decision geometry metrics defined in Appendix E. For each of M, C, B, and V, its Δ is the value with Label-Guided Decision Module minus the value with Visual Uncertainty-Calibrated Alignment Module. For example,

$$\Delta M = M_{\text{Label-Guided Decision Module}} - M_{\text{Visual Uncertainty-Calibrated Alignment Module}}, \tag{F.1}$$

and similarly define

$$\Delta\mathrm{MCR} = \mathrm{MCR}_{\text{Label-Guided Decision Module}} - \mathrm{MCR}_{\text{Visual Uncertainty-Calibrated Alignment Module}}. \tag{F.2}$$

Table F.1: Changes in decision geometry after adding Label-Guided Decision Module on six datasets.

Dataset	ΔM	ΔC	ΔB	ΔV	ΔMCR
Skin8-T4	+6.382	+2.611	+1.311	-0.081	+6.09
Skin40-T20	+15.077	+5.843	+4.079	-0.095	+9.50
Path16-T7	+15.898	+4.528	+4.011	-0.017	+1.73
MedMNIST-T4	+30.223	+9.428	+7.345	-0.075	+11.83
COVID-T6	+19.507	+8.160	+7.164	-0.056	+4.11
Blood-T4	+17.902	+7.337	+6.222	-0.017	+2.10
Mean	+17.498	+6.318	+5.022	-0.057	+5.89

The $\Delta\mathrm{MCR}$ values are percentage points.

All six datasets satisfy

$$\Delta M > 0, \qquad \Delta C > 0, \qquad \Delta B > 0, \qquad \Delta V < 0, \tag{F.3}$$

Each dataset also shows a positive MCR gain. The average improvement is not driven by a few datasets. Decision geometry changes across class counts, stage counts, and medical image distributions.

Geometric improvements do not map directly to MCR gains. COVID-T6 and Blood-T4 have large ΔC and ΔB. Their MCR gains are $+4.11$ and $+2.10$. Skin8-T4 has smaller geometric changes, yet its violation rate falls by $\Delta V = -0.081$, and its MCR increases by $+6.09$ percentage points.

Decision geometry is continuous, while predictions are discrete. A larger margin can improve a correct sample without changing its prediction. A small change can alter the prediction of a sample near a competing boundary. Thus, M, C, B, and V complement MCR. They describe class separation, competing boundaries, and sample decisions.

F.2 Representative Decision Correction Cases

Appendix F.1 shows that Label-Guided Decision Module increases class separation and sample margins. It also reduces violations. Here, we examine three Skin8-T4 cases from the basal cell carcinoma (BCC) class. They show how geometry changes correct predictions.

The three cases show a consistent progression. Under zero shot prediction, the true BCC class ranks fifth in every case, with a response near zero. After Visual Uncertainty-Calibrated Alignment Module, it rises to rank two, with responses near 0.13.

