RETHINKING CLASS REWEIGHTING THROUGH TEMPORAL EVIDENCE FOR LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-tailed audio recognition typically measures class imbalance by training sample counts per class. This view overlooks how useful information is distributed within each sample: the labeled sound may occupy only a small portion, while persistent or repeated patterns make other portions redundant. Classes with identical sample counts therefore differ in the amount of irrelevant or repeated content. To account for recording content in class reweighting, we propose Constrained Effective Temporal Evidence Learning (CETE) for long-tailed audio recognition. CETE introduces a temporal head to identify temporal features that support the class decision. The head is trained with a loss that combines classification constraints with a penalty on the total temporal weight. This loss encourages correct classification after feature weighting while discouraging large weights on unnecessary background or repeated features. A teacher model selects correctly and confidently classified examples for learning temporal weights and computing class statistics. CETE aggregates their temporal weights within each class and combines the resulting statistics with sample counts to determine class loss weights, reducing bias toward common classes. Experiments on VGGSound-LT, iNatSound, and BirdCLEF demonstrate CETE's consistent superiority over the baselines and its robustness across controlled imbalance ratios and naturally long-tailed data.

1 INTRODUCTION

Many real-world audio datasets exhibit long-tailed class distributions, where a small number of common sound classes contain many training examples while many rare classes contain only a few (Ebbers et al., 2025; Fonseca et al., 2022). This imbalance causes recognition task to be dominated by common classes(Menon et al., 2021), biasing learning toward these classes and resulting in poorer recognition performance on rare classes.

Existing long-tailed learning methods can be broadly divided into data-oriented and training-oriented approaches. Data-oriented methods alter the examples used for learning through sampling or augmentation. Sampling-based methods rebalance the training data by changing how often examples from each class are selected. For example, class-balanced sampling selects rare-class examples more frequently (Kang et al., 2020). Augmentation-based methods increase data diversity by constructing new training examples from existing ones. Mixup combines pairs of examples, Remix adjusts the combined labels in favor of rare classes, DODA adaptively selects augmentation operations for each class, and Supervised Exploratory Learning (SEL) constructs exploratory examples for rare classes (Zhang et al., 2018; Chou et al., 2020; Wang et al., 2024; Jian et al., 2025).

Training-oriented methods modify training objectives, classifiers, or optimization strategies to reduce bias toward frequent classes. Objective-based methods modify the objective to control the influence of different classes. Focal Loss reduces the contribution of well-classified examples, Class-Balanced Loss reweights classes according to the effective number of samples, Balanced Softmax (BS) incorporates class frequencies into softmax normalization, LDAM assigns larger margins to rare classes, and LOS reduces bias toward common classes through label over-smoothing (Lin et al., 2017; Cui et al., 2019; Ren et al., 2020; Cao et al., 2019; Sun et al., 2025). Classifier-based methods adjust prediction scores or classifier parameters. Logit Adjustment corrects prediction scores using class priors, while Weight Balancing reduces class-dependent differences in classifier weights (Menon et al., 2021; Alshammari et al., 2022). Optimization-based methods modify the training procedure,

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.

as represented by the class-dependent optimization of CC-SAM and the two-stage multi-objective fine-tuning of TS-MOF (Zhou et al., 2023; Zhao et al., 2025). Despite these differences, both data-oriented and training-oriented methods generally treat each audio example as a complete unit. They do not distinguish the parts containing the labeled sound from background content or account for repeated information, and therefore cannot determine how much useful content supports recognition.

Class reweighting methods for long-tailed recognition typically determine class loss weights from the number of training examples in each class to reduce bias toward common classes (Cui et al., 2019). However, audio labels often indicate only which sound is present in a sample, without specifying its timing or duration. The labeled sound may occupy only a short interval, while the rest contains silence, background noise, or other sounds. In addition, persistent, periodic, or repeated sounds cause different parts of a sample to contain nearly identical acoustic features. Consequently, classes with identical sample counts may differ in the duration of the labeled sounds and the amount of irrelevant content in their recordings. These differences are overlooked when class loss weights depend solely on sample counts.

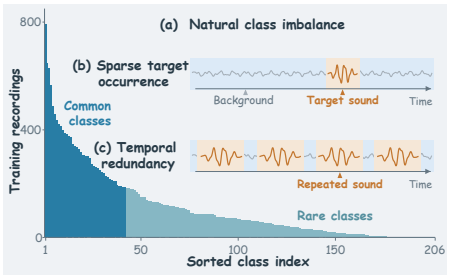

Figure 1: Motivation for CETE. Recording counts alone do not fully reflect differences in sparse target occurrence and temporal redundancy across classes.

