BEYOND CLASS IMBALANCE: TPT-ROBUST COST-MATRIX EMBEDDING FOR COST-SENSITIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cost-sensitive classification is critical when prediction errors incur unequal costs, but class-frequency correction and scalar loss reweighting can obscure the directed pairwise structure of the task. A class may be frequent yet safety-critical, and the cost of confusing one class with another need not equal the reverse error. We introduce NICME (Non-Identical Cost-Matrix Embeddings), a cost-sensitive training objective for deep visual classification with non-identical, asymmetric cost matrices. NICME embeds each true-label row of the cost matrix into logit space through pairwise cost margins, together with a normalized expected-cost regularizer on the original predictive distribution. An exact finite-loss characterization shows how above-baseline costs increase the raw-logit separation required from costly wrong classes, including at arbitrarily small positive loss levels. The objective retains ordinary raw-logit inference. Experiments combine balanced binary and multiclass tasks, calibrated minimum-cost decision baselines, an eight-seed component study, and six dense asymmetric cost matrices on CIFAR-100. NICME provides favorable cost–utility operating points on the retained PMI-20 and BreaKHis comparisons and lowers mean raw-argmax cost in five of the six CIFAR-100 settings. Component and sensitivity studies identify the directed margin as a useful mechanism and characterize the role of regularization, calibration, and cost scale.

1 INTRODUCTION

Modern image classifiers are typically trained as if all errors are equally undesirable (Bartlett et al., 2006). This convention is mathematically convenient, but it is rarely the decision problem faced at deployment (Elkan, 2001; Bach et al., 2006; Chen et al., 2025). For example, identifying a poisonous mushroom as edible is generally more costly than the opposite (White et al., 2019). This scenario, along with many others such as medical screening, security inspection, harmful species recognition, and industrial defect detection, illustrates that the same overall accuracy can correspond to very different operational risks (Kumagai et al., 2024; Chen et al., 2025). Standard accuracy-oriented classification hides these distinctions by treating every off-diagonal confusion as a unit-cost error (Bartlett et al., 2006). Cost-sensitive learning makes this implicit assumption explicit: a classifier should be trained and evaluated with respect to a task-specific matrix C, where $C_{y,k}$ is the cost of predicting class k when the true class is y (Elkan, 2001; Domingos, 1999; Bach et al., 2006; Zhou and Liu, 2010).

Despite its practical importance, cost-sensitive learning remains difficult to reconcile with modern deep neural networks (Narasimhan and Menon, 2021; Chen et al., 2025). A large body of work handles complex objectives by reducing them to cost-sensitive subproblems, often solved by reweighting examples or classes (Agarwal et al., 2018; Cotter et al., 2019; Narasimhan and Menon, 2021). This approach fits naturally within surrogate empirical-risk minimization and stochastic-gradient training (Vapnik, 1999). However, overparameterized neural networks can often interpolate the training set (Zhang et al., 2017; Arpit et al., 2017; Belkin et al., 2019; Bartlett et al., 2020). Once a model drives the training error close to zero, multiplying the loss of a memorized example by a larger scalar may no longer meaningfully change the learned decision boundary. In this regime, a cost-weighted objective can share nearly the same empirical minimizers as ordinary accuracy maximization, even though

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.

their deployment goals differ. Thus, the central question is not only how to define a cost matrix, but how to preserve its influence during optimization in high-capacity models.

A second obstacle is that cost sensitivity is frequently conflated with class imbalance (He and Garcia, 2009; Johnson and Khoshgoftaar, 2019; Menon et al., 2021). Long-tailed recognition asks a model to correct for skewed class frequencies; cost-sensitive recognition asks a model to avoid harmful directed confusions (Buda et al., 2018; Elkan, 2001). These are distinct problems. A rare class need not be safety-critical, a common class may be involved in severe errors, and the cost of confusing y with k need not equal the cost of confusing k with y. Class-prior correction, resampling, and diagonal class weighting can therefore mask the pairwise structure of the application cost (Chawla et al., 2002; Cui et al., 2019; Cao et al., 2019; Menon et al., 2021). Balanced-class experiments are consequently an essential diagnostic rather than a simplification: they remove the class-prior effect and test whether a method actually uses the cost matrix.

This paper proposes NICME, a cost-sensitive training method for deep visual classification with non-identical, asymmetric cost matrices. NICME is based on a simple principle: the cost matrix should alter the geometry of the logits, not merely rescale the empirical loss after the fact. For each training example with true label y, NICME applies a pairwise cost-dependent margin to every non-target logit k, so that the loss requires a larger raw-logit gap between the true-class logit and the logit of every high-cost wrong class (Menon et al., 2021; Narasimhan and Menon, 2021). A normalized expected-cost regularizer is then applied to the model's original predictive distribution, penalizing probability mass assigned to costly wrong classes (Dmochowski et al., 2010; Scott, 2012). The resulting objective combines a cost-margin mechanism with a normalized expected-cost mechanism, both of which remain active beyond the point at which ordinary cross-entropy has merely classified the training example correctly.

Our contributions are as follows.

- We formulate pairwise cost-sensitive visual classification as a problem distinct from long-tailed recognition, and argue that balanced-class evaluation is necessary to isolate cost sensitivity from class-prior effects.
- We introduce NICME, a cost-matrix embedding objective that pairs cost-dependent logit margins with a normalized expected-cost regularizer. The cost matrix specifies an entry-wise finite-loss margin requirement, giving a direct geometric interpretation even at small loss values.
- We evaluate binary and multiclass tasks, six dense CIFAR-100 cost matrices, calibration and minimum-cost decisions, and controlled component and tuning studies. These experiments characterize both useful operating points and practical parameter selection.

2 Observations and Motivations

2.1 Cost Sensitivity Versus Class Imbalance

A central observation in this work is that cost sensitivity and class imbalance are often implemented with the same mathematical primitive, even though they describe different objects. Cost-sensitive learning begins from a decision-theoretic loss matrix C, where $C_{y,k}$ is the cost incurred when the true class is y and the classifier predicts k (Duda and Hart, 1973, pp. 10–34); see also Elkan (2001), Bach et al. (2006), Scott (2012), and Ramaswamy and Agarwal (2016). For a classifier $h : \mathcal{X} \to [K]$, where $\mathcal{X}$ is the input space and $[K] = \{1, \ldots, K\}$ is the K-class label set, the population cost risk under deployment distribution P is

$$R_C(h) = \mathbb{E}_{X \sim P_X}\left[\sum_{c=1}^{K} C_{c,h(X)} P(Y = c \mid X)\right] = \mathbb{E}_{(X,Y)\sim P}\left[C_{Y,h(X)}\right]. \tag{1}$$

Here, X is an input random variable, Y is its true label, $h(X)$ is the predicted label, and $C_{Y,h(X)}$ is the cost incurred by that prediction. Class imbalance, in contrast, concerns the true-label prior $\pi_y = P(Y = y)$, or its empirical counterpart $\widehat{\pi}_y = n_y/n$, and seeks to prevent the learner from underperforming on rare labels (He and Garcia, 2009; Buda et al., 2018; Johnson and Khoshgoftaar, 2019). Common long-tailed and imbalanced-learning methods therefore modify the data distribution,

the loss, the margin, or the prediction rule using class-frequency-derived quantities rather than pairwise deployment costs (Chawla et al., 2002; Lin et al., 2017; Cui et al., 2019; Cao et al., 2019; Kang et al., 2020; Ren et al., 2020; Menon et al., 2021).

This distinction is obscured in modern deep learning because cost information and class-frequency information are usually collapsed into a single weight, cost, or gain object inside a surrogate empirical risk. For example, inverse-prior weighting, balanced accuracy, and worst-case recall can be expressed through diagonal gain or cost matrices, whereas pairwise cost-sensitive classification uses a full matrix whose off-diagonal entries encode directed harms (Brodersen et al., 2010; Narasimhan and Menon, 2021; Chen et al., 2025). Narasimhan and Menon (2021) show that many non-decomposable objectives are optimized by reducing them to a sequence of cost-sensitive learning problems, with a gain matrix G that may be diagonal for class-recall objectives or non-diagonal for coverage and constraint objectives. At the primitive optimization level, both class imbalance and cost sensitivity can be written as surrogate empirical risk minimization (ERM) or as distributionally robust surrogate ERM (Appendix C). A shared optimization form can hide what a weighting factor represents. A large weight may reflect sample scarcity, severe deployment harm, or both, even though these are different modeling assumptions.

This motivates our balanced-class protocol. When the empirical training prior $\widehat{\pi}_y$ is uniform, class-prior corrections become constant, whereas directed deployment costs remain nonuniform. Balanced evaluation therefore isolates the cost-matrix channel; a separate natural-prior experiment examines their composition.

2.2 Class Imbalance Model-Setting Collapse

A second consequence of distinguishing cost sensitivity from class imbalance is that many class-imbalance objectives collapse when the empirical class prior is balanced, whereas the cost-sensitive task does not. This is easiest to see from the common form of frequency-based losses with weighted cross-entropy:

$$\ell_{\mathrm{wCE}}(y, z) = a_y \, \ell_{\mathrm{CE}}(y, z), \tag{2}$$

where $y \in [K]$ is the true label, $z \in \mathbb{R}^K$ is the logit vector, ℓ_{CE} is ordinary cross-entropy, and $a_y = g(\widehat{\pi}_y, n_y) > 0$ is a class weight derived from the empirical training-set true-label prior $\widehat{\pi}_y = n_y/n$ or class count n_y. If n_y is constant across classes, then a_y is constant and minimizing Eq. (2) is equivalent to minimizing ordinary cross-entropy up to a scalar. Similarly, class-prior logit adjustment and Balanced Softmax insert terms derived from $\widehat{\pi}_y$ or n_y into the logits; under balanced sampling, these terms are constant across candidate classes and cancel inside the softmax (Ren et al., 2020; Menon et al., 2021). Cost sensitivity is different: changing the number of images per class does not change the semantic object $C_{y,k}$. A balanced medication, medical-screening, security, or defect-inspection dataset can still contain directed errors whose consequences are asymmetric and severe. This observation motivates NICME: instead of relying on imbalance-derived weights, it embeds each pairwise cost into raw-logit margins and normalized expected-cost regularization.

3 Related Work

Cost-sensitive learning and class imbalance. Cost-sensitive learning replaces uniform classification error with a task loss matrix and has classical roots in expected-cost prediction, MetaCost-style relabeling, post-hoc cost-sensitive decision rules, and calibrated surrogate-risk analysis (Domingos, 1999; Zadrozny and Elkan, 2001; Elkan, 2001; Bartlett et al., 2006; Scott, 2012). Class imbalance is a separate distributional problem, where the label prior is skewed and rare classes are underrepresented (Chawla et al., 2002; He and Garcia, 2009; Johnson and Khoshgoftaar, 2019). Deep long-tail methods typically modify data, losses, margins, or predictions using class-frequency information rather than a user-specified pairwise cost matrix (Lin et al., 2017; Cui et al., 2019; Cao et al., 2019; Kang et al., 2020; Ren et al., 2020; Menon et al., 2021).

Modern cost-aware deep learning. Logit adjustment has emerged as an effective alternative to scalar reweighting for long-tailed and constrained objectives: Menon et al. (2021) shift logits according to label frequencies, while Narasimhan and Menon (2021) extend logit-adjusted losses to non-decomposable objectives and general gain matrices. Another strategy is to add cost-aware

regularization to a standard loss, consistent with broader neural-network regularization practice (Krogh and Hertz, 1992; Srivastava et al., 2014; Galdran et al., 2020). CSADA is the most direct prior work for the overparameterized setting: Chen et al. (2025) generate targeted adversarial examples for costly pairs to change the decision boundary. NICME targets the same failure mode without adversarial image generation: the cost matrix is embedded directly into clean-data logits and expected-cost probability mass.

4 Proposed Method

We now introduce NICME, a cost-sensitive surrogate loss for deep multiclass classification with fixed non-identical cost matrices. The method has two components. First, a pairwise cost-margin cross-entropy term embeds each true-label row of the cost matrix into logit space. Second, a normalized expected-cost regularizer penalizes probability mass assigned to costly wrong classes.

Notation. Throughout this section, K denotes the number of classes and $[K] = \{1, \dots, K\}$ denotes the label set. The training set is $S = \{(x_i, y_i)\}_{i=1}^{n}$, where n is the number of training examples, $x_i \in \mathcal{X}$ is the ith input, and $y_i \in [K]$ is its label. The model is a logit function $f_\theta : \mathcal{X} \to \mathbb{R}^K$, where θ denotes all trainable network parameters. For any input x, we write $z = f_\theta(x)$ for its logit vector and $z_k = f_{\theta,k}(x)$ for the logit of class k. The cost matrix is $C \in \mathbb{R}_+^{K \times K}$, with $C_{y,k}$ denoting the cost of predicting class k when the true class is y.

4.1 Problem Setting

A model f_θ produces softmax probabilities

$$p_\theta(k \mid x) = \mathrm{softmax}_k(f_\theta(x)) = \frac{\exp(f_{\theta,k}(x))}{\sum_{j=1}^{K} \exp(f_{\theta,j}(x))}. \tag{3}$$

Here, $p_\theta(k \mid x)$ is the predicted probability of class k for input x. The associated argmax classifier is $h_\theta(x) \in \arg\max_{k \in [K]} f_{\theta,k}(x)$. We assume a nonnegative cost matrix C with $C_{y,y} = 0$ for every class y. Entry $C_{y,k}$ is the cost of predicting k when the true class is y. We allow the matrix to be non-identical and asymmetric, so $C_{y,k}$ may differ from $C_{k,y}$, and two errors from the same true class may have different costs. The population objective is to minimize the cost-sensitive risk $R_C(h)$ in Eq. (1); the equivalent confusion-matrix expression and Bayes rule are provided in Appendix A.