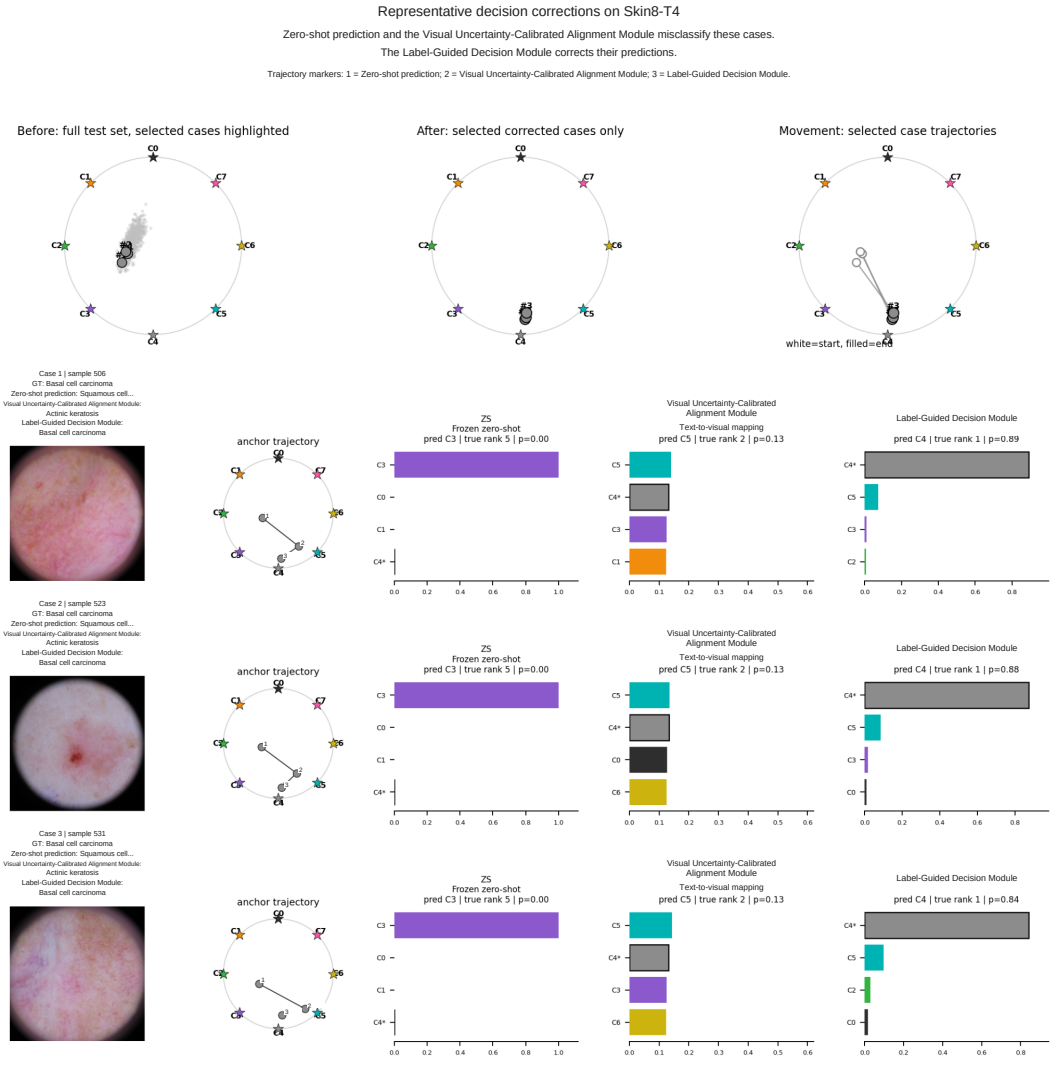

Figure F.1: Representative decision correction cases. Top: selected positions and decision paths in class anchor space. Bottom: medical images and class responses under zero shot prediction, Visual Uncertainty-Calibrated Alignment Module, and Label-Guided Decision Module.

Statistical modeling across modalities improves the relation between each sample and its true class. However, the main competitor still ranks first. After Label-Guided Decision Module, the true class ranks first in all three cases. Its responses are approximately 0.89, 0.88, and 0.84.

These cases show two stages of decision adjustment. Visual Uncertainty-Calibrated Alignment Module moves the true class into close competition. Label-Guided Decision Module then changes the class relations and moves each sample across the previous decision boundary.