To address this limitation, we propose Constrained Effective Temporal Evidence Learning (CETE) for long-tailed audio recognition. An audio classifier typically uses an encoder to extract temporal features from each sample and a classification head to predict its class. To account for recording content when determining class weights, we identify temporal information supporting the class decision, which we call temporal evidence. We therefore introduce a temporal head to estimate how much each temporal feature contributes to classification, expressing these contributions through continuous temporal weights. To learn these weights, we design a loss that combines classification constraints with a penalty on the total temporal weight. The constraints encourage correct classification after feature weighting, while the penalty discourages assigning large weights to every feature. Together, they encourage lower weights on background or repeated features unnecessary for classification. Notably, the classification head may misclassify examples early in training or when handling difficult samples. In such cases, its predictions do not provide a reliable reference for learning. CETE therefore introduces a teacher model to select correctly and confidently classified examples as training progresses. Only these examples are used to train the temporal head and compute class statistics. We aggregate their temporal weights within each class and combine the statistics with sample counts to determine class loss weights. This allows classifier training to account for both sample counts and recording content, with the aim of learning decision boundaries that better distinguish classes.

Our contributions are summarized as follows:

- *Classwise temporal evidence:* we define classwise temporal evidence to characterize how temporal weights and correlations affect the precision of classwise mean estimation. Under a fixed-weight model and the stated tail condition, we derive an estimation error bound that explains why recording counts alone do not determine estimation precision.
- *Temporal loss design:* combines complementary classification constraints with a penalty on the total weight, enabling CETE to learn continuous temporal weights for class reweighting. These weights complement sample counts, allowing training to account for differences in recording content.
- *Superior empirical performance:* CETE achieves the highest balanced accuracy among all compared methods on VGGSound-LT, iNatSound, and BirdCLEF, demonstrating its effectiveness under both controlled and naturally occurring class imbalance.

2 Classwise Temporal Information in Long-Tailed Audio Learning

2.1 Problem Setup

We consider long-tailed audio classification over C classes. The training set is $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an audio recording, $y_i \in \{1, \ldots, C\}$ is its label, and $n_c = |\{i : y_i = c\}|$ is the number

of training recordings in class c. A model parameterized by θ consists of a temporal encoder and a classifier. The encoder maps x_i to a sequence of T temporal features $H_i = (h_{i,1}, \ldots, h_{i,T})$, and the classifier maps H_i to class logits $z(H_i) = f_\theta(H_i)$. For notational simplicity, θ denotes all trainable parameters of the encoder and classifier. Let $\ell_i(\theta) = \ell(z(H_i), y_i)$ denote the classification loss. For any temporal representation H, we define the margin of class c as

$$m_c(H) = z_c(H) - \max_{k \neq c} z_k(H). \tag{1}$$

A larger margin indicates a stronger preference for class c over the competing classes. Cui et al. (2019) use n_c to determine the contribution of each class during training. We instead examine the effective information underlying each class decision.

2.2 Temporal Weights under Margin Constraints

Because temporal features contribute unequally to the classifier's decision, we assign a weight to each feature and evaluate these weights through their effect on the class margin. For a fixed classifier, consider the temporal features $H = (h_1, \ldots, h_T)$ of a recording with label c. Let $a = (a_1, \ldots, a_T)$ be a candidate weight vector. We construct two complementary representations:

$$H^+(a) = a \odot H, H^-(a) = (1 - a) \odot H, \tag{2}$$

where each temporal weight multiplies every component of the corresponding feature vector. In $H^+(a)$, feature h_t is multiplied by a_t; in $H^-(a)$, the same feature is multiplied by $1 - a_t$. Thus, increasing a_t gives this feature more weight in $H^+(a)$ and less weight in $H^-(a)$. We choose the weights by minimizing their total value subject to two requirements on the class margins.