A direct smooth surrogate is

$$\widehat{R}_C^{\mathrm{prob}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} C_{y_i,k} p_\theta(k \mid x_i). \tag{4}$$

However, when an overparameterized network interpolates the training set, it enters the terminal phase of training (TPT), in which $p_\theta(y_i \mid x_i) \to 1$ and $p_\theta(k \mid x_i) \to 0$ for every $k \neq y_i$ (Papyan et al., 2020). Since $C_{y_i,y_i} = 0$, the empirical expected cost can approach zero in TPT regardless of the off-diagonal entries of C (Chen et al., 2025). NICME is constructed to avoid this collapse by embedding $C_{y,k}$ into the logits of each training example.

4.2 NICME Formulation

TPT-safe logit adjustment. For a sample (x, y), let $z = f_\theta(x)$ be the raw logit vector. For the margin term, costs are expressed on a unit-baseline scale: correct predictions have cost 0, ordinary noncritical off-diagonal errors have cost 1, and costs greater than 1 denote errors that should require additional separation. Define the clamped pairwise cost

$$M_{y,k} = \max(C_{y,k}, 1). \tag{5}$$

The clamping prevents diagonal or subunit entries from producing undefined or negative logarithmic offsets, while unit-baseline errors receive zero additional offset. For a cost-to-margin scale parameter $\alpha \geq 0$, NICME defines adjusted logits

$$\bar{z}_k = z_k + \alpha(1 - \delta_{k,y}) \log M_{y,k}, \tag{6}$$

where $\delta_{k,y}$ is the Kronecker delta. Thus, the true-class logit is unchanged, while each non-target logit receives a cost-dependent offset controlled by α. NICME applies ordinary cross-entropy to $\bar{z}$, giving the pairwise cost-margin loss

$$\ell_{\mathrm{pcm}}(y, z; C, \alpha) = -\log \mathrm{softmax}_y(\bar{z}) = \log\left(1 + \sum_{k \neq y} \exp(z_k - z_y) M_{y,k}^{\alpha}\right). \tag{7}$$

When $\alpha = 0$, Eq. (7) reduces exactly to ordinary cross-entropy. For $\alpha > 0$ and $C_{y,k} > 1$, the wrong class k receives a larger contribution, increasing the separation needed to achieve the same loss tolerance.

We make precise the sense in which the objective retains pairwise cost dependence at small loss.

Definition 1 (Finite-loss TPT robustness). For fixed $\alpha > 0$, we use *TPT robustness* to mean that the necessary raw-logit margin at any positive loss tolerance contains an additive term increasing with each above-baseline directed cost. This is a finite-loss property of the surrogate; it does not assert a particular asymptotic decision boundary.

Proposition 1 (Cost-dependent margin). Fix $\alpha \geq 0$, a true label y, and $\varepsilon > 0$. If $\ell_{\mathrm{pcm}}(y, z; C, \alpha) \leq \varepsilon$, then for every $k \neq y$,

$$z_y - z_k \geq \alpha \log M_{y,k} - \log(e^{\varepsilon} - 1). \tag{8}$$

Indeed, Eq. (7) implies $\sum_{k \neq y} \exp(z_k - z_y) M_{y,k}^{\alpha} \leq e^{\varepsilon} - 1$; bounding each nonnegative summand and taking logarithms proves the result. Thus increasing an above-baseline cost by a factor q raises its necessary margin by $\alpha \log q$. The guarantee is entry-wise: it connects the declared cost to a required logit gap at every finite tolerance, including arbitrarily small ones. It is not a proof of posterior calibration, Bayes consistency, or lower population cost risk. Weighted losses may still influence optimization; NICME supplies an explicit directed-margin construction. Appendix E also gives the exact adjusted-odds equivalence and the regularizer gradient.

Cost-sensitive regularizer. The logit-adjusted term addresses pairwise separation, while the regularizer addresses probability mass. Define normalized costs

$$\widetilde{C} = \frac{C}{C_{\max} + \varepsilon_{\mathrm{num}}}, \qquad C_{\max} = \max_{a,b \in [K]} C_{a,b}, \tag{9}$$

where $\varepsilon_{\mathrm{num}} > 0$ prevents division by zero when all costs are zero. Let $p = \mathrm{softmax}(z)$ denote the probability vector computed from the original, unadjusted logits. The normalized expected-cost regularizer is

$$r_C(y, z) = \widetilde{C}_{y,\cdot}^{\top} p = \sum_{k=1}^{K} \widetilde{C}_{y,k} p_k. \tag{10}$$

The regularizer gradient, derived in Appendix E, lowers logits for classes whose row cost is above the model's current expected cost.

The final single-example NICME loss is

$$\ell_{\mathrm{NICME}}(y, z; C, \alpha, \lambda) = \ell_{\mathrm{pcm}}(y, z; C, \alpha) + \lambda r_C(y, z), \tag{11}$$

where $\lambda \geq 0$ is the expected-cost regularization weight. The empirical NICME objective is

$$\widehat{L}_{\mathrm{NICME}}(\theta; C, \alpha, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left[\log\left(1 + \sum_{k \neq y_i} \exp(f_{\theta,k}(x_i) - f_{\theta,y_i}(x_i)) M_{y_i,k}^{\alpha}\right) + \lambda \sum_{k=1}^{K} \widetilde{C}_{y_i,k} \, \mathrm{softmax}_k(f_\theta(x_i)) \right]. \tag{12}$$

This objective is a drop-in replacement for cross-entropy: it uses the same logits and argmax inference, requires no adversarial inner loop, and introduces only two cost-strength hyperparameters, α for pairwise margins and λ for expected-cost regularization. At inference, $\bar{z}$ is not used: predictions and probabilities come from z. Gathering one cost row per example adds $O(BK)$ loss work for a batch of size B; a dense matrix uses $O(K^2)$ storage. A 1,000-class float32 matrix requires 4 MB. These are computational properties, separate from empirical large-class performance.

Cost-aware decisions and parameter selection. Given estimated probabilities $\widehat{p}$, a complementary plug-in rule is

$$\widehat{h}_C(x) \in \arg\min_j \sum_y \widehat{p}(y \mid x) C_{y,j}. \tag{13}$$

With the true posterior this is the Bayes rule; its estimated version provides a direct practical baseline. We evaluate it after held-out calibration. For training, costs above one have an explicit margin interpretation. If an application declares an added-margin budget $b > 0$ and $M_{\max} > 1$, then $\alpha \leq b/\log M_{\max}$ is an engineering cap. Select λ and the operating point on validation data under declared cost and utility criteria. The normalized regularizer is invariant to uniform positive cost scaling apart from its stabilizer; the clamped log-margin is not. Cost units and the unit baseline must therefore be declared.

5 Experiments

We evaluate whether a model uses directed costs after class-prior effects have been controlled, then examine calibration, larger class counts, and joint prior–cost settings. Cost and recognition metrics are reported together: a favorable operating point should avoid harmful errors while retaining useful classification performance.

5.1 Datasets

PMI-20 is the primary multiclass benchmark: 20 medication classes with sparse, directed high-cost confusions following the pharmacist-reviewed PMI/CSADA protocol (Chen et al., 2025). Spider and BreaKHis provide binary stress tests. Spider distinguishes black widow from false widow; BreaKHis distinguishes benign from malignant breast histopathology (Spanhol et al., 2016). Their matrices are respectively $[0, 8; 1, 0]$ and $[0, 1; 7, 0]$, in those class orders. These transparent decision-context ratios encode asymmetric experimental objectives, not externally validated clinical or toxicological utilities. Appendix F gives their construction, splits, and access conditions.

CIFAR-100 (Krizhevsky, 2009) tests dense costs across 100 fine classes and 20 superclasses. Let $s(y)$ denote the official coarse-class index. Costs are zero on the diagonal, one for within-superclass errors, four for ordinary cross-superclass errors, and $4r$ in an emphasized direction, with $r \in \{2, 4, 8\}$. We test both $s(k) > s(y)$ (up) and $s(k) < s(y)$ (down). These index-based directions are synthetic controls, not semantic harm rankings. Each matrix has 100 zeros, 400 ones, 4,750 fours, and 4,750 emphasized entries. The official training set is stratified into 40,000/5,000/5,000 train/validation/calibration examples; the 10,000-image test set is retained unchanged.

5.2 Comparison Methods and Metrics

Comparisons include CE, cost-weighted CE, expected-cost regularization, Menon class-prior logit adjustment, a CSADA-inspired adversarial adaptation, the general gain-matrix hybrid of Narasimhan and Menon (2021), and calibrated CE with Eq. (13). The gain-matrix model is evaluated by score argmax, as intended by that loss. We use raw-logit argmax for NICME unless another decision rule is explicitly named.

Average test cost is $\mathrm{ATC} = N^{-1} \sum_i C_{y_i, \widehat{y}_i}$; $\mathrm{nATC} = \mathrm{ATC}/C_{\max}$ for the nonzero matrices studied. Normalization expresses cost relative to each matrix's maximum; it does not equate the utilities of different tasks or cost matrices. We report balanced accuracy (BA), macro-F1, critical-pair error counts, and target-minimum recall, the minimum recall among designated costly-source classes. Appendix G distinguishes this from target-macro recall used in the historical comparisons. Calibration uses 15-bin top-label ECE, NLL, and class-summed Brier score.

5.3 Experimental Setting

All vision experiments use ImageNet-pretrained ConvNeXt-Base (Liu et al., 2022) with 224-pixel inputs. The retained task comparisons use three checkpoints per method. Separate controlled PMI-20 experiments use eight aligned training/data seeds (42–49), and matched searches and CIFAR-100 use three seeds (42–44). We report mean $\pm$ sample SD on fixed splits, not uncertainty over new dataset

Table 1: Cost and utility on the retained models. Values are mean ± sample SD over three retained runs. Scaling uses temperature on PMI-20 and affine Platt scaling on the binary tasks, fitted on held-out calibration data. These descriptive runs are separate from the controlled-seed experiments in Table 3. Lower nATC is better.

Dataset	Method / decision	nATC	Bal. acc.	Target-min
PMI-20	CE, argmax	0.00568 ± 0.00310	0.9714 ± 0.0174	0.9167 ± 0.0180
PMI-20	CE, scaled cost-min	0.00547 ± 0.00118	0.9594 ± 0.0217	0.8854 ± 0.0180
PMI-20	NICME, argmax	0.00370 ± 0.00095	0.9771 ± 0.0106	0.9167 ± 0.0180
Spider	CE, argmax	0.02056 ± 0.00792	0.9367 ± 0.0100	0.9711 ± 0.0168
Spider	CE, scaled cost-min	0.01375 ± 0.00451	0.9056 ± 0.0234	0.9956 ± 0.0038
Spider	NICME, argmax	0.01875 ± 0.00439	0.9356 ± 0.0158	0.9756 ± 0.0139
BreaKHis	CE, argmax	0.05902 ± 0.02736	0.8954 ± 0.0251	0.8972 ± 0.0662
BreaKHis	CE, scaled cost-min	0.04728 ± 0.01305	0.8676 ± 0.0421	0.9338 ± 0.0168
BreaKHis	NICME, argmax	0.03259 ± 0.00919	0.8641 ± 0.0461	0.9693 ± 0.0364

Table 2: CIFAR-100 across six directed cost matrices. Raw-logit argmax; nATC is mean ± sample SD over three paired seeds. Balanced accuracy (BA) and emphasized-direction error counts are means. The same CE anchors are rescored under each matrix; matrix rows are not independent replications.

Direction, r	CE nATC	NICME nATC	CE BA	NICME BA	CE err.	NICME err.
up, 2	0.03090 ± 0.00214	0.03169 ± 0.00194	0.9224	0.9197	172.3	178.7
up, 4	0.02406 ± 0.00165	0.02352 ± 0.00065	0.9224	0.9251	172.3	168.3
up, 8	0.02065 ± 0.00146	0.01996 ± 0.00059	0.9224	0.9245	172.3	165.7
down, 2	0.03043 ± 0.00251	0.02936 ± 0.00269	0.9224	0.9229	163.0	147.7
down, 4	0.02336 ± 0.00215	0.02122 ± 0.00039	0.9224	0.9246	163.0	137.0
down, 8	0.01983 ± 0.00199	0.01768 ± 0.00044	0.9224	0.9149	163.0	133.0

draws. The reference NICME setting is $(\alpha, \lambda) = (0.5, 0.1)$, with a two-epoch regularizer warmup where configured.

PMI-20 controlled training uses LR 5×10^{-5}, weight decay 0.005, batch size 16 with accumulation 4, a cosine schedule, a 32-epoch ceiling, and patience 5. Checkpoints maximize validation target-min recall with earliest-epoch ties. CIFAR-100 uses a 16-epoch ceiling and patience 3; NICME checkpoints minimize validation nATC subject to BA ≥ 0.70, with no CIFAR hyperparameter search. CE anchors are fixed before evaluation. Selection details and distinct historical adaptation protocols appear in Appendix H.

5.4 Cost, Calibration, and Decision Rules

Table 1 compares training-time cost embedding with calibrated post-hoc decisions. NICME has lower mean nATC than scaled CE cost-min on PMI-20 and BreaKHis. On PMI-20, its lower cost accompanies higher balanced accuracy and target-min recall; on BreaKHis, it raises target-min recall with similar balanced accuracy. Scaled CE achieves lower cost on Spider, illustrating the value of evaluating both training and decision mechanisms. The complete historical trained-baseline comparisons are in Appendix I.

NICME is compatible with probability calibration because its label-conditioned transformation is confined to training. On retained PMI-20 models, raw ECE decreases from 0.02417 for CE to 0.01737 for NICME, NLL from 0.10550 to 0.07928, and Brier score from 0.04659 to 0.03356. Binary results depend on the metric and calibration map: after Platt scaling, CE has lower mean ECE on both tasks, while NICME has lower NLL and Brier score on BreaKHis. Applying temperature-scaled cost-min prediction to NICME further yields PMI-20 nATC 0.00297. These results support combining cost-aware training with calibration rather than treating them as competing requirements. Appendix J reports every evaluated map and decision rule.