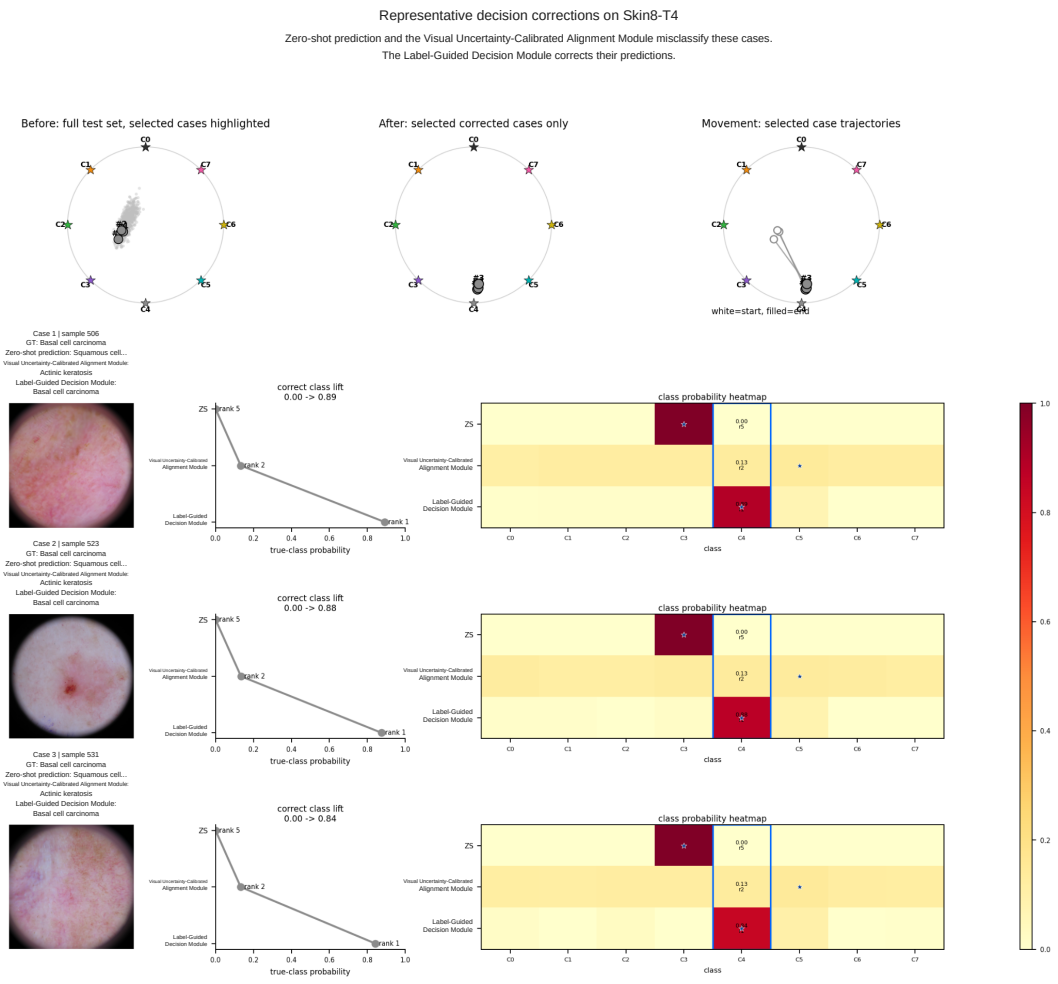

Figure F.2: Class responses in representative decision correction cases. Left: true class responses and ranks under zero shot prediction, Visual Uncertainty-Calibrated Alignment Module, and Label-Guided Decision Module. Right: full class response distributions.

The full response distributions show the same pattern. Visual Uncertainty-Calibrated Alignment Module raises the true class from a low rank into close competition, but does not change the prediction. Label-Guided Decision Module increases its advantage and moves it to rank 1.

These cases agree with Appendix F.1. Better decision geometry does not always change a discrete prediction immediately. Further adjustment can correct a sample near a competing boundary.

F.3 Temporal Variation across Dimensions

We track a diagnostic for each dimension as classes arrive. It uses frozen visual features $v_{\tau,i}$ with $d_v = 512$. This auxiliary diagnostic inspects the visual variance profile. The model matrix U^t is computed from expanded features defined in Section 3. We denote the balanced variance across classes at stage t by Var_v^t. For a dataset with final stage $t_\star > 1$, define its temporal mean and median scale by

$$\overline{\mathrm{Var}}_{v,j} = \frac{1}{t_\star} \sum\nolimits_{t=1}^{t_\star} \mathrm{Var}_{v,j}^t, \qquad m_{\mathrm{dyn}} = \mathrm{median}_{1 \le k \le d_v} \overline{\mathrm{Var}}_{v,k}. \tag{F.4}$$