Definition 1 (Constrained Temporal Weights)**.** *For $\varepsilon \geq 0$ and $\Delta > 0$, assuming that the feasible set is nonempty, we define the temporal weights q as any solution to*

$$\begin{aligned} q \in \arg\min_{a \in [0,1]^T} \quad & \frac{1}{T} \sum_{t=1}^{T} a_t \\ \text{s.t.} \quad & m_c\big(H^+(a)\big) \geq m_c(H) - \varepsilon, \\ & m_c\big(H^+(a)\big) - m_c\big(H^-(a)\big) \geq \Delta. \end{aligned} \tag{3}$$

The solution $q = (q_1, \ldots, q_T)$ is the temporal weight vector, where q_t weights temporal feature h_t. A larger q_t retains more of h_t in $H^+(q)$ and removes more of it in $H^-(q)$. We call this selection temporal evidence for class c: the retained margin is at least the original margin minus ε and exceeds the complementary margin by at least Δ. Minimizing the overall selected weight favors compact temporal selections satisfying both constraints.

2.3 Classwise Temporal Evidence and Estimation Precision

The temporal weights above describe how much of each feature is retained in a recording. Class reweighting uses a shared loss weight for all recordings of the same class. We therefore examine the weighted features from these recordings together. Because correlated features may carry repeated information, we consider their correlations as well as their weights.

Class c contains n_c recordings, each with T temporal features, giving $M = n_c T$ temporal units. We index these units by $j = 1, \ldots, M$, and let q_j denote the weight assigned to unit j. For clarity, we examine one fixed component of the feature vectors and denote its value in unit j by S_j. We consider a fixed-weight model in which $\mathbb{E}[S_j] = \mu_c$ and $\mathrm{Var}(S_j) = \sigma_c^2 > 0$, where μ_c is the expected value of this feature component for class c. Assuming $\sum_{j=1}^{M} q_j > 0$, we estimate this mean by

$$\widehat{\mu}_c = \frac{\sum_{j=1}^{M} q_j S_j}{\sum_{j=1}^{M} q_j}. \tag{4}$$

Correlations between temporal features affect the variance of this estimate. In particular, positively correlated feature values provide less additional precision than independent values with the same weights and marginal variances. We account for this dependence through the following definition.

Definition 2. *Let* $\boldsymbol{q} = (q_1, \ldots, q_M)^\top$, $\boldsymbol{e} = (1, \ldots, 1)^\top$, *and* $\boldsymbol{\Sigma}_c = [\mathrm{Cov}(S_j, S_k)]_{j,k=1}^M$. *Assuming* $\boldsymbol{q}^\top \boldsymbol{\Sigma}_c \boldsymbol{q} > 0$, *we define the temporal evidence for estimating* μ_c *as*

$$n_c^{\mathrm{e}} = \frac{\sigma_c^2 (\boldsymbol{e}^\top \boldsymbol{q})^2}{\boldsymbol{q}^\top \boldsymbol{\Sigma}_c \boldsymbol{q}}. \tag{5}$$

Under the fixed-weight model, this definition gives

$$\mathrm{Var}(\widehat{\mu}_c) = \frac{\boldsymbol{q}^\top \boldsymbol{\Sigma}_c \boldsymbol{q}}{(\boldsymbol{e}^\top \boldsymbol{q})^2} = \frac{\sigma_c^2}{n_c^{\mathrm{e}}}. \tag{6}$$

An average of K independent observations, each with variance σ_c^2, has variance σ_c^2/K. Thus, n_c^{e} expresses the precision of the weighted estimate as the number of independent observations needed to achieve the same variance. For a fixed σ_c^2, a larger n_c^{e} means a more precise estimate of μ_c.

Proposition 1. *If the temporal units are independent, then*

$$n_c^{\mathrm{e}} = \frac{\left(\sum_{j=1}^M q_j\right)^2}{\sum_{j=1}^M q_j^2}. \tag{7}$$

In particular, equal nonzero weights give $n_c^{\mathrm{e}} = M$, *while binary weights* $q_j \in \{0, 1\}$ *give* $n_c^{\mathrm{e}} = \sum_{j=1}^M q_j$. *With fixed weights and marginal variances, increasing the covariance between positively weighted units reduces* n_c^{e}, *provided the other covariances remain unchanged.*

This result explains why recording counts alone do not determine temporal evidence. The count n_c determines the number of temporal units, but not their weights or correlations. Even under independence, continuous weights generally require both their sum and their squared sum to determine n_c^{e}. We next relate temporal evidence to the deviation of the estimated mean from its expected value.