Table 3: Controlled PMI-20 component study: eight seeds (42–49), shared training envelope and selection criterion, raw argmax. Values are mean ± sample SD. Parameters in parentheses are (α, λ).

Objective	nATC	Bal. acc.	Target-min
CE	0.00703 ± 0.00422	0.9631 ± 0.0140	0.8945 ± 0.0552
Menon adjustment	0.00580 ± 0.00228	0.9672 ± 0.0129	0.9141 ± 0.0277
Margin only $(.5, 0)$	0.00410 ± 0.00110	0.9766 ± 0.0088	0.9297 ± 0.0221
Regularizer only $(0, .1)$	0.00727 ± 0.00505	0.9604 ± 0.0272	0.8789 ± 0.0540
NICME $(.5, .1)$	0.00723 ± 0.00693	0.9490 ± 0.0571	0.8906 ± 0.0442

Table 4: Separate matched-search PMI-20 protocol. Nine configurations per trainable family; three final seeds; mean ± sample SD. Search counts and training envelopes are matched, while realized computation and selection criteria differ.

Method / decision	nATC	Bal. acc.	Target-min
CE	0.00510 ± 0.00237	0.9760 ± 0.0106	0.9063 ± 0.0313
CE, temperature cost-min	0.00490 ± 0.00163	0.9698 ± 0.0086	0.8750 ± 0.0000
NICME	0.01036 ± 0.00899	0.9188 ± 0.0943	0.8854 ± 0.0361
Regularized CE	0.00734 ± 0.00151	0.9583 ± 0.0213	0.8958 ± 0.0180
Gain-matrix hybrid	0.04078 ± 0.04620	0.5922 ± 0.4620	0.5729 ± 0.5081

5.5 Dense Asymmetric Costs on CIFAR-100

Table 2 extends evaluation to six dense 100-class matrices. NICME lowers mean raw-argmax nATC in five settings. For down, $r = 4$, mean nATC falls from 0.023365 to 0.021223 (9.17%), with a favorable difference in every paired seed. Emphasized-direction errors decrease from 163 to 137 (15.95%), while balanced accuracy increases from 0.9224 to 0.9246. Within- and cross-superclass error rates also decline, from 0.044100 to 0.042100 and from 0.033533 to 0.033267. This operating point combines lower directed cost with slightly stronger recognition.

The complete family shows how the effect depends on the declared asymmetry: up, $r = 2$ favors CE; down, $r = 8$ attains the largest relative cost reduction in this family with a recognition tradeoff. With temperature-scaled cost-min decisions for both methods, NICME has lower mean cost in four of six settings. We report these as paired descriptive effects over three seeds; Appendix L includes uncertainty summaries, all decision modes, and cross-matrix evaluations.

5.6 Component and Tuning Studies

The eight-seed study in Table 3 isolates the two mechanisms under one training envelope. The margin-only configuration achieves the lowest mean cost and highest mean balanced accuracy and target-min recall among these variants. This supports directed separation as a useful component and identifies regularization strength as an operating-point choice. Full NICME improves on CE in six of eight seed pairs, but its mean cost is similar and more variable; the paired mean difference is $+0.000195$, with a 95% paired-bootstrap interval $[-0.004648, 0.006621]$. The two predeclared exact tests do not establish full-objective superiority (Appendix K). All seeds are retained.

Table 4 reports a separate nine-configuration search for each trainable family. Calibrated CE is a strong reference in this protocol, and the gain-matrix comparator is highly variable. This panel is not pooled with the retained-model comparison. Together the studies motivate selecting a cost–utility operating point using held-out criteria, rather than prescribing a universally preferred regularizer strength. An exploratory eight-seed calibration extension using saved logits is included in Appendix K.

Hyperparameters, cost scale, and class priors. The historical 49-point validation grid contains 23 distinct metric vectors, with 45 points attaining target-min recall 1.0; the reference setting was preselected and then evaluated in the grid. A separate nine-setting neighborhood, each with three seeds, satisfies mean validation floors of target-min recall 0.90 and BA 0.80. Ordering within this feasible region varies by seed. Completed scale and cost-uncertainty studies characterize changes under different declared costs (Appendix M).

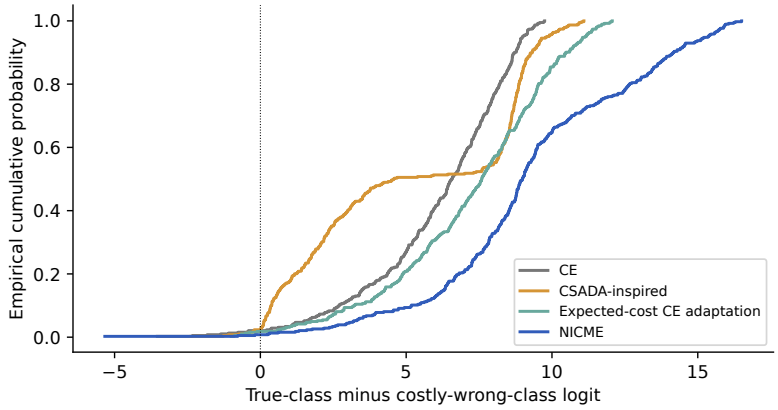

Figure 1: Empirical distributions of raw critical-pair margins on the historical PMI-20 test examples. The curve aggregates the three retained runs; it describes trained-model separation rather than an independent proof of the finite-loss proposition. Method labels distinguish expected-cost adaptation and CSADA-inspired adaptation from controlled component variants.

On natural PMI-20, adding a Menon-style training prior offset before NICME's pairwise offsets raises mean target-min recall from 0.9492 to 0.9613 and reduces its SD from 0.0339 to 0.0070. Mean nATC also rises, from 0.00370 to 0.00431. This provides a selectable recall–cost tradeoff and demonstrates that prior and directed-cost mechanisms can be composed; Appendix N reports the full comparison. Figure 1 complements the cost results by visualizing raw critical-pair separation in the historical PMI-20 models.

6 Conclusion

NICME embeds a non-identical, asymmetric cost matrix into deep-classifier logit geometry through directed pairwise margins and normalized expected-cost regularization. Its finite-loss characterization makes the contribution of each above-baseline cost explicit while retaining ordinary inference. Balanced task comparisons, calibrated decision baselines, and dense CIFAR-100 studies demonstrate useful cost–utility operating points beyond class-frequency correction. Controlled component and sensitivity studies further identify how the margin, regularizer, calibration map, and declared cost scale shape those operating points. These results support treating directed cost structure as a first-class part of deep-classifier training and evaluation.

AI USE STATEMENT

Generative AI tools were used to assist with editing for readability, literature-search support, and formatting of the manuscript into the ICLR 2027 LaTeX template. All AI-assisted text and formatting were reviewed by the authors, and citations, equations, claims, and manuscript structure were checked against the underlying sources and experimental artifacts. Generative AI tools are not part of NICME, the model-training pipeline, or the experimental evaluation reported in this paper. The authors take responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses existing image datasets and does not collect new human-subject data. Spider and BreaKHis costs are transparent experimental decision-context constructions, not validated deployment utilities. Application-specific use requires expert cost elicitation and evaluation under the intended population and workflow. Raw third-party imagery and pretrained weights are not redistributed; dataset access conditions and provenance are documented in Appendix F.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix E specify the objective and proofs. Appendices F–H document data, cost matrices, metrics, and training. The anonymous supplement contains implementation code, configurations, split manifests, run-level results, numeric logits for the eight-seed reanalysis, and a CPU script that verifies aggregates and regenerates the new tables. It also includes paired statistics and complete CIFAR-100 confusion counts. Raw images must be obtained separately; the Spider source pool is not independently redistributed. Main-text captions distinguish retained runs from controlled seeds and separate experimental protocols.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. *International Conference on Machine Learning*, pp. 60–69, 2018. https://proceedings.mlr.press/v80/agarwal18a.html.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. *International Conference on Machine Learning*, pp. 233–242, 2017. https://proceedings.mlr.press/v70/arpit17a.html.

Francis R. Bach, David Heckerman, and Eric Horvitz. Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7(63):1713–1741, 2006. https://jmlr.org/papers/v7/bach06a.html.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. https://doi.org/10.1198/016214505000000907.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. https://doi.org/10.1073/pnas.1907378117.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. https://doi.org/10.1073/pnas.1903070116.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. *International Conference on Pattern Recognition*, pp. 3121–3124, 2010. https://doi.org/10.1109/ICPR.2010.764.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. https://doi.org/10.1016/j.neunet.2018.07.011.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32, 2019. https://papers.nips.cc/paper/2019/hash/621461af90cadfdaf0e8d4cc25129f91-Abstract.html.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. https://doi.org/10.1613/jair.953.

Qiyuan Chen, Raed Al Kontar, Maher Nouiehed, X. Jessie Yang, and Corey Lester. Rethinking cost-sensitive classification in deep learning via adversarial data augmentation. *INFORMS Journal on Data Science*, 4(1):1–19, 2025. https://doi.org/10.1287/ijds.2022.0033.

Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Seungil You, Karthik Sridharan, and Maya R. Gupta. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019. https://jmlr.org/papers/v20/18-616.html.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019. https://openaccess.thecvf.com/content_CVPR_2019/html/Cui_Class-Balanced_Loss_Based_on_Effective_Number_of_Samples_CVPR_2019_paper.html.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010. https://doi.org/10.1287/opre.1090.0741.

Jacek P. Dmochowski, Paul Sajda, and Lucas C. Parra. Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. *Journal of Machine Learning Research*, 11:3313–3332, 2010. https://jmlr.org/papers/v11/dmochowski10a.html.

Pedro Domingos. MetaCost: A general method for making classifiers cost-sensitive. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164, 1999. https://doi.org/10.1145/312129.312220.

Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, New York, 1973. https://www.peterhart.net/Home/selected-publications.

Charles Elkan. The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

Joann G. Elmore, Gary M. Longton, Patricia A. Carney, Berta M. Geller, Tracy Onega, Anna N. A. Tosteson, Heidi D. Nelson, Margaret S. Pepe, Kimberly H. Allison, Stuart J. Schnitt, Frances P. O'Malley, and Donald L. Weaver. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*, 313(11):1122–1132, 2015. https://doi.org/10.1001/jama.2015.1405.

Adrian Galdran, Jose Dolz, Hadi Chakor, Hervé Lombaert, and Ismail Ben Ayed. Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 665–674, 2020. https://doi.org/10.1007/978-3-030-59722-1_64.

Andis Graudins, Narendra Gunja, Kevin W. Broady, and Graham M. Nicholson. Clinical and in vitro evidence for the efficacy of Australian red-back spider (*Latrodectus hasselti*) antivenom in the treatment of envenomation by a cupboard spider (*Steatoda grossa*). *Toxicon*, 40(6):767–775, 2002. https://doi.org/10.1016/S0041-0101(01)00280-X.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning*, pp. 1321–1330, 2017. https://proceedings.mlr.press/v70/guo17a.html.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. https://doi.org/10.1109/TKDE.2008.239.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. https://doi.org/10.1109/CVPR.2016.90.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019. https://doi.org/10.1186/s40537-019-0192-5.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *International Conference on Learning Representations*, 2020. https://openreview.net/forum?id=r1g3H6NtwB.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34, 2021. https://proceedings.neurips.cc/paper/2021/hash/9dfcf16f0adbc5e2a55ef02db36bac7f-Abstract.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems*, 4:950–957, 1992. https://papers.nips.cc/paper/1991/hash/8eefcfdf5990e441f0fb6f3fad709e21-Abstract.html.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS Tutorials in Operations Research*, pp. 130–166, 2019. https://doi.org/10.1287/educ.2019.0198.

Matjaž Kukar and Igor Kononenko. Cost-sensitive learning with neural networks. *Proceedings of the 13th European Conference on Artificial Intelligence*, pp. 445–449, 1998. https://www.ida.liu.se/ext/acres/cf/ecai/1998/indlis.html.

Atsutoshi Kumagai, Tomoharu Iwata, Hiroshi Takahashi, Taishi Nishiyama, and Yasuhiro Fujiwara. AUC maximization under positive distribution shift. *Advances in Neural Information Processing Systems*, 37:36071–36096, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/3f8c7eb848ffec848f3ed2b7ca44915d-Abstract-Conference.html.

Laboratório Visão Robótica e Imagem, Universidade Federal do Paraná. Breast Cancer Histopathological Database (BreaKHis), 2022. Dataset page: https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/.

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with Wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018. https://papers.neurips.cc/paper/7534-minimax-statistical-learning-with-wasserstein-distances.pdf.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2017. https://doi.org/10.1109/ICCV.2017.324.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. https://doi.org/10.1109/CVPR52688.2022.01167.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018. https://openreview.net/forum?id=rJzIBfZAb.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=37nvvqkCo5.

Merck Manual Professional Edition. Spider bites, 2025. https://www.merckmanuals.com/professional/injuries-poisoning/bites-and-stings/spider-bites.

Montana State University Extension IPM Program. Cobweb weavers and false widows (*Steatoda* spp.), 2026. https://www.montana.edu/extension/ipm/entomology/resources/fact-sheets/cob-web-weavers.html.

Andrew A. Monte, Becki Bucher-Bartelson, and Kennon J. Heard. A US perspective of symptomatic *Latrodectus* spp. envenomation and treatment: A national poison data system review. *The Annals of Pharmacotherapy*, 45(12):1491–1498, 2011. https://doi.org/10.1345/aph.1Q424.