The normalized score for temporal variation in dimension j is

$$\delta_j^{\mathrm{dyn}} = \frac{\dfrac{1}{t_\star - 1} \sum\nolimits_{t=2}^{t_\star} \left| \mathrm{Var}_{v,j}^t - \mathrm{Var}_{v,j}^{t-1} \right|}{\overline{\mathrm{Var}}_{v,j} + m_{\mathrm{dyn}} + \epsilon_{\mathrm{dyn}}}, \qquad \epsilon_{\mathrm{dyn}} = 10^{-12}. \tag{F.5}$$

Here, $t_\star$ is the number of stages. The numerator measures the mean absolute change between adjacent stages. The denominator uses the dimension's temporal mean and a dataset median scale. We report the Median, Mean, 90th percentile (Q90), and Max of δ_j^{dyn} across dimensions.

Table F.2: Distribution of normalized temporal variation scores δ_j^{dyn} across feature dimensions on six datasets.

Dataset	Median	Mean	Q_{90}	Max
Skin8-T4	0.0203	0.0232	0.0421	0.1135
Skin40-T20	0.0168	0.0180	0.0268	0.0523
Path16-T7	0.0399	0.0462	0.0767	0.3362
MedMNIST-T4	0.0151	0.0183	0.0330	0.1091
COVID-T6	0.0640	0.0710	0.1153	0.2942
Blood-T4	0.0177	0.0210	0.0386	0.0694
Mean	0.0290	0.0330	0.0554	0.1625

Across all six datasets, the Mean of δ_j^{dyn} exceeds its Median, indicating a right tail of dimensions with larger relative changes during class arrival. Q90 is approximately 1.60-2.19 times the Median, so between stages changes are not uniform across feature dimensions.

The normalized change differs among datasets. COVID-T6 has the highest Median, Mean, and Q90, whereas Skin40-T20 and MedMNIST-T4 have lower central scores. These values show that temporal variability differs among datasets.

The Max values identify dimensions with larger changes. Across datasets, Max is about 3.11 to 8.43 times the Median. Path16-T7 has Median 0.0399 and Max 0.3362. MedMNIST-T4 has Median 0.0151 and Max 0.1091. A modest dataset median does not mean every dimension remains equally stable.

These results show that the diagnostic can vary unevenly across dimensions as new classes arrive.

F.4 Uncertainty under Continual Class Arrival

Appendix F.3 summarizes dimension level variation across stages. Here, we examine the mean and distribution at each stage.

At incremental stage t, let $\mathcal{C}^{(1:t)}$ be the observed class set and $K^t = |\mathcal{C}^{(1:t)}|$. Using only frozen visual training features available through the current stage, we define the balanced variance across classes