Theorem 1. *For each class c, let* $\boldsymbol{S} = (S_1, \ldots, S_M)^\top$. *Under the fixed-weight model, suppose that, for every* $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda \boldsymbol{q}^\top (\boldsymbol{S} - \mu_c \boldsymbol{e})\right)\right] \le \exp\left(\frac{\lambda^2}{2} \boldsymbol{q}^\top \boldsymbol{\Sigma}_c \boldsymbol{q}\right). \tag{8}$$

Then, for any $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$|\widehat{\mu}_c - \mu_c| \le \sigma_c \sqrt{\frac{2 \log(2C/\delta)}{n_c^{\mathrm{e}}}} \qquad \text{for all } c \in \{1, \ldots, C\}. \tag{9}$$

Under the condition in Eq. (8), larger n_c^{e} yields a tighter mean estimation error bound for fixed σ_c. Classes with identical recording counts can therefore have different bounds because their temporal weights and correlations differ. This analysis assumes fixed weights; applying the bound to weights learned from the same features requires accounting for the dependence between weights and features. CETE aggregates learned temporal weights to adjust recording counts as a practical correction, without assuming that this aggregate equals the covariance-dependent quantity n_c^{e}.

Proposition 2. *For binary weights* $q_j \in \{0, 1\}$, *independence gives* $n_c^{\mathrm{e}} = \sum_{j=1}^M q_j$. *With fixed weights and marginal variance, positive temporal covariance reduces* n_c^{e}. *Since* n_c *determines* $M = n_c T$ *but not this covariance, neither* n_c *nor* M *generally determines estimation precision.*

Proposition 2 shows that selected temporal counts characterize estimation precision only under independent binary selection. CETE therefore uses the normalized selected weight as a tractable classwise correction, rather than an exact estimator of n_c^{e}.

3 Constrained Effective Temporal Evidence Learning

CETE combines recording counts with learned temporal weights for class reweighting. A temporal head predicts class-conditional weights and is trained using the margin constraints introduced in Section 2.2. An exponential moving average (EMA) teacher selects recordings for both temporal weight learning and classwise aggregation.

3.1 Learning Temporal Weights

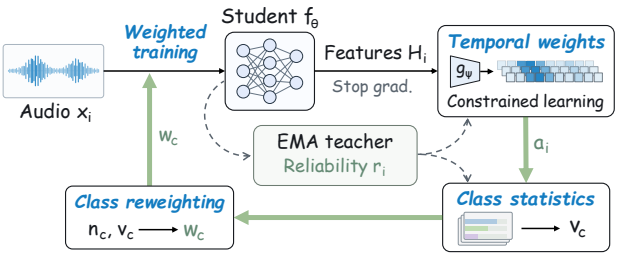

Figure 2: Overview of CETE. CETE combines reliable temporal information with recording counts to determine class weights for long-tailed audio learning.

Solving the constrained problem in Definition 1 separately for each recording would require repeated optimization during training. The optimal weights can also change as the encoder and classifier are updated. We therefore train a shared temporal head g_ψ to predict class-conditional weights directly from the temporal features. The head utilizes the same parameters across recordings and produces weights for each input in a single forward pass. Its training objective combines the mean temporal weight with penalties for violating the two margin constraints, encouraging smaller weights while discouraging margin violations. Given $H_i = (h_{i,1}, \ldots, h_{i,T})$, we compute

$$A_i = g_\psi(H_i), \qquad a_i = A_{i,:,y_i}, \tag{10}$$

where $A_{i,t,c}$ assigns a weight to temporal feature $h_{i,t}$ for class c, and a_i contains the weights for the observed class y_i. The same feature can therefore receive different weights for different classes. We use a_i in Eq. (2) to construct the complementary representations $H_i^+(a_i)$ and $H_i^-(a_i)$.

Definition 3 (Margin Constraint Violation)**.** *For a recording (H_i, y_i) and a candidate temporal weight vector $a \in [0,1]^T$, we define the total violation of the two margin constraints as*

$$\begin{aligned}\mathcal{V}_i(a) = \; & \left[m_{y_i}(H_i) - m_{y_i}\left(H_i^+(a)\right) - \varepsilon\right]_+ \\ & + \left[\Delta - m_{y_i}\left(H_i^+(a)\right) + m_{y_i}\left(H_i^-(a)\right)\right]_+ ,\end{aligned} \tag{11}$$

where $[u]_+ = \max(u, 0)$.