Harikrishna Narasimhan and Aditya Krishna Menon. Training over-parameterized models with non-decomposable objectives. *Advances in Neural Information Processing Systems*, 34, 2021. https://papers.nips.cc/paper/2021/hash/9713faa264b94e2bf346a1bb52587fd8-Abstract.html.

National Cancer Institute. Breast cancers between mammograms have aggressive features, 2011. https://www.cancer.gov/types/breast/research/interval-breast-cancer.

National Cancer Institute. Breast cancer screening (PDQ), 2025. https://www.cancer.gov/types/breast/hp/breast-screening-pdq.

National Cancer Institute. Cancer screening overview (PDQ), 2026. https://www.cancer.gov/about-cancer/screening/hp-screening-overview-pdq.

Heidi D. Nelson, Miranda Pappas, Amy Cantor, Jessica Griffin, Monica Daeges, and Linda Humphrey. Harms of breast cancer screening: Systematic review to update the 2009 U.S. Preventive Services Task Force recommendation. *Annals of Internal Medicine*, 164(4):256–267, 2016. https://doi.org/10.7326/M15-0970.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. https://doi.org/10.1073/pnas.2015509117.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019. https://arxiv.org/abs/1908.05659.

Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(14):1–45, 2016. https://jmlr.org/papers/v17/14-316.html.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33:4175–4186, 2020. https://papers.nips.cc/paper/2020/hash/2ba61cc3a8f44143e1f2f13b2b729ab3-Abstract.html.

Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012. https://doi.org/10.1214/12-EJS699.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *International Conference on Learning Representations*, 2018. https://openreview.net/forum?id=Hk6kPgZA-.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. https://jmlr.org/papers/v19/18-188.html.

Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016. https://doi.org/10.1109/TBME.2015.2496264.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. https://jmlr.org/papers/v15/srivastava14a.html.

University of California Statewide Integrated Pest Management Program. Widow spiders and their relatives, 2017. https://ipm.ucanr.edu/PMG/PESTNOTES/pn74149.html.

Vladimir N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999. https://doi.org/10.1109/72.788640.

Julian White, Scott A. Weinstein, Luc De Haro, Régis Bédry, Andreas Schaper, Barry H. Rumack, and Thomas Zilker. Mushroom poisoning: A proposed new clinical classification. *Toxicon*, 157:53–65, 2019. https://doi.org/10.1016/j.toxicon.2018.11.007.

Ross Wightman. PyTorch image models, 2019. Software library: https://github.com/huggingface/pytorch-image-models.

Mollie Williams, Neha Sehgal, and Thomas M. Nappe. Black widow spider toxicity. *StatPearls*, National Center for Biotechnology Information Bookshelf, 2023. https://www.ncbi.nlm.nih.gov/books/NBK499987/.

Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 204–213, 2001. https://doi.org/10.1145/502512.502540.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017. https://openreview.net/forum?id=Sy8gdB9xx.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006. https://doi.org/10.1109/TKDE.2006.17.

Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010. https://doi.org/10.1111/j.1467-8640.2010.00358.x.

A Additional Notation and Cost-Sensitive Risk

Let K denote the number of classes and let $[K] = \{1, \ldots, K\}$ be the label set. The input space is $\mathcal{X}$, and P denotes the population distribution over $\mathcal{X} \times [K]$. A training sample is $S = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}$ is the ith input and $y_i \in [K]$ is its label. A neural classifier is represented by a logit function $f_\theta : \mathcal{X} \to \mathbb{R}^K$ with parameters θ. For an input x, $z = f_\theta(x)$ is the logit vector and $z_k = f_{\theta,k}(x)$ is the logit for class k. The predicted probability of class k is $p_\theta(k \mid x) = \mathrm{softmax}_k(f_\theta(x))$.

The cost matrix is $C \in \mathbb{R}_+^{K \times K}$. We use the row-true, column-predicted convention: $C_{y,k}$ is the cost of predicting class k when the true class is y. Correct classifications have zero cost, $C_{y,y} = 0$. The matrix may be non-identical and asymmetric, so $C_{y,k}$ need not equal $C_{k,y}$. This convention follows classical cost-sensitive decision theory, where a task loss matrix specifies application-level consequences for each directed prediction pair (Elkan, 2001; Bach et al., 2006; Zhou and Liu, 2010; Scott, 2012; Ramaswamy and Agarwal, 2016).

For a classifier $h : \mathcal{X} \to [K]$, the population cost risk is

$$R_C(h) = \mathbb{E}_{(X,Y)\sim P}\left[C_{Y,h(X)}\right]. \tag{14}$$

When $C_{y,k} = 1$ for every $y \neq k$ and $C_{y,y} = 0$, Eq. (14) reduces to ordinary misclassification risk. For nonuniform off-diagonal costs, it encodes expected deployment cost rather than expected error rate (Elkan, 2001; Scott, 2012).

Define the population confusion matrix $\Gamma[h] \in \mathbb{R}_+^{K\times K}$ by

$$\Gamma_{y,k}[h] = P(Y = y, h(X) = k). \tag{15}$$

Then

$$R_C(h) = \langle C, \Gamma[h]\rangle = \sum_{y=1}^{K}\sum_{k=1}^{K} C_{y,k}\Gamma_{y,k}[h]. \tag{16}$$

This form separates the application-specified cost matrix C from the classifier-induced error distribution $\Gamma[h]$. A diagonal or row-wise weighting scheme can change class-level emphasis, but a full cost matrix can distinguish individual directed confusions.

Let $\eta_y(x) = P(Y = y \mid X = x)$ denote the conditional probability of class y at input x. The cost-sensitive Bayes decision is

$$h^\star_C(x) \in \arg\min_{j\in[K]} \sum_{y=1}^{K} \eta_y(x) C_{y,j}. \tag{17}$$

The summation is the conditional expected cost of predicting j. This rule differs from maximum-posterior prediction whenever the most likely class is not the lowest-cost decision (Elkan, 2001; Bartlett et al., 2006; Scott, 2012; Ramaswamy and Agarwal, 2016). Equivalently, one may use a gain matrix G and maximize $\langle G, \Gamma[h]\rangle$, as in reduction-based optimization of non-decomposable metrics (Agarwal et al., 2018; Cotter et al., 2019; Narasimhan and Menon, 2021). We use the cost convention because the experimental metrics are expressed as average test cost.

B Classical Risk and Surrogate Empirical Risk

This section clarifies the relationship between the classical decision-theoretic risk formulation of Duda and Hart (1973, pp. 10–34) and the surrogate empirical-risk formulation used by NICME (Vapnik, 1999). In classical Bayesian decision theory, choosing action k at input x incurs conditional risk

$$R_C(k \mid x) = \sum_{y=1}^{K} C_{y,k}\eta_y(x), \qquad \eta_y(x) = P(Y = y \mid X = x). \tag{18}$$

The Bayes decision chooses the action with minimum conditional risk, which is Eq. (17). For any fixed classifier h, the corresponding total risk is

$$\mathbb{E}_X[R_C(h(X) \mid X)] = \mathbb{E}_X\left[\sum_{y=1}^{K} C_{y,h(X)}P(Y = y \mid X)\right] = \mathbb{E}_{(X,Y)\sim P}[C_{Y,h(X)}]. \tag{19}$$

Thus, our population cost risk is a direct multiclass cost-matrix instance of the Duda–Hart total risk. We use the word *cost* to emphasize application consequences, but mathematically C is the Duda–Hart loss matrix, and risk is the expectation of the incurred loss under the data distribution.

The training objective used by NICME differs from directly applying the Bayes decision rule. The Duda–Hart rule requires posterior probabilities $\eta_y(x)$ and then chooses the prediction k that minimizes the conditional risk. Modern neural-network training instead usually follows the statistical-learning-theory empirical-risk principle: replace the unknown population distribution P by the empirical distribution $\widehat{P}_n$ and optimize an average loss over the sample (Vapnik, 1999). Because the discrete task loss $C_{Y,h(X)}$ is not a convenient differentiable objective for deep-network optimization, NICME minimizes the surrogate empirical risk

$$\widehat{L}_{\mathrm{NICME}}(\theta; C, \alpha, \lambda) = \mathbb{E}_{(X,Y)\sim\widehat{P}_n}\left[\ell_{\mathrm{NICME}}(Y, f_\theta(X); C, \alpha, \lambda)\right]. \tag{20}$$

The normalized expected-cost term is

$$r_C(y, z) = \sum_{k=1}^{K} \widetilde{C}_{y,k} p_k, \qquad p_k = \mathrm{softmax}_k(z). \tag{21}$$

This is an expectation over the model's predictive distribution on candidate predicted classes k, conditional on the observed training label y. It is therefore not the same object as the Duda–Hart conditional risk, which is an expectation over possible true classes y under the posterior $\eta_y(x)$ for a fixed candidate action k. The two are connected by the same cost matrix, but they play different roles: Duda–Hart conditional risk defines the optimal cost-sensitive decision rule when the true posteriors are available, whereas $r_C(y, z)$ is a differentiable training regularizer that discourages probability mass on costly wrong labels.

NICME also does not explicitly use class-prior probabilities $\pi_y = P(Y = y)$. Class priors enter only through the sampled training distribution, unless one deliberately combines the cost matrix with a separate class-imbalance method. Balanced experiments isolate this channel; Appendix N additionally evaluates composition with prior correction on natural PMI-20.

C Cost Sensitivity, Class Imbalance, and Surrogate ERM

Cost sensitivity and class imbalance often share the same algebraic implementation even though they encode different information. A common surrogate empirical-risk form is

$$\min_{\theta} \; \mathbb{E}_{(X,Y)\sim\widehat{P}_n} \left[\ell_\Omega(Y, f_\theta(X)) \right], \tag{22}$$

where Ω is a coefficient object. Depending on the method, Ω may be a class-prior weight vector, a diagonal gain matrix, a full pairwise cost matrix, or a cost-weighted adversarial objective (Vapnik, 1999; Dmochowski et al., 2010; Menon et al., 2021; Narasimhan and Menon, 2021; Chen et al., 2025).

A distributionally robust surrogate version is

$$\min_{\theta} \sup_{Q\in\mathcal{U}(\widehat{P}_n)} \mathbb{E}_{(X,Y)\sim Q} \left[\ell_\Omega(Y, f_\theta(X)) \right], \tag{23}$$

where $\mathcal{U}(\widehat{P}_n)$ is an ambiguity set around the empirical distribution. Standard ERM is the special case $\mathcal{U}(\widehat{P}_n) = \{\widehat{P}_n\}$. Adversarial training can be interpreted as optimizing over a neighborhood of perturbed examples, and distributionally robust optimization provides a general lens for objectives that minimize worst-case expected surrogate risk over such neighborhoods (Delage and Ye, 2010; Madry et al., 2018; Sinha et al., 2018; Lee and Raginsky, 2018; Rahimian and Mehrotra, 2019; Kuhn et al., 2019). This view does not make all methods equivalent; it identifies the common ingredients: an empirical distribution, an ambiguity set, and a surrogate loss.

For class-weighted cross-entropy,

$$\ell_{\mathrm{wCE}}(y, z) = a_y \ell_{\mathrm{CE}}(y, z), \tag{24}$$

where $a_y = g(\widehat{\pi}_y, n_y)$ is a frequency-derived class weight. If the dataset is balanced and $a_y \equiv a$ is constant, this is ordinary cross-entropy up to a positive scalar. Class-prior logit adjustment and Balanced Softmax similarly add terms such as $\tau \log \widehat{\pi}_y$ or $\log n_y$ to logits; these terms become constant under uniform empirical priors and cancel inside the softmax (Menon et al., 2021; Ren et al., 2020). Pairwise cost sensitivity does not vanish under balanced sampling because the entries $C_{y,k}$ remain nonuniform even when $\widehat{\pi}_y$ is uniform. This is the reason the main experiments use balanced splits to isolate the cost-matrix channel.

D Overparameterization and Terminal-Phase Collapse

The overparameterized regime is central to the motivation for NICME. Modern neural networks can fit training labels exactly, including settings with random labels or heavy overparameterization (Zhang et al., 2017; Arpit et al., 2017; Belkin et al., 2019; Bartlett et al., 2020). After the training error

reaches zero, optimization can continue to change margins and representations; this terminal phase is associated with implicit-bias and neural-collapse phenomena in separable deep models (Soudry et al., 2018; Papyan et al., 2020; Kini et al., 2021). For cost-sensitive learning, the relevant issue is not merely interpolation itself, but the inability of zero empirical cost alone to distinguish different cost-aware decision geometries.

The smooth empirical expected-cost surrogate is

$$\widehat{R}_C^{\mathrm{prob}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} C_{y_i,k}\, p_\theta(k \mid x_i). \tag{25}$$

If the model interpolates the training set in probability, then

$$p_\theta(y_i \mid x_i) \to 1, \qquad p_\theta(k \mid x_i) \to 0 \quad (k \neq y_i). \tag{26}$$

Since $C_{y_i,y_i} = 0$, the surrogate tends to zero independently of the off-diagonal entries of C. Let $h_\theta(x) \in \arg\max_{k\in[K]} f_{\theta,k}(x)$. The empirical discrete cost $n^{-1}\sum_i C_{y_i,h_\theta(x_i)}$ also vanishes. Thus, a cost-sensitive empirical objective evaluated only on perfectly fit training examples may not distinguish a cost-aware classifier from an accuracy-maximizing classifier (Chen et al., 2025).

Weighted CE has a related limitation. If $\ell_{\mathrm{CE}}(y_i, f_\theta(x_i)) \to 0$ for every training point, then for every finite positive scalar a_{y_i},

$$a_{y_i}\ell_{\mathrm{CE}}(y_i, f_\theta(x_i)) \to 0. \tag{27}$$

This does not imply that weights have no effect during all of training. It implies that after examples become separable, scalar weights do not directly specify which pairwise margins should keep increasing. Narasimhan and Menon (2021) demonstrate this issue for overparameterized ResNet models trained with reweighted cost-sensitive steps for non-decomposable objectives and propose logit-adjusted losses because the adjustment changes the required score geometry rather than only multiplying the loss value. Kini et al. (2021) formalize related behavior for label-imbalanced and group-sensitive classification under overparameterization, showing that terminal-phase dynamics can be critical for minority or group-sensitive performance.