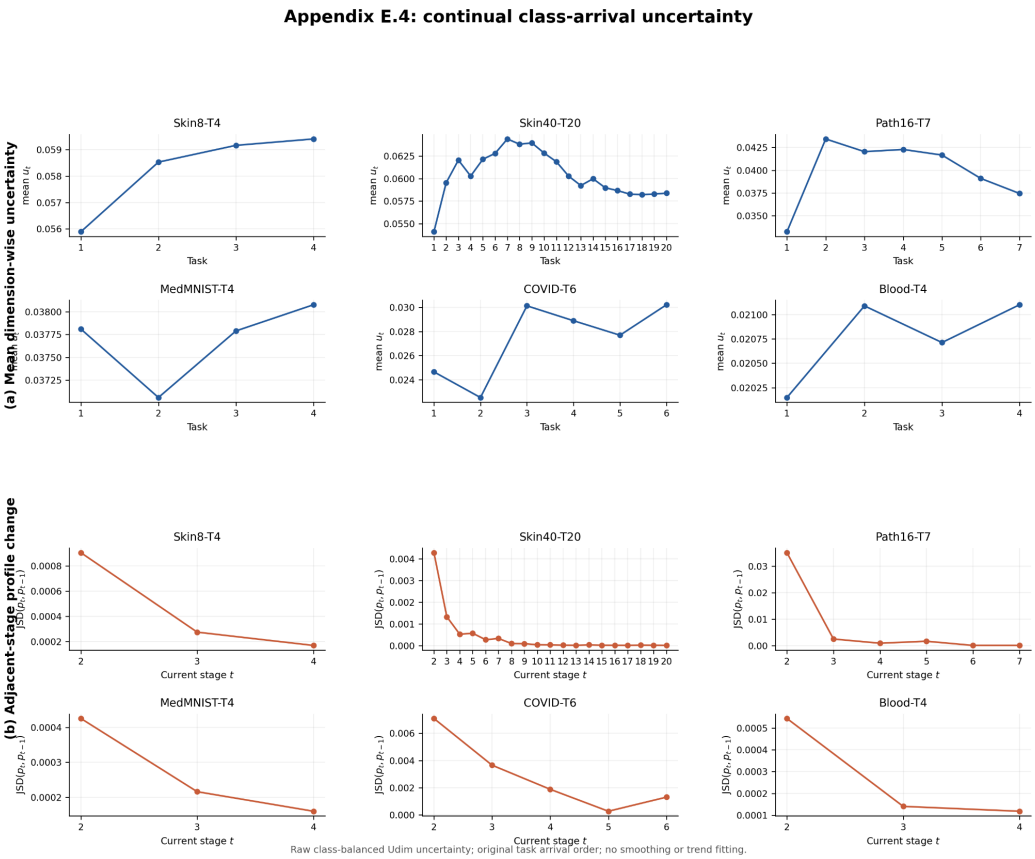

Figure F.3: Stage by stage diagnostic values on six datasets. Top: mean variance across dimensions, $\overline{\mathrm{Var}}_v^t$. Bottom: change between adjacent profiles, $J_t^{\mathrm{unc}} = \mathrm{JSD}(q_t^{\mathrm{unc}}, q_{t-1}^{\mathrm{unc}})$. Each stage uses available training data and the original class order.

for each dimension as

$$\mathrm{Var}_v^t = \frac{1}{K^t} \sum_{\tau=1}^{t} \sum_{c \in \mathcal{C}_\tau} \frac{1}{n_{\tau,c} - 1} \sum_{i \in \mathcal{I}_{\tau,c}} (v_{\tau,i} - \mu_{v,\tau,c}) \odot (v_{\tau,i} - \mu_{v,\tau,c}). \tag{F.6}$$

Here $\mu_{v,\tau,c} = n_{\tau,c}^{-1} \sum_{i \in \mathcal{I}_{\tau,c}} v_{\tau,i}$ is the frozen visual feature mean of class c in task τ. Each class has $n_{\tau,c} > 1$, and the squared difference is computed element by element.

To summarize the overall uncertainty level at the current stage, define

$$\overline{\mathrm{Var}}_v^t = \frac{1}{d_v} \sum_{j=1}^{d_v} \mathrm{Var}_{v,j}^t. \tag{F.7}$$

The mean does not capture changes in the dimension profile. We therefore define the normalized profile

$$q_t^{\mathrm{unc}} = \frac{\mathrm{Var}_v^t + \epsilon_{\mathrm{p}} \mathbf{1}}{\sum_{j=1}^{d_v} \mathrm{Var}_{v,j}^t + d_v \epsilon_{\mathrm{p}}}, \qquad \epsilon_{\mathrm{p}} = 10^{-12}, \tag{F.8}$$

and use the Jensen-Shannon divergence between adjacent stages,

$$J_t^{\mathrm{unc}} = \mathrm{JSD}(q_t^{\mathrm{unc}}, q_{t-1}^{\mathrm{unc}}), \qquad t \geq 2, \tag{F.9}$$

to measure changes in the profile after new classes arrive. We use natural logarithms in the Jensen-Shannon divergence.

The mean $\overline{\mathrm{Var}}_v^t$ varies across stages on all six datasets, with no common monotonic pattern. Skin8-T4 increases as classes arrive. Skin40-T20 changes rapidly early, then declines and stabilizes. Path16-T7

rises early and later declines. MedMNIST-T4, COVID-T6, and Blood-T4 alternate between increases and decreases.