The first term penalizes a decrease in the original class margin greater than ε. The second penalizes a margin difference between $H_i^+(a)$ and $H_i^-(a)$ smaller than Δ. Thus, $\mathcal{V}_i(a) = 0$ exactly when both constraints in Definition 1 are satisfied. We combine this violation with the mean temporal weight to define the temporal loss:

$$\ell_i^{\mathrm{tw}}(a) = s(a) + \eta \mathcal{V}_i(a), \qquad s(a) = \frac{1}{T}\sum_{t=1}^{T} a_t, \tag{12}$$

where $\eta > 0$ controls the penalty strength. The mean weight term encourages smaller temporal weights, while the constraint penalty discourages reductions that violate either margin constraint.

Proposition 3. *For recording i, let $\mathcal{F}_i = \{a \in [0,1]^T : \mathcal{V}_i(a) = 0\}$. Suppose that $\mathcal{F}_i$ is nonempty and that the class margins are continuous with respect to a. For each $\eta > 0$, let a_η be a global minimizer of $\ell_i^{\mathrm{tw}}(a)$ over $[0,1]^T$. Then every accumulation point $\bar{a}$ of a_η as $\eta \to \infty$ satisfies $\bar{a} \in \arg\min_{a \in \mathcal{F}_i} s(a)$, and hence solves the constrained problem in Definition 1.*

The proof is provided in Appendix A.5. This proposition describes the limiting behavior of globally optimized weights for individual recordings. CETE trains a shared temporal head g_ψ with a finite η, so its predictions are not assumed to solve the constrained problem exactly.

3.2 Class Reweighting with Temporal Weights

CETE aggregates learned temporal weights to adjust recording counts. Under fixed weights and uncorrelated responses across recordings, Appendix A.4 interprets the normalized weight sum as a lower bound on temporal evidence. An EMA teacher $f_{\bar{\theta}}$ of the student encoder and classifier selects recordings for temporal weight learning and aggregation. Its parameters $\bar{\theta}$ are updated from the student parameters θ by

$$\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta, \tag{13}$$

where $\mu \in [0, 1)$ is the teacher momentum. For recording i, let $\bar{p}_{i,y_i}$ and $\bar{m}_{i,y_i}$ denote the teacher probability and margin for the observed class. We define the selection indicator as

$$r_i = \mathbb{I}[\bar{p}_{i,y_i} \geq \tau_p \;\wedge\; \bar{m}_{i,y_i} \geq \tau_m], \tag{14}$$

where $\mathbb{I}[\cdot]$ equals one when its condition holds and zero otherwise. The probability threshold is $\tau_p \in (0, 1)$, and the margin threshold is $\tau_m \geq 0$. The indicator r_i controls temporal weight learning and aggregation; all recordings retain their observed labels and contribute to the classification loss. Let $s(a_i) = T^{-1} \sum_{t=1}^{T} a_{i,t}$ be the mean temporal weight defined in Eq. (12). At the current model parameters, the total contribution from the selected recordings of class c is

$$v_c^\star = \sum_{i:y_i=c} r_i s(a_i) = \frac{1}{T} \sum_{i:y_i=c} \sum_{t=1}^{T} r_i a_{i,t}, \tag{15}$$

with $0 \leq v_c^\star \leq n_c$. We maintain an EMA estimate v_c of this total with momentum β. For each class c present in the current batch B, let B_c denote the indices of its recordings within the batch. We compute the batch estimate $\widehat{v}_c$ of the class total and update v_c by

$$\widehat{v}_c = \frac{n_c}{|B_c|} \sum_{i \in B_c} r_i s(a_i), \quad v_c \leftarrow \beta v_c + (1 - \beta)\widehat{v}_c, \tag{16}$$

Definition 4 (Temporal-Adjusted Class Reweighting)**.** *Given the recording count n_c and the running estimate v_c, CETE defines the adjusted count $\widetilde{n}_c$ and normalized class weight w_c as*

$$\begin{aligned} \widetilde{n}_c &= (1 - \alpha) n_c + \alpha v_c, \\ w_c &= \frac{(\widetilde{n}_c + \epsilon_w)^{-\gamma}}{\frac{1}{C} \sum_{k=1}^{C} (\widetilde{n}_k + \epsilon_w)^{-\gamma}}, \end{aligned} \tag{17}$$

where $\alpha \in [0, 1]$ controls the temporal correction, $\gamma \geq 0$ controls the reweighting strength, and $\epsilon_w > 0$ ensures numerical stability.