CSADA provides a direct cost-sensitive DNN illustration. It compares conventional accuracy maximization with a training-set expected-cost objective and argues that, if a DNN can reveal the true label for every training example, then training-set cost becomes zero and the cost matrix no longer changes the empirical optimum (Chen et al., 2025). It also reports a pairwise MNIST experiment in which a baseline ResNet reaches perfect training accuracy and an extreme critical-pair penalty produces little change in pairwise error patterns, motivating boundary-local adversarial augmentation. NICME uses a different mechanism: it makes the cost matrix a condition on raw-logit separation, so low loss requires more than argmax correctness for costly wrong classes.

E NICME Derivations

E.1 Cross-Entropy on Adjusted Logits

For an example (x, y), let $z = f_\theta(x)$. Define

$$M_{y,k} = \max(C_{y,k}, 1), \tag{28}$$

and for scale $\alpha \geq 0$ construct

$$\bar{z}_k = z_k + \alpha(1 - \delta_{k,y}) \log M_{y,k}. \tag{29}$$

The target logit is unchanged, $\bar{z}_y = z_y$, while each non-target logit receives a cost-dependent offset. The pairwise cost-margin loss is ordinary cross-entropy on $\bar{z}$:

$$\ell_{\mathrm{pcm}}(y, z; C, \alpha) = \mathrm{CE}(y, \bar{z}) = -\log \mathrm{softmax}_y(\bar{z}) \tag{30}$$

$$= \log\left(1 + \sum_{k\neq y} \exp(z_k - z_y) M_{y,k}^{\alpha}\right). \tag{31}$$

When $\alpha = 0$, or when $M_{y,k} = 1$ for every $k \neq y$, this reduces to ordinary cross-entropy. This construction follows the logit-adjustment principle that class or cost information should change score geometry rather than only rescale the sample loss (Menon et al., 2021; Narasimhan and Menon, 2021).

E.2 Cost-Dependent Margins

Proposition 2 (Adjusted-odds margin). Fix a true label y, a wrong label $k \neq y$, $\alpha \geq 0$, and $M_{y,k} \geq 1$. Then

$$\frac{\exp(\bar{z}_k)}{\exp(\bar{z}_y)} = \exp(z_k - z_y) M_{y,k}^{\alpha}. \tag{32}$$

If this adjusted odds contribution is at most $\rho \in (0, 1]$, then

$$z_y - z_k \geq \alpha \log M_{y,k} + \log(1/\rho). \tag{33}$$

Proof. For $k \neq y$, $\bar{z}_k = z_k + \alpha \log M_{y,k}$ and $\bar{z}_y = z_y$. Taking logarithms and rearranging yields the result.

Proposition 3 (Small-loss margin). Fix a true label y, $\alpha \geq 0$, and $M_{y,k} \geq 1$ for every wrong label $k \neq y$. If $\ell_{\mathrm{pcm}}(y, z; C, \alpha) \leq \varepsilon$ for $\varepsilon > 0$, then for every $k \neq y$,

$$z_y - z_k \geq \alpha \log M_{y,k} - \log(e^{\varepsilon} - 1). \tag{34}$$

Proof. From the pairwise form of the loss,

$$1 + \sum_{k \neq y} \exp(z_k - z_y) M_{y,k}^{\alpha} \leq e^{\varepsilon}. \tag{35}$$

Each summand is nonnegative, so for every $k \neq y$,

$$\exp(z_k - z_y) M_{y,k}^{\alpha} \leq e^{\varepsilon} - 1, \tag{36}$$

and taking logarithms and rearranging proves the claim. Correct classification only requires $z_y > z_k$ for every $k \neq y$; a small NICME loss requires larger true-versus-wrong margins when $M_{y,k}$ is larger.

E.3 Normalized Expected-Cost Regularizer

Let

$$\widetilde{C} = \frac{C}{C_{\max} + \varepsilon_{\mathrm{num}}}, \qquad C_{\max} = \max_{a,b \in [K]} C_{a,b}, \tag{37}$$

where $\varepsilon_{\mathrm{num}} > 0$ prevents division by zero. The normalized expected-cost regularizer is

$$r_C(y, z) = \widetilde{C}_{y,\cdot}^{\top} \operatorname{softmax}(z) = \sum_{k=1}^{K} \widetilde{C}_{y,k} p_k, \qquad p = \operatorname{softmax}(z). \tag{38}$$

Since $\widetilde{C}_{y,k} \in [0, 1]$ and $\sum_k p_k = 1$, $r_C(y, z) \in [0, 1]$. The normalization makes λ comparable across datasets whose cost matrices have different maximum entries. Cost-sensitive expected-probability penalties of this type are related to smooth cost-sensitive surrogate risks and cost-sensitive regularization approaches (Dmochowski et al., 2010; Scott, 2012; Galdran et al., 2020; Chen et al., 2025).

Proposition 4 (Expected-cost gradient). For $r_C(y, z) = \sum_{k=1}^{K} \widetilde{C}_{y,k} p_k$ with $p = \operatorname{softmax}(z)$,

$$\frac{\partial r_C(y, z)}{\partial z_q} = p_q \left(\widetilde{C}_{y,q} - r_C(y, z) \right) = p_q \left(\widetilde{C}_{y,q} - \sum_{k=1}^{K} \widetilde{C}_{y,k} p_k \right). \tag{39}$$

Proof. Using $\partial p_k / \partial z_q = p_k(\delta_{k,q} - p_q)$,

$$\frac{\partial r_C}{\partial z_q} = \sum_{k=1}^{K} \widetilde{C}_{y,k} p_k (\delta_{k,q} - p_q) \tag{40}$$

$$= \widetilde{C}_{y,q} p_q - p_q \sum_{k=1}^{K} \widetilde{C}_{y,k} p_k = p_q (\widetilde{C}_{y,q} - r_C). \tag{41}$$

For gradient descent, a class q with $\widetilde{C}_{y,q} > r_C(y, z)$ receives a positive derivative and its logit is decreased. A class with $\widetilde{C}_{y,q} < r_C(y, z)$ receives a negative derivative and may gain relative probability mass. The cross-entropy term remains the label-fitting anchor, while the regularizer shapes residual probability mass under the original logits.

E.4 Final Objective

The single-example loss is

$$\ell_{\mathrm{NICME}}(y, z; C, \alpha, \lambda) = \ell_{\mathrm{pcm}}(y, z; C, \alpha) + \lambda r_C(y, z), \tag{42}$$

and the empirical objective is Eq. (12). The first term enforces cost-dependent score separation; the second penalizes expected cost under the original predictive distribution.

F Datasets and Cost-Matrix Construction

F.1 PMI-20

PMI-20 is the primary multiclass benchmark because it contains visually confusable medication classes and a pairwise medication-error cost matrix inherited from the PMI/CSADA protocol (Chen et al., 2025). The original PMI dataset is a subset of a larger pharmaceutical image collection with more than 300 medications, and the CSADA protocol selected 20 National Drug Code (NDC) classes on which a baseline model made the most mistakes. In that protocol, images were randomly split within each NDC into training, validation, and test partitions at approximately a 6:2:2 ratio, and expert costs were assigned to critical pairs after pharmacist review of baseline errors. Noncritical pairs were assigned unit cost in the original protocol, while critical pairs received larger directed costs based on potential danger (Chen et al., 2025).

The paper-facing NICME experiments use a balanced PMI-20 split prepared from the PMI pills data. The current split root records 1,940 training images, 320 validation images, 300 calibration images, and 640 test images, with four cared classes used for target-recall metrics. The main no-calibration protocol trains on the training split, selects using validation, and reports argmax decisions on the test split. The calibration partition is retained in the prepared data tree but is not used for the main argmax comparison.

F.2 Spider

The Spider task distinguishes black widow from false widow images. The paper-facing binary classes are class 0 black widow and class 1 false widow. The local two-class source pool contains 1,500 black-widow images and 1,499 false-widow images; an additional three-class source pool contains 1,500 images each for *Latrodectus hesperus*, *Steatoda grossa*, and *Steatoda nobilis*. The binary experiments use balanced stratified train/validation/calibration/test splits generated with a fixed random seed. The split has 2,098/300/300/300 train/validation/calibration/test images. The high-cost directed error is black widow $\rightarrow$ false widow.

The primary matrix is

$$C_{\mathrm{spider}} = \begin{bmatrix} 0 & 8 \\ 1 & 0 \end{bmatrix}, \tag{43}$$

with rows and columns ordered as black widow and false widow. The matrix is constructed as a decision-context matrix rather than a claimed universal toxicological utility. The construction has three steps.

First, public toxicology evidence establishes the high-cost direction. Monte et al. (2011) analyzed *Latrodectus* exposures reported to U.S. poison centers and describe black-widow envenomation as the most clinically significant spider envenomation in the United States. Williams et al. (2023) likewise summarize clinically significant *Latrodectus* envenomation and latrodectism. The Merck Manual Professional Edition (2025) lists treatment for severe or sometimes moderate widow bites as potentially involving parenteral opioids, benzodiazepines, and antivenom. These sources support treating black widow $\rightarrow$ false widow as a high-cost under-warning error.

Second, the lower-cost direction is not assigned zero. *Steatoda grossa* is a false-widow species that can be confused with true widows because it shares a rounded-abdomen body form and cobweb-making traits (UC Statewide IPM Program, 2017; Montana State University Extension, 2026). Graudins et al. (2002) report clinical and in-vitro evidence for medically meaningful *Steatoda grossa* envenomation and antivenom cross-reactivity, supporting a nonzero lower-direction harm. Thus, false widow

→ black widow is assigned the unit-baseline cost 1, reflecting precaution, anxiety, and possible unnecessary evaluation rather than no harm.

Third, the numeric ratio combines a harm-index estimate and a threshold-style action estimate. The harm-index scores are 12 for black widow → false widow and 2 for false widow → black widow, giving

$$R_{\text{harm}} = 12/2 = 6. \tag{44}$$

The threshold estimator is $R_{\text{threshold}} = 10$. The final primary ratio is the nearest integer to the geometric mean,

$$R_{\text{spider}} = \text{round}\left(\sqrt{R_{\text{harm}} R_{\text{threshold}}}\right) = \text{round}(\sqrt{60}) = 8. \tag{45}$$

The geometric mean treats harm severity and action threshold as multiplicative sources of evidence while avoiding dominance by either estimator. These ratios are transparent elicitation heuristics for sensitivity-audited experiments rather than externally validated toxicological utility estimates. The broad sensitivity audit evaluates $R \in \{1, 2, 5, 10, 20\}$, with $R = 1$ as a symmetric-cost control.

F.3 BreaKHis

BreaKHis is a breast histopathological image dataset for benign-versus-malignant classification (Spanhol et al., 2016). The dataset contains 7,909 RGB microscopy images from 82 patients, with 2,480 benign images from 24 patients and 5,429 malignant images from 58 patients. Images are available at four magnification factors: $40\times$, $100\times$, $200\times$, and $400\times$, and the standard binary task distinguishes benign from malignant tumors (Spanhol et al., 2016; UFPR, 2022). The label hierarchy includes four benign subtypes and four malignant subtypes. The NICME experiments use a balanced benign-versus-malignant binary setting to isolate asymmetric cost behavior from class-prior imbalance. The prepared split has 3,448/410/538/564 train/validation/calibration/test images. Patient identifiers are separated across partitions. The high-cost directed error is malignant → benign.

The primary matrix is

$$C_{\text{breakhis}} = \begin{bmatrix} 0 & 1 \\ 7 & 0 \end{bmatrix}, \tag{46}$$

with rows and columns ordered as benign and malignant. It is a decision-context matrix for histopathology decision support, not a universal clinical utility.

First, the high-cost direction follows the clinical semantics of the task. A malignant tumor is cancerous and can invade adjacent structures or metastasize, while benign lesions are comparatively localized in the BreaKHis task definition (Spanhol et al., 2016; UFPR, 2022). A malignant → benign error can delay additional review, diagnosis, or treatment in a decision-support setting. NCI breast-screening materials describe false negatives as a source of false reassurance and possible delay in cancer diagnosis, and interval cancers as a clinically important category that may be more aggressive than screen-detected cancers (National Cancer Institute, 2025; 2011). Although screening mammography is not identical to histopathology classification, these sources provide public clinical evidence for the directionality of missed-malignancy harm.

Second, the lower-cost direction is not assigned zero. A benign → malignant error can generate anxiety, additional workup, and possible biopsy or review burden. NCI materials summarize false positives, additional diagnostic procedures, biopsies, overdiagnosis, and psychological harms as real harms of cancer screening and diagnostic pathways (National Cancer Institute, 2026; 2025). The USPSTF harms review similarly summarizes false-positive and false-negative results, additional imaging, biopsies, overdiagnosis, anxiety, pain, and radiation exposure (Nelson et al., 2016). Therefore, benign → malignant is assigned the unit-baseline cost 1 rather than zero.

Third, the numeric ratio combines a harm-index estimate and an action-threshold estimate. The harm-index scores are 18 for malignant → benign and 3 for benign → malignant, giving $R_{\text{harm}} = 18/3 = 6$. The threshold estimator uses a 10% action threshold for possible malignancy,

$$R_{\text{threshold}} = \frac{1 - 0.10}{0.10} = 9. \tag{47}$$

The final primary ratio is

$$R_{\text{breakhis}} = \text{round}(\sqrt{6 \cdot 9}) = \text{round}(\sqrt{54}) = 7. \tag{48}$$

These ratios are transparent elicitation heuristics for sensitivity-audited experiments rather than externally validated clinical utility estimates. Pathology-specific diagnostic-variability evidence further supports the relevance of decision support in this domain (Elmore et al., 2015). The sensitivity audit again evaluates $R \in \{1, 2, 5, 10, 20\}$, with $R = 1$ as a symmetric-cost control.