New classes change the contributions to the balanced class aggregation along each dimension. Thus, the profile can change as the observed class set grows.

Adjacent stage J_t^{unc} shows changes in the distribution across dimensions. All six datasets have nonzero profile changes as classes arrive. Larger changes often occur early and weaken later. This trend is not monotonic. Path16-T7 rises at an intermediate stage. COVID-T6 changes again late in the sequence.

The quantities $\overline{\mathrm{Var}}_v^t$ and J_t^{unc} capture different changes. Some adjacent stages have similar means but nonzero J_t^{unc}. Their profiles differ even when the overall variance remains similar.

Together with the temporal scores in Appendix F.3, these curves show changes in both the mean level and the distribution across dimensions.

G Additional Experimental Analysis

G.1 Dataset and Continual Learning Protocols

We summarize the datasets and incremental protocols used in the main experiments. New classes arrive across tasks. We evaluate all observed classes after each stage.

Table G.1: Medical continual learning protocols used in the main text.

Dataset	Setting
Skin8	$T4$
Skin40	$T20$
Path16	$T7$
MedMNIST	$T4$
COVID	$T6$
Blood	$T4$

We examine how Random Anchors settings, statistics implementations, and training set size affect performance. Unless stated otherwise, experiments use six datasets, frozen OpenCLIP ViT-B/16 features, and the main text method. The ablation names follow the main table. ‘Cross-modal alignment‘ uses the alignment stage. ‘+ Uncertainty calibration‘ adds U^t. ‘+ Random Anchor‘ adds the random feature expansion. ‘+ Label-guided decisions‘ is the full setting and adds Label-Guided Decision Module. Results report final MCR using images only across observed classes.

G.2 Robustness to Random Anchors Dimension and Initialization

RA maps frozen visual and text representations into a fixed feature space. We vary the projection dimension and random seed separately. This tests whether performance depends on one RA setting.

The expanded feature dimension is d_x. We vary it over

$$d_x \in \{512, 1024, 2560, 5120\}. \tag{G.1}$$

Table G.2 reports final MCR on six datasets, with each cell showing Visual Uncertainty-Calibrated Alignment Module/ Full.

As d_x increases, mean Visual Uncertainty-Calibrated Alignment Module MCR rises from 65.18% to 74.42%. Full rises from 74.01% to 80.31%. Larger RA spaces generally provide more capacity. Gains vary by dataset. Skin40-T20 and COVID-T6 improve at higher dimensions. Full on Path16-T7 is nearly stable above $d_x = 2560$.

Visual Uncertainty-Calibrated Alignment Module does not increase at every lower dimension. Path16-T7 and MedMNIST-T4 decline at 1024 dimensions. Full improves on all six datasets across the tested range. Label decisions show a steadier trend. The 5120 setting gives the highest mean across six datasets.

Table G.2: Final MCR at different RA dimensions. Each cell reports Visual Uncertainty-Calibrated Alignment Module/ Full.

Dataset	512	1024	2560	5120
Skin8-T4	53.14 / 60.61	52.84 / 63.27	58.38 / 66.55	61.74 / 67.83
Skin40-T20	54.00 / 69.00	59.25 / 72.75	66.75 / 75.50	69.75 / 79.25
Path16-T7	68.07 / 70.19	59.12 / 72.20	68.49 / 73.12	71.62 / 73.35
MedMNIST-T4	56.75 / 66.33	52.26 / 68.94	57.74 / 71.55	61.50 / 73.33
COVID-T6	77.58 / 88.46	84.70 / 90.89	86.69 / 93.18	90.06 / 94.17
Blood-T4	81.54 / 89.44	89.14 / 91.66	91.22 / 93.00	91.85 / 93.95
Mean	65.18 / 74.01	66.22 / 76.62	71.55 / 78.82	74.42 / 80.31

We next fixed

$$d_x = 5120, \tag{G.2}$$

We used five random initializations. Table G.3 reports the mean, standard deviation, and range of Full MCR.