For $\gamma > 0$, the inverse-power form assigns larger loss weights to recordings from classes with smaller adjusted counts. Increasing γ strengthens this relative weighting, while $\gamma = 0$ gives uniform weights. Setting $\alpha = 0$ recovers count-based reweighting. The denominator preserves the weight ratios and normalizes their mean across classes to one. Let B_r denote the indices of recordings in B with $r_i = 1$. Using these class weights, the training objective is

$$\mathcal{L} = \frac{1}{|B|} \sum_{i \in B} w_{y_i} \ell_i(\theta) + \frac{\lambda_{\text{tw}}}{|B_r|} \sum_{i \in B_r} \ell_i^{\text{tw}}(a_i), \tag{18}$$

where $\ell_i(\theta)$ is the classification loss and $\lambda_{\text{tw}} \geq 0$ controls the temporal loss. The second term is omitted when B_r is empty. The classification and temporal losses update only the student parameters θ and temporal head parameters ψ, respectively. The EMA teacher, v_c, and w_c receive no gradients. Only the student encoder and classifier are retained at inference.

4 Experiments

We evaluate CETE on controlled and natural long-tailed audio benchmarks. We report comparisons with existing methods, followed by ablations and analyses of results.

4.1 Experimental Setup

Datasets. We evaluate CETE on three long-tailed audio benchmarks:

- *VGGSound-LT* consists of three controlled long-tailed variants that we construct from the VGG-Sound training set (Chen et al., 2020), with different imbalance ratios (IRs). These variants allow us to study increasing class imbalance within the same source dataset.
- *iNatSound* is a naturally imbalanced bioacoustic benchmark containing recordings collected under diverse acoustic conditions (Chasmai et al., 2024). Its class distribution reflects the natural frequency variation among sound-producing species.
- *BirdCLEF* is a large-scale bird-sound recognition benchmark with a naturally long-tailed class distribution and substantial background interference (Klinck et al., 2025). It provides a realistic setting in which rare acoustic events may occupy only a limited portion of each recording.

Table 1: Comparison on VGGSound-LT with imbalance ratios of 10, 50, and 100. Many, Medium, and Few report accuracy (%) for classes grouped by training-set frequency. Gray-shaded ALL columns present balanced accuracy (%), with CETE's results highlighted in bold. Values in parentheses indicate percentage-point improvements over the highest-scoring baseline in ALL per setting.

Method	Reference	IR=10				IR=50				IR=100			
		Many	Medium	Few	**ALL**	Many	Medium	Few	**ALL**	Many	Medium	Few	**ALL**
CE	–	61.0	50.1	42.6	51.1	57.2	47.0	19.6	43.8	57.8	42.8	14.1	40.6
Focal	ICCV17	58.7	50.6	44.5	51.2	53.7	48.3	21.2	44.3	55.0	43.7	6.7	39.5
CB-CE	CVPR19	57.9	51.4	47.2	52.1	51.9	49.6	29.4	46.3	51.8	46.3	21.4	43.2
LDAM	NeurIPS19	57.0	50.5	45.8	51.0	50.6	47.7	25.9	44.3	51.8	43.8	19.5	41.1
cRT	ICLR20	58.1	51.6	47.1	52.2	52.0	49.6	29.7	46.4	51.8	46.3	21.5	43.2
τ-norm	ICLR20	60.8	50.2	42.9	51.2	57.1	47.4	21.3	44.4	57.5	43.2	16.4	41.1
LWS	ICLR20	60.8	50.7	43.6	51.6	56.6	47.8	21.6	44.6	57.4	44.1	14.4	41.4
NCM	ICLR20	57.7	51.2	47.4	52.0	52.1	49.5	30.7	46.6	51.3	46.4	23.5	43.5
LogitAdjust	ICLR21	56.5	50.9	48.8	52.0	52.3	50.9	33.1	47.9	52.1	46.5	27.9	44.4
LAS	CVPR21	54.4	51.0	51.5	52.2	48.4	50.4	34.6	47.2	47.0	47.5	27.1	44.2
BS	NeurIPS20	58.1	51.5	48.5	52.6	52.2	49.6	35.2	47.5	51.3	45.9	31.1	44.4
CC-SAM	CVPR23	57.5	50.1	43.6	50.3	52.6	47.2	24.2	44.0	53.3	43.4	17.1	40.8
DODA