F.4 Asset Licenses and Access Conditions

We document the main third-party assets used in the experiments so that asset provenance is tied to explicit license or access information rather than only to citations. For BreaKHis, we cite both the original dataset paper and the official UFPR dataset page; the official page describes the database and its citation and access requirements (Spanhol et al., 2016; UFPR, 2022). Our experiments use BreaKHis for supervised training and evaluation and do not redistribute the raw microscopy images.

For the PMI-20 experiments, the balanced PMI-20 split is a paper-facing split prepared from PMI pills imagery following the task definition and pairwise-cost protocol of Chen et al. (2025). Because the current submission does not own the underlying pill imagery and does not introduce a new public PMI image release, we treat PMI-20 as a non-redistributed derived benchmark: the accompanying reproducibility package provides split manifests, cost matrices, configurations, and scripts, while requiring users to obtain the underlying images from the original PMI source rather than from this paper.

For Spider, the paper-facing benchmark is a local research image pool assembled for this study from externally hosted spider photographs. We do not claim copyright over those raw images, do not assign them a new open license, and do not redistribute them with the submission. Any later release should therefore be limited to derived artifacts such as split manifests, class labels, cost matrices, and instructions, with the raw images remaining subject to the copyright and access conditions of their original host sources.

For software and pretrained models, the implementation uses the `timm` codebase and the ConvNeXt-Base checkpoint `convnext_base.fb_in22k_ft_in1k` (Liu et al., 2022; Wightman, 2019). The `timm` project documents its code as Apache 2.0 licensed, and its license notes further state that ImageNet-pretrained weights should be treated as subject to the original ImageNet research-use conditions; accordingly, we use these components only as cited research dependencies and do not relicense or redistribute third-party pretrained weights in the submission package.

G Baselines and Metrics

The comparisons cover CE, scalar cost weighting, expected-cost regularization, Menon prior adjustment, CSADA-inspired adaptation, a general gain-matrix hybrid, and NICME. Menon adjustment uses class priors rather than a full pairwise matrix (Menon et al., 2021). The gain-matrix loss provides a trainable non-diagonal comparator (Narasimhan and Menon, 2021). Calibrated CE with minimum expected-cost decisions is included directly in the main comparisons (Elkan, 2001; Zadrozny and Elkan, 2001; Guo et al., 2017). Historical expected-cost adaptation is distinct from the controlled $(\alpha, \lambda) = (0, 0.1)$ ablation; Appendix I records that distinction.

Let $T \subseteq [K]$ denote the nonempty set of cared or target classes, and let $\widehat{y}_i$ denote the predicted test label for example i. Recall for class c is

$$\mathrm{Recall}(c) = \frac{\#\{i : y_i = c, \widehat{y}_i = c\}}{\#\{i : y_i = c\}}. \tag{49}$$

Target-min recall and target-macro recall are

$$\mathrm{TargetMin} = \min_{c \in T} \mathrm{Recall}(c), \qquad \mathrm{TargetMacro} = \frac{1}{|T|} \sum_{c \in T} \mathrm{Recall}(c). \tag{50}$$

Target-min recall measures worst-case cared-class recall and is used because cost-sensitive and non-decomposable objectives often aim to protect the weakest cared class rather than only average performance (Narasimhan and Menon, 2021). Target-macro recall measures average cared-class recall and exposes tradeoffs where the worst cared class improves while others degrade (Brodersen et al., 2010; Narasimhan and Menon, 2021).

Average test cost is

$$\mathrm{ATC} = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} C_{y_i,\widehat{y}_i}, \tag{51}$$

and normalized ATC is

$$\mathrm{nATC} = \frac{\mathrm{ATC}}{C_{\max}}, \qquad C_{\max} > 0. \tag{52}$$

Normalization expresses incurred cost relative to the maximum entry of the evaluation matrix. It does not make different application utilities interchangeable. For an all-zero matrix the implementation returns ATC, which is zero. Critical-pair errors count predictions belonging to a predeclared set of high-cost directed confusions; this metric is useful when a small number of rare but important pairwise errors carries most of the application cost (Chen et al., 2025). Balanced accuracy is mean recall over all classes, and macro-F1 is the unweighted mean of per-class F1 scores. These utility metrics are reported because a cost-sensitive method should not appear strong solely by overpredicting cared classes or collapsing coverage (Brodersen et al., 2010; He and Garcia, 2009; Johnson and Khoshgoftaar, 2019).

H Training, Implementation, and Experimental Protocol

H.1 Mini-Batch Procedure and Reproduction

For each mini-batch, compute raw logits, gather the true-label rows of C, add the wrong-class offsets in Eq. (6), and apply CE to the adjusted logits. Compute the expected-cost term from softmax of the original logits, then optimize their weighted sum. The implementation dispatches `nicme_logit_adjustment` to the pairwise margin loss and `nicme_hybrid` to the margin-plus-regularizer loss. Legacy loss variants are separate implementations and are not substituted for these dispatch names. A two-epoch linear warmup is applied to the regularizer where specified in the supplied configurations. Primary inference uses the original logits without the true label or any label-conditioned cost row.

The anonymous supplement contains the implementation, training entry points, experiment runners, effective configurations, prepared split manifests, cost matrices, and result summaries. Running `python reproduce_tables.py` uses bundled numeric artifacts to recompute the additional calibration analysis, validate aggregates, reproduce the paired statistics, and regenerate the new manuscript tables. This CPU path requires no model downloads or raw images. Training reproduction additionally requires the original datasets and the pretrained backbone; acquisition and path-resolution instructions are included. Images and pretrained weights are not included in the archive.

H.2 Distinct Experimental Protocols

Retained-model comparisons. The historical PMI-20 reference uses $(\alpha, \lambda) = (0.5, 0.1)$ and LR 5×10^{-5}, with 224-pixel inputs, batch 16, accumulation 4, cosine scheduling, weight decay 0.005, and warmup ratio 0.10. The direct-training ceiling is 32 epochs with patience 5. Some historical baselines adapt a CE checkpoint with a shorter schedule. The original run labels are 42–44, but some recorded Trainer seeds remained 42. We therefore report three retained runs, rather than treating those checkpoint differences as verified independent training seeds. The controlled experiments below separately align the declared seed, Trainer seed, and data seed.

Eight-seed component panel. The 40 models comprise CE, Menon adjustment, margin only $(0.5, 0)$, regularizer only $(0, 0.1)$, and full NICME $(0.5, 0.1)$ over seeds 42–49. All use the fixed balanced PMI-20 split and the shared direct-training envelope stated in the main text. The selection metric is validation target-min recall; ties retain the earliest checkpoint. Initialization specifications are shared, but identical realized initial weight tensors were not independently hashed. The shared fixed split and seed correspondence support paired training-run analysis; they do not supply independent dataset replications.

Matched search. Each of CE, NICME, regularized CE, and the gain-matrix hybrid receives nine validation-only candidates. CE and the gain hybrid search LR $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$ and

weight decay $\{0.001, 0.005, 0.01\}$. NICME searches $\alpha \in \{0.3, 0.5, 0.7\}$ and $\lambda \in \{0.05, 0.1, 0.2\}$. Regularized CE searches the same LR set and coefficients $\{0.1, 1, 5\}$. Candidates meeting target-min recall 0.90 and BA 0.80 are ranked by validation nATC, followed by target-macro recall, BA, critical errors, elapsed time, and configuration hash. The supplied search specification defines the fallback when no candidate is feasible. Selected identities are frozen before final evaluation over seeds 42–44. Thus configuration counts and training envelopes are matched, but the searched dimensions, checkpoint criteria, and realized budgets are not identical. Search training times are approximately 1.060, 0.910, 1.034, and 1.275 GPU-hours for CE, NICME, regularized CE, and the gain hybrid, respectively.

The gain construction uses $G = B\mathbf{1}\mathbf{1}^\top - C$ with the configured $B = C_{\max}$, rescales each column by its diagonal gain, and applies the general hybrid objective of Narasimhan and Menon (2021). We retain score argmax for this comparator; its scores are not treated as calibrated class probabilities. The complete configurations and all final rows are provided in the supplement.

Compute. The studies use a single NVIDIA RTX 5090 with 32,607 MiB visible memory. The corrected eight-seed panel used approximately 4.67 training GPU-hours. The completed CIFAR-100 direction/strength extension used approximately 14.74 fresh-training GPU-hours; reused model evaluations do not count as additional independent trainings. These figures describe the named campaigns rather than a total including all exploratory work. A dense 100-class float32 matrix occupies 40,000 bytes. Recorded CIFAR-100 peak allocated memory is about 8,590 MiB, and inference measurements are about 1.56–1.63 ms per example under the recorded evaluation setup. Exact run-level measurements are included in the supplement.

I Historical Comparisons and Method Identity

These tables preserve the historical trained-model operating points with explicit method identities. "Expected-cost CE adaptation" uses coefficient 5, a CE initialization, LR 10^{-5}, a 10-epoch ceiling, and patience 3. It is not the $(\alpha, \lambda) = (0, 0.1)$ component ablation, which is evaluated separately in Table 3. The CSADA-inspired implementation uses targeted adaptation toward a selected costly class rather than reproducing every part of the original published CSADA procedure. Its configured adaptation uses five attack steps, adversarial weight 2, and cost temperature 3; the supplement contains the exact perturbation and optimizer settings. Cost-weighted CE uses fixed normalized row-mean costs and the common direct-training setting, without a separate cost-weight grid.

Table 5: Historical PMI-20 trained-model comparison. Raw argmax, mean ± sample SD over three retained runs. Target recall here is target-macro; adaptation protocols differ from direct training.

Method	nATC	ATC	BA	Target-macro
Expected-cost CE adaptation	0.00505 ± 0.00181	0.05052 ± 0.01811	0.9802 ± 0.0033	0.9661 ± 0.0090
Menon adjustment	0.00552 ± 0.00317	0.05521 ± 0.03165	0.9729 ± 0.0185	0.9688 ± 0.0078
CE	0.00568 ± 0.00310	0.05677 ± 0.03103	0.9714 ± 0.0174	0.9688 ± 0.0078
Cost-weighted CE	0.00734 ± 0.00149	0.07344 ± 0.01491	0.9547 ± 0.0149	0.9714 ± 0.0045
CSADA-inspired	0.02526 ± 0.00610	0.25260 ± 0.06104	0.7755 ± 0.0577	0.9714 ± 0.0045
NICME	0.00370 ± 0.00095	0.03698 ± 0.00955	0.9771 ± 0.0106	0.9740 ± 0.0090

Table 6: Historical Spider trained-model comparison. Raw argmax, mean ± sample SD over three retained runs. Target recall here is target-macro; adaptation protocols differ from direct training.

Method	nATC	ATC	BA	Target-macro
CSADA-inspired	0.04222 ± 0.01087	0.33778 ± 0.08694	0.6622 ± 0.0869	1.0000 ± 0.0000
Cost-weighted CE	0.06139 ± 0.00192	0.49111 ± 0.01540	0.5089 ± 0.0154	1.0000 ± 0.0000
NICME	0.01875 ± 0.00439	0.15000 ± 0.03512	0.9356 ± 0.0158	0.9756 ± 0.0139
CE	0.02056 ± 0.00792	0.16444 ± 0.06336	0.9367 ± 0.0100	0.9711 ± 0.0168
Expected-cost CE adaptation	0.02236 ± 0.00536	0.17889 ± 0.04286	0.9533 ± 0.0100	0.9622 ± 0.0102
Menon adjustment	0.02569 ± 0.00421	0.20556 ± 0.03372	0.9267 ± 0.0100	0.9622 ± 0.0102

Table 7: Historical Breakhis trained-model comparison. Raw argmax, mean ± sample SD over three retained runs. Target recall here is target-macro; adaptation protocols differ from direct training.

Method	nATC	ATC	BA	Target-macro
CSADA-inspired	0.07084 ± 0.00102	0.49586 ± 0.00717	0.5041 ± 0.0072	1.0000 ± 0.0000
NICME	0.03259 ± 0.00919	0.22813 ± 0.06435	0.8641 ± 0.0461	0.9693 ± 0.0364
Cost-weighted CE	0.05226 ± 0.00872	0.36584 ± 0.06107	0.7370 ± 0.0293	0.9657 ± 0.0143
Expected-cost CE adaptation	0.04534 ± 0.01607	0.31738 ± 0.11252	0.8848 ± 0.0116	0.9326 ± 0.0338
CE	0.05902 ± 0.02736	0.41312 ± 0.19151	0.8954 ± 0.0251	0.8972 ± 0.0662
Menon adjustment	0.05902 ± 0.02736	0.41312 ± 0.19151	0.8954 ± 0.0251	0.8972 ± 0.0662

J Calibration and Minimum-Expected-Cost Decisions

Following Guo et al. (2017), one scalar temperature is fitted to minimize NLL on each held-out calibration partition. The deterministic 79-point grid has spacing 0.05 over $[0.5, 3.0]$ and spacing 0.25 over $[3.25, 10.0]$, with $T = 1$ retained. Binary affine Platt scaling is fitted to $z_1 - z_0$ with L-BFGS-B using calibration labels only. Checkpoint inference for the retained-model analysis uses float32 without autocast. Calibration and decision rules are not selected using test labels.

The Brier score is $N^{-1} \sum_i \sum_k (p_{ik} - \mathbf{1}[y_i = k])^2$; in a binary task it is twice the scalar positive-class convention. ECE uses the top prediction and confidence in 15 equal-width bins, while NLL uses the true-label probability. Temperature scaling preserves raw argmax but can change minimum-cost decisions. Affine binary scaling can also change argmax. We therefore report probability quality and decisions separately. A favorable ECE value alone does not establish calibration for every subgroup or event.