Table G.3: Full MCR across RA random seeds.

Dataset	Mean $\pm$ Std.	Minimum-Maximum	Range
Skin8-T4	66.750 ± 1.221	65.09-67.83	2.74
Skin40-T20	77.700 ± 1.165	76.50-79.25	2.75
Path16-T7	73.374 ± 0.553	72.57-74.13	1.56
MedMNIST-T4	73.426 ± 0.208	73.15-73.70	0.55
COVID-T6	94.124 ± 0.300	93.71-94.52	0.81
Blood-T4	93.922 ± 0.345	93.50-94.42	0.92

The mean MCR values across six datasets for the five seeds are

$$80.3133, \quad 79.6183, \quad 79.8917, \quad 79.6633, \quad 79.9267, \tag{G.3}$$

with an overall summary of

$$79.8827 \pm 0.2763. \tag{G.4}$$

Random initialization causes limited variation on individual datasets, especially Skin8-T4 and Skin40-T20. Across six datasets, mean MCR has a standard deviation of 0.2763 percentage points. Thus, the overall RA benefit does not depend on one favorable random projection.

G.3 Consistency across Sample Orders and Data Partitions

With frozen encoders and fixed RA mappings, we compared sample orders and data partitions. This measures finite precision effects on accumulated statistics and model outputs.

On Skin40-T20, we fixed $d_x = 5120$. We compared the canonical order, random sample permutations, and different data chunks. For each accumulated statistic or matrix valued model quantity Θ, let Θ_{ref} be its value in the canonical setting. We compute the relative difference

$$E_{\mathrm{rel}} = \frac{\|\Theta - \Theta_{\mathrm{ref}}\|_F}{\|\Theta_{\mathrm{ref}}\|_F}. \tag{G.5}$$

For sufficient statistics, we apply the norm to their concatenated entries. MCR entries show absolute differences in percentage points. We report the maximum difference across orders and partitions.

All settings yield

$$\mathrm{MCR}_{\text{Alignment only}} = 69.75, \qquad \mathrm{MCR}_{\mathrm{Full}} = 79.25. \tag{G.6}$$

Relative differences in sufficient statistics and U^t remain near 10^{-7}. Later linear solves amplify these differences. The largest relative difference in the Visual Uncertainty-Calibrated Alignment

Table G.4: Maximum numerical differences across sample orders and chunk partitions.

Quantity	Maximum difference
Sufficient statistics	1.7044×10^{-7}
U^t	2.5051×10^{-7}
Visual Uncertainty-Calibrated Alignment Module mapping	1.8947×10^{-2}
Label-Guided Decision Module weights	1.403×10^{-4}
Visual Uncertainty-Calibrated Alignment Module MCR	0.00
Full MCR	0.00

Module mapping is 1.8947×10^{-2}. However, the reported final MCR is unchanged. Both Visual Uncertainty-Calibrated Alignment Module and Full give the same MCR across the tested orders and partitions.

The tested orders and partitions produce identical final MCR. Intermediate statistics and model quantities still show small numerical differences.

G.4 Data Accumulation and Representation Capacity

We ask whether more training data improve classification and whether RA capacity limits these gains. We therefore varied training set size and RA dimension together.

We tested these training data fractions:

$$\varrho \in \{10\%, 25\%, 50\%, 75\%, 100\%\}, \tag{G.7}$$

using strictly nested subsets balanced across classes:

$$\mathcal{D}_{10} \subset \mathcal{D}_{25} \subset \mathcal{D}_{50} \subset \mathcal{D}_{75} \subset \mathcal{D}_{100}. \tag{G.8}$$

We kept validation and test splits fixed. We evaluated

$$d_x \in \{512, 1024, 2560, 5120\}. \tag{G.9}$$

Performance differences therefore reflect training data and RA capacity.

More training data generally improve performance. However, datasets differ in the rate of improvement and onset of saturation. Skin8-T4 and Skin40-T20 retain gains at higher fractions. Path16-T7, MedMNIST-T4, and Blood-T4 stabilize earlier.