Table 8: Probability calibration on PMI-20. Three retained runs; mean ± sample SD. ECE uses 15 equal-width top-label bins; Brier is summed across classes.

Method	Scaling	ECE	NLL	Brier
CE	raw	0.02417 ± 0.01446	0.10550 ± 0.04101	0.04659 ± 0.02184
CE	temperature	0.01420 ± 0.00341	0.09698 ± 0.03323	0.04580 ± 0.02048
NICME	raw	0.01737 ± 0.00812	0.07928 ± 0.02334	0.03356 ± 0.01390
NICME	temperature	0.01679 ± 0.00690	0.07933 ± 0.01793	0.03418 ± 0.01387
EC-CE adaptation	raw	0.03784 ± 0.00900	0.12269 ± 0.01740	0.05472 ± 0.00848
EC-CE adaptation	temperature	0.01462 ± 0.00274	0.10776 ± 0.01378	0.05285 ± 0.00737

Table 9: Probability calibration on Spider. Three retained runs; mean ± sample SD. ECE uses 15 equal-width top-label bins; Brier is summed across classes.

Method	Scaling	ECE	NLL	Brier
CE	platt	0.03320 ± 0.00545	0.14784 ± 0.00910	0.08223 ± 0.00789
CE	raw	0.04579 ± 0.00244	0.26621 ± 0.07999	0.09952 ± 0.01287
CE	temperature	0.02379 ± 0.00781	0.17289 ± 0.01851	0.09432 ± 0.01194
NICME	platt	0.03769 ± 0.00640	0.15163 ± 0.02134	0.08520 ± 0.01506
NICME	raw	0.05748 ± 0.01437	0.31926 ± 0.16824	0.11066 ± 0.02906
NICME	temperature	0.02847 ± 0.00324	0.17468 ± 0.04049	0.09765 ± 0.02297

Table 10: Probability calibration on BreaKHis. Three retained runs; mean ± sample SD. ECE uses 15 equal-width top-label bins; Brier is summed across classes.

Method	Scaling	ECE	NLL	Brier
CE	platt	0.04157 ± 0.01037	0.26468 ± 0.05675	0.14829 ± 0.03462
CE	raw	0.08863 ± 0.00704	0.31188 ± 0.03892	0.18115 ± 0.02956
CE	temperature	0.03999 ± 0.01060	0.27127 ± 0.05273	0.16208 ± 0.03181
NICME	platt	0.04619 ± 0.02669	0.23598 ± 0.01662	0.13817 ± 0.02296
NICME	raw	0.12117 ± 0.05054	0.69693 ± 0.59493	0.24714 ± 0.09851
NICME	temperature	0.04267 ± 0.01692	0.32833 ± 0.10196	0.20621 ± 0.07038

Table 11: Complete retained-run decision comparison on pmi20. Mean ± sample SD over three runs.

Method	Scaling	Rule	nATC	BA	Target-min
CE	raw	argmax	0.00568 ± 0.00310	0.9714 ± 0.0174	0.9167 ± 0.0180
CE	raw	cost-min	0.00568 ± 0.00164	0.9526 ± 0.0234	0.8333 ± 0.0477
CE	temp.	argmax	0.00568 ± 0.00310	0.9714 ± 0.0174	0.9167 ± 0.0180
CE	temp.	cost-min	0.00547 ± 0.00118	0.9594 ± 0.0217	0.8854 ± 0.0180
NICME	raw	argmax	0.00370 ± 0.00095	0.9771 ± 0.0106	0.9167 ± 0.0180
NICME	raw	cost-min	0.00313 ± 0.00072	0.9734 ± 0.0143	0.9167 ± 0.0180
NICME	temp.	argmax	0.00370 ± 0.00095	0.9771 ± 0.0106	0.9167 ± 0.0180
NICME	temp.	cost-min	0.00297 ± 0.00056	0.9750 ± 0.0128	0.9167 ± 0.0180
EC-CE adaptation	raw	argmax	0.00661 ± 0.00133	0.9667 ± 0.0055	0.8958 ± 0.0180
EC-CE adaptation	raw	cost-min	0.00620 ± 0.00106	0.9474 ± 0.0119	0.8333 ± 0.0477
EC-CE adaptation	temp.	argmax	0.00661 ± 0.00133	0.9667 ± 0.0055	0.8958 ± 0.0180
EC-CE adaptation	temp.	cost-min	0.00583 ± 0.00065	0.9557 ± 0.0104	0.8750 ± 0.0313

Table 12: Complete retained-run decision comparison on spider. Mean ± sample SD over three runs.

Method	Scaling	Rule	nATC	BA	Target-min
CE	platt	argmax	0.02181 ± 0.00670	0.9500 ± 0.0067	0.9644 ± 0.0139
CE	platt	cost-min	0.01375 ± 0.00451	0.9056 ± 0.0234	0.9956 ± 0.0038
CE	raw	argmax	0.02056 ± 0.00792	0.9367 ± 0.0100	0.9711 ± 0.0168
CE	raw	cost-min	0.01333 ± 0.00289	0.9167 ± 0.0231	0.9933 ± 0.0000
CE	temp.	argmax	0.02056 ± 0.00792	0.9367 ± 0.0100	0.9711 ± 0.0168
CE	temp.	cost-min	0.02014 ± 0.00479	0.8544 ± 0.0278	0.9956 ± 0.0038
NICME	platt	argmax	0.02208 ± 0.00737	0.9400 ± 0.0115	0.9667 ± 0.0176
NICME	platt	cost-min	0.01486 ± 0.00173	0.9122 ± 0.0164	0.9911 ± 0.0038
NICME	raw	argmax	0.01875 ± 0.00439	0.9356 ± 0.0158	0.9756 ± 0.0139
NICME	raw	cost-min	0.01375 ± 0.00125	0.9289 ± 0.0190	0.9889 ± 0.0077
NICME	temp.	argmax	0.01875 ± 0.00439	0.9356 ± 0.0158	0.9756 ± 0.0139
NICME	temp.	cost-min	0.01764 ± 0.00457	0.8822 ± 0.0434	0.9933 ± 0.0067

Table 13: Complete retained-run decision comparison on breakhis. Mean ± sample SD over three runs.

Method	Scaling	Rule	nATC	BA	Target-min
CE	platt	argmax	0.06966 ± 0.01415	0.9060 ± 0.0246	0.8688 ± 0.0248
CE	platt	cost-min	0.04728 ± 0.01305	0.8676 ± 0.0421	0.9338 ± 0.0168
CE	raw	argmax	0.05902 ± 0.02736	0.8954 ± 0.0251	0.8972 ± 0.0662
CE	raw	cost-min	0.04078 ± 0.01352	0.7465 ± 0.0706	0.9894 ± 0.0123
CE	temp.	argmax	0.05902 ± 0.02736	0.8954 ± 0.0251	0.8972 ± 0.0662
CE	temp.	cost-min	0.04601 ± 0.01428	0.8162 ± 0.0541	0.9539 ± 0.0409
NICME	platt	argmax	0.04593 ± 0.01064	0.9054 ± 0.0302	0.9243 ± 0.0341
NICME	platt	cost-min	0.03664 ± 0.00944	0.8676 ± 0.0238	0.9586 ± 0.0297
NICME	raw	argmax	0.03259 ± 0.00919	0.8641 ± 0.0461	0.9693 ± 0.0364
NICME	raw	cost-min	0.03842 ± 0.01435	0.7417 ± 0.1005	0.9965 ± 0.0035
NICME	temp.	argmax	0.03259 ± 0.00919	0.8641 ± 0.0461	0.9693 ± 0.0364
NICME	temp.	cost-min	0.04998 ± 0.01156	0.6856 ± 0.1182	0.9882 ± 0.0175

K Controlled Components, Paired Statistics, and Saved-Logit Extension

The two predeclared tests compare full NICME against CE and against the true regularizer-only component on nATC. Paired effects are NICME minus comparator, so negative values favor NICME. The eight aligned seeds are retained, including the higher-cost full-NICME result at seed 44. Each exact two-sided sign-flip test enumerates all $2^8 = 256$ sign assignments using absolute mean paired difference, with sign symmetry as the null reference assumption. Holm adjustment applies to the two-test family. Paired percentile-bootstrap intervals use 100,000 shared resamples with RNG seed 20260728.

Table 14: Predeclared paired nATC effects over eight seeds on one fixed split. Intervals quantify variation across the observed training runs, not across independently sampled datasets.

Comparator	Mean difference	95% bootstrap interval	Exact p	Holm p
CE	$+0.000195$	$[-0.004648, 0.006621]$	0.9765625	1.0
Regularizer only	-0.000039	$[-0.006406, 0.006777]$	0.9921875	1.0

An additional exploratory analysis fits temperatures to the saved calibration logits of all 40 models and evaluates the corresponding test logits. It introduces no new model training or model selection. Tables 15 and 16 show raw argmax and temperature-scaled cost-min endpoints. The raw metrics reproduce the original component panel. With calibrated cost-min decisions, mean nATC is 0.005313 for CE, 0.004004 for margin only, and 0.006523 for full NICME. These results further support studying the directed margin and calibration jointly, while retaining the distinction between component evidence and a superiority claim for the fixed full objective. All probability metrics, temperatures, and exploratory paired effects are included as machine-readable files. The exploratory intervals use 100,000 paired bootstrap resamples with RNG seed 20260925 and do not expand the predeclared test family.

Table 15: Exploratory eight-seed reanalysis from saved logits: raw argmax. Mean $\pm$ sample SD; each temperature is fitted only on the calibration partition.

Objective	nATC	BA	ECE	NLL
CE	0.00703 ± 0.00422	0.9631 ± 0.0140	0.0282 ± 0.0182	0.1170 ± 0.0393
NICME $(.5, .1)$	0.00723 ± 0.00693	0.9490 ± 0.0571	0.0234 ± 0.0204	0.1488 ± 0.1418
Margin only $(.5, 0)$	0.00410 ± 0.00110	0.9766 ± 0.0088	0.0145 ± 0.0053	0.0824 ± 0.0264
Regularizer only $(0, .1)$	0.00727 ± 0.00505	0.9604 ± 0.0272	0.0389 ± 0.0370	0.1368 ± 0.0966
Menon adjustment	0.00580 ± 0.00228	0.9672 ± 0.0129	0.0212 ± 0.0110	0.1028 ± 0.0327

Table 16: Exploratory eight-seed reanalysis from saved logits: temperature-scaled minimum-cost decisions. Mean $\pm$ sample SD; each temperature is fitted only on the calibration partition.

Objective	nATC	BA	ECE	NLL
CE	0.00531 ± 0.00273	0.9535 ± 0.0200	0.0147 ± 0.0046	0.1086 ± 0.0301
NICME $(.5, .1)$	0.00652 ± 0.00649	0.9400 ± 0.0658	0.0211 ± 0.0178	0.1442 ± 0.1288
Margin only $(.5, 0)$	0.00400 ± 0.00161	0.9723 ± 0.0114	0.0112 ± 0.0034	0.0827 ± 0.0254
Regularizer only $(0, .1)$	0.00547 ± 0.00306	0.9523 ± 0.0321	0.0152 ± 0.0058	0.1189 ± 0.0668
Menon adjustment	0.00457 ± 0.00214	0.9596 ± 0.0181	0.0145 ± 0.0040	0.0993 ± 0.0275

L Dense CIFAR-100 Matrices and Complete Results

For fine classes y, k, the down-direction matrix is

$$C^{\downarrow,r}_{y,k} = \begin{cases} 0, & y = k, \\ 1, & y \neq k,\ s(y) = s(k), \\ 4r, & s(y) > s(k), \\ 4, & s(y) < s(k). \end{cases} \tag{53}$$

The up-direction matrix interchanges the final two cases. The ordering is the official coarse-label index, provided with the exact fine-to-coarse mapping. Each matrix is dense off the diagonal. Its largest entry is $4r$, so nATC comparisons are interpreted within the same matrix. A symmetric control with all cross-superclass costs equal to four is also included in the evidence files: mean raw-argmax nATC is 0.044733 for CE and 0.044225 for NICME.

The stratified split seed is 20260728. CE and NICME share the ConvNeXt-Base architecture and data partitions. NICME uses $\alpha = 0.5$, $\lambda = 0.1$, a two-epoch regularizer warmup, a 16-epoch ceiling, patience 3, and validation nATC checkpoint selection subject to BA ≥ 0.70. The detailed optimizer and augmentation settings are supplied as effective configurations. The six-matrix extension comprises 15 fresh NICME models, three reused up-$r = 2$ NICME models, and three CE anchors. The same anchors are rescored across matrices; 21 models are not interpreted as 36 independent trainings.

Each frozen model is evaluated under all six asymmetric matrices, yielding 252 model/matrix/decision rows for two decision rules. Main-text results are the diagonal evaluations, where NICME is evaluated under its training matrix. Cross-matrix results in the supplement show how a fixed trained model responds when the evaluation cost changes. Frozen identities and selection records precede test evaluation. Full confusion matrices permit independent reconstruction of cost, recognition, and directional error counts.

Table 17: Complete upward CIFAR-100 diagonal evaluation. Mean ± sample SD over three paired seeds.

r	Method	Decision	nATC	BA	High-dir. errors
2	CE	argmax	0.03090 ± 0.00214	0.9224 ± 0.0055	172.3 ± 13.2
2	CE	T cost-min	0.03007 ± 0.00253	0.9196 ± 0.0066	127.0 ± 10.4
4	CE	argmax	0.02406 ± 0.00165	0.9224 ± 0.0055	172.3 ± 13.2
4	CE	T cost-min	0.02007 ± 0.00167	0.9152 ± 0.0069	91.0 ± 8.7
8	CE	argmax	0.02065 ± 0.00146	0.9224 ± 0.0055	172.3 ± 13.2
8	CE	T cost-min	0.01351 ± 0.00104	0.9063 ± 0.0083	63.0 ± 2.6
2	NICME	argmax	0.03169 ± 0.00194	0.9197 ± 0.0072	178.7 ± 7.5
2	NICME	T cost-min	0.03048 ± 0.00172	0.9172 ± 0.0071	130.0 ± 3.0
4	NICME	argmax	0.02352 ± 0.00065	0.9251 ± 0.0034	168.3 ± 3.1
4	NICME	T cost-min	0.01925 ± 0.00004	0.9193 ± 0.0036	89.7 ± 5.9
8	NICME	argmax	0.01996 ± 0.00059	0.9245 ± 0.0015	165.7 ± 6.5
8	NICME	T cost-min	0.01391 ± 0.00019	0.9085 ± 0.0054	69.3 ± 7.6

Table 18: Complete downward CIFAR-100 diagonal evaluation. Mean ± sample SD over three paired seeds.

r	Method	Decision	nATC	BA	High-dir. errors
2	CE	argmax	0.03043 ± 0.00251	0.9224 ± 0.0055	163.0 ± 18.4
2	CE	T cost-min	0.02983 ± 0.00223	0.9194 ± 0.0057	117.7 ± 10.0
4	CE	argmax	0.02336 ± 0.00215	0.9224 ± 0.0055	163.0 ± 18.4
4	CE	T cost-min	0.02008 ± 0.00147	0.9140 ± 0.0065	86.0 ± 6.0
8	CE	argmax	0.01983 ± 0.00199	0.9224 ± 0.0055	163.0 ± 18.4
8	CE	T cost-min	0.01351 ± 0.00088	0.9045 ± 0.0072	60.7 ± 3.1
2	NICME	argmax	0.02936 ± 0.00269	0.9229 ± 0.0079	147.7 ± 13.4
2	NICME	T cost-min	0.02883 ± 0.00248	0.9200 ± 0.0080	105.3 ± 7.1
4	NICME	argmax	0.02122 ± 0.00039	0.9246 ± 0.0014	137.0 ± 3.6
4	NICME	T cost-min	0.01854 ± 0.00061	0.9165 ± 0.0026	70.3 ± 2.1
8	NICME	argmax	0.01768 ± 0.00044	0.9149 ± 0.0092	133.0 ± 7.5
8	NICME	T cost-min	0.01304 ± 0.00093	0.8945 ± 0.0104	46.7 ± 9.5

M HYPERPARAMETERS, COST SCALE, AND COST UNCERTAINTY

The historical 49-point grid is a validation sensitivity study. It contains 23 distinct metric vectors and 45 points with target-min recall 1.0. The $(0.5, 0.1)$ reference was selected before that grid and is highlighted as a reference point, not as its optimum. Exact rankings within tied or near-tied feasible regions should not be interpreted as unique prescriptions.

The local panel uses $\alpha \in \{0.3, 0.5, 0.7\}$ and $\lambda \in \{0.05, 0.1, 0.2\}$ with three seeds each. All nine mean validation results satisfy target-min recall 0.90 and BA 0.80. Seed-pair ranking correlations are

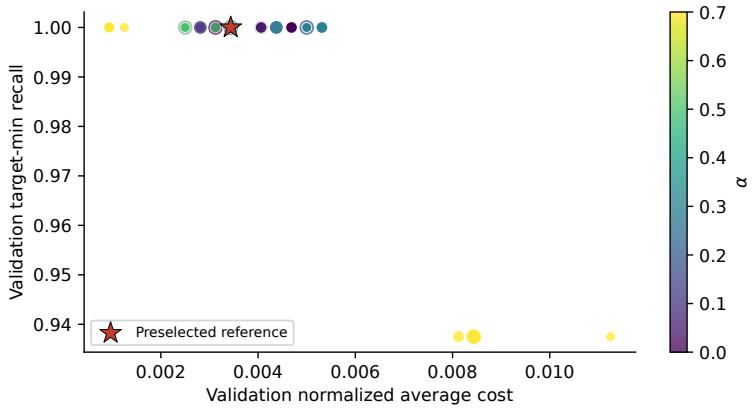

Figure 2: Historical validation grid. Color indicates α and point size indicates λ. The star marks the preselected $(0.5, 0.1)$ reference. Test results are not used to rank the plotted configurations.

-0.4430, -0.4418, and 0.0261. We report both validation and test results to distinguish feasibility from stable fine-grained ordering. A practical procedure is to declare cost units, the unit baseline, validation utility floors, and the selection criterion before fitting; then retain a feasible operating point under those criteria.

Table 19: Local alpha lambda on PMI-20 (validation, canonical evaluation matrix). Mean $\pm$ sample SD over three seeds.

Setting	nATC	BA	Target-min
α=0.3,λ=0.05	0.00677 ± 0.00308	0.9417 ± 0.0148	0.9583 ± 0.0361
α=0.3,λ=0.1	0.00667 ± 0.00262	0.9427 ± 0.0100	0.9583 ± 0.0361
α=0.3,λ=0.2	0.00615 ± 0.00255	0.9479 ± 0.0095	0.9583 ± 0.0361
α=0.5,λ=0.05	0.00906 ± 0.00922	0.9188 ± 0.0760	0.9375 ± 0.0625
α=0.5,λ=0.1	0.00844 ± 0.00976	0.9250 ± 0.0814	0.9583 ± 0.0722
α=0.5,λ=0.2	0.00875 ± 0.00950	0.9219 ± 0.0789	0.9375 ± 0.0625
α=0.7,λ=0.05	0.00896 ± 0.01081	0.9198 ± 0.0921	0.9167 ± 0.0722
α=0.7,λ=0.1	0.01042 ± 0.00989	0.9146 ± 0.0874	0.9167 ± 0.0361
α=0.7,λ=0.2	0.00906 ± 0.01069	0.9188 ± 0.0909	0.9167 ± 0.0722

Table 20: Local alpha lambda on PMI-20 (test, canonical evaluation matrix). Mean $\pm$ sample SD over three seeds.

Setting	nATC	BA	Target-min
α=0.3,λ=0.05	0.00880 ± 0.00166	0.9438 ± 0.0124	0.9063 ± 0.0313
α=0.3,λ=0.1	0.00865 ± 0.00197	0.9453 ± 0.0095	0.9063 ± 0.0313
α=0.3,λ=0.2	0.00672 ± 0.00230	0.9505 ± 0.0152	0.9271 ± 0.0180
α=0.5,λ=0.05	0.01146 ± 0.00970	0.9182 ± 0.0889	0.8750 ± 0.0541
α=0.5,λ=0.1	0.01120 ± 0.01011	0.9198 ± 0.0916	0.8854 ± 0.0651
α=0.5,λ=0.2	0.01161 ± 0.00970	0.9167 ± 0.0889	0.8750 ± 0.0541
α=0.7,λ=0.05	0.01036 ± 0.00899	0.9188 ± 0.0943	0.8854 ± 0.0361
α=0.7,λ=0.1	0.01135 ± 0.00821	0.9099 ± 0.0900	0.8750 ± 0.0313
α=0.7,λ=0.2	0.01047 ± 0.00911	0.9177 ± 0.0951	0.8854 ± 0.0361

Every wrong-class cost in the original task matrices is at least one, so their margin clamp is inactive. Uniform scale experiments multiply all wrong-class entries by $s \in \{0.5, 1, 2\}$ and retrain, evaluating both the training matrix and the canonical matrix. Scaling by 0.5 puts 376 ordinary PMI-20 wrong-class entries below one and activates the clamp. Even without clamp activation, uniform scaling changes all wrong-class logit offsets; the full loss is not scale invariant. The normalized regularizer alone is invariant apart from the numerical stabilizer.

Critical-cost uncertainty multiplies only the four designated PMI-20 entries by $\gamma \in \{0.5, 0.75, 1, 1.25, 1.5\}$. Canonical-matrix scoring holds the evaluation objective fixed while the training cost is perturbed. The frozen CE reference has mean canonical nATC 0.009375 at every

multiplier. NICME is favorable at the two tested extremes, while the complete response is non-monotonic. These curves describe sensitivity to declared costs and do not establish a universal robustness guarantee. All configurations and seeds are retained; the test curves are not used to choose the manuscript setting.

Table 21: Uniform scale on PMI-20 (test, canonical evaluation matrix). Mean ± sample SD over three seeds.

Setting	nATC	BA	Target-min
scale=0.5	0.00599 ± 0.00142	0.9542 ± 0.0182	0.9167 ± 0.0180
scale=1	0.01115 ± 0.01013	0.9167 ± 0.0889	0.8750 ± 0.0541
scale=2	0.01172 ± 0.00818	0.9182 ± 0.0845	0.8854 ± 0.0477

Table 22: Critical cost uncertainty on PMI-20 (test, canonical evaluation matrix). Mean ± sample SD over three seeds.

Method / setting	nATC	BA	Target-min
CE reference, γ=0.5	0.00938 ± 0.00641	0.9599 ± 0.0201	0.8750 ± 0.0827
CE reference, γ=0.75	0.00938 ± 0.00641	0.9599 ± 0.0201	0.8750 ± 0.0827
CE reference, γ=1	0.00938 ± 0.00641	0.9599 ± 0.0201	0.8750 ± 0.0827
CE reference, γ=1.25	0.00938 ± 0.00641	0.9599 ± 0.0201	0.8750 ± 0.0827
CE reference, γ=1.5	0.00938 ± 0.00641	0.9599 ± 0.0201	0.8750 ± 0.0827
NICME, γ=0.5	0.00693 ± 0.00242	0.9484 ± 0.0189	0.9167 ± 0.0180
NICME, γ=0.75	0.01099 ± 0.00983	0.9229 ± 0.0902	0.8750 ± 0.0541
NICME, γ=1	0.01115 ± 0.01013	0.9167 ± 0.0889	0.8750 ± 0.0541
NICME, γ=1.25	0.00995 ± 0.00650	0.9615 ± 0.0142	0.8438 ± 0.0827
NICME, γ=1.5	0.00594 ± 0.00134	0.9677 ± 0.0119	0.9167 ± 0.0180

N JOINT CLASS IMBALANCE AND DIRECTED COSTS

Under label shift, $P_T(X \mid Y) = P_S(X \mid Y)$ and $P_T(Y = y) > 0 \Rightarrow P_S(Y = y) > 0$, a target-distribution surrogate risk can be expressed using source examples weighted by $P_T(Y)/P_S(Y)$. This identity separates a change in label prior from a change in application cost.

We additionally evaluate an exploratory training composition on natural PMI-20. Add $\tau \log \widehat{\pi}_k$ to each class logit before applying NICME's pairwise offsets, with $\tau = 1$ and Laplace-smoothed training priors $\widehat{\pi}_k = (n_k + 1)/(N + K)$. Counts come only from the natural training partition ($N = 8,393$). The expected-cost regularizer remains on the original, unadjusted probabilities, and primary inference remains raw-logit argmax. This composition is an experimental prior-correction mechanism; it is not asserted to be identical to importance weighting the full NICME loss.

Table 23 reports all evaluated methods and decisions. The composition raises target-min recall and reduces its seed variation relative to NICME alone, accompanied by higher cost and lower BA. CE and calibrated CE are strong references in this natural-prior setting. The result demonstrates that prior and directed-cost terms can be composed while exposing a concrete recall–cost choice.

Table 23: Natural PMI-20: joint prior and cost experiments. Mean ± sample SD over three seeds.

Objective	Scaling	Decision	nATC	BA	Target-min
CE	raw	argmax	0.00232 ± 0.00110	0.9919 ± 0.0012	0.9675 ± 0.0022
CE	temperature	cost-min	0.00136 ± 0.00036	0.9905 ± 0.0021	0.9525 ± 0.0147
Regularized CE	raw	argmax	0.00249 ± 0.00182	0.9844 ± 0.0148	0.9675 ± 0.0022
Regularized CE	temperature	cost-min	0.00219 ± 0.00140	0.9806 ± 0.0164	0.9508 ± 0.0310
Menon adjustment	raw	argmax	0.00412 ± 0.00157	0.9757 ± 0.0144	0.9729 ± 0.0154
Menon adjustment	temperature	cost-min	0.00372 ± 0.00163	0.9647 ± 0.0246	0.9362 ± 0.0319
NICME	raw	argmax	0.00370 ± 0.00365	0.9891 ± 0.0058	0.9492 ± 0.0339
NICME	temperature	cost-min	0.00231 ± 0.00201	0.9869 ± 0.0067	0.9329 ± 0.0506
NICME + prior	raw	argmax	0.00431 ± 0.00255	0.9788 ± 0.0140	0.9613 ± 0.0070
NICME + prior	temperature	cost-min	0.00319 ± 0.00191	0.9754 ± 0.0169	0.9250 ± 0.0557

O SCOPE AND PRACTICAL USE

NICME provides an explicit finite-loss margin interpretation for directed costs, with empirical evidence on fixed vision splits. The studies support useful operating points and a practical selection procedure rather than uniform superiority across every matrix, seed, and decision rule. The normalized expected-cost penalty and the margin-only variant should be considered alongside calibrated CE under a declared validation objective. Further split replications would characterize sampling variation beyond the training-seed variation measured here.

PMI-20 uses a medication-error cost protocol; the binary ratios and CIFAR-100 directions are transparent experimental constructions. Application-specific deployment requires domain-expert utilities, calibration assessment under the target population, and consideration of error redistribution. The source pool for Spider is not publicly redistributed, which limits independent reconstruction from images; its manifests, cost matrix, and derived results are provided. These access boundaries do not affect the bundled CPU reproduction of the numerical tables and controlled saved-logit analysis.