To separate the effects of data size and representation capacity, define

$$\Delta_{\text{data}} = \text{MCR}_{100\%,5120} - \text{MCR}_{50\%,5120}, \tag{G.10}$$

and

$$\Delta_{\text{cap}} = \text{MCR}_{100\%,5120} - \text{MCR}_{100\%,2560}. \tag{G.11}$$

The first measures the gain from more data at high capacity. The second measures the gain from more RA capacity with all training data.

Skin40-T20 shows the strongest gains from data and capacity. At $d_x = 5120$, raising the training fraction from 50% to 100% improves MCR by 7.25 points. With all data, raising RA dimension from 2560 to 5120 adds 3.75 points. A larger space can retain useful structure when the statistics have not saturated.

By contrast, for Path16-T7,

$$\Delta_{\text{data}} = +0.43, \qquad \Delta_{\text{cap}} = +0.23, \tag{G.12}$$

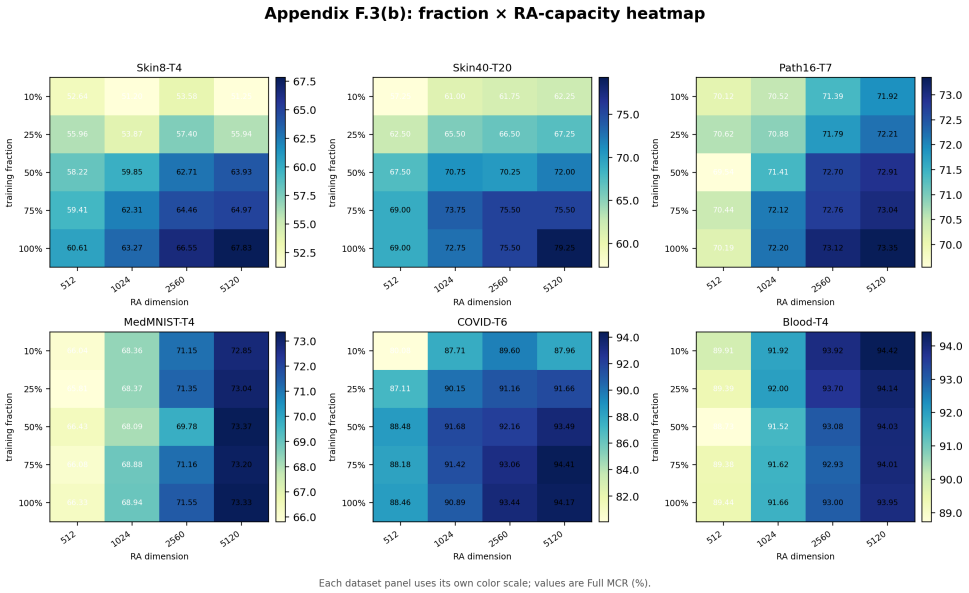

Figure G.1: Full MCR by training fraction and RA dimension. Each panel shows one dataset. Subsets are nested. Validation and test splits remain fixed.

Table G.5: MCR differences attributable to training data size and RA capacity.

Dataset	Δ_{data}	Δ_{cap}
Skin8-T4	+3.90	+1.28
Skin40-T20	+7.25	+3.75
Path16-T7	+0.43	+0.23
MedMNIST-T4	-0.04	+1.79
COVID-T6	+0.68	+0.72
Blood-T4	-0.08	+0.95

These values indicate limited sensitivity to more data or capacity in this setting. MedMNIST-T4 and Blood-T4 have near zero Δ_{data}. Yet, a larger RA dimension improves MCR by 1.79 and 0.95 points. Data saturation and capacity saturation are distinct.

The six datasets have no common saturation threshold. A dataset may be limited by training statistics or representation capacity. It may also be stable with respect to both. RA capacity controls how much statistical structure can support classification. This effect varies by dataset.

These results support the main setting of $d_x = 5120$. It gives the highest mean across six datasets and reduces early plateaus when additional useful statistics remain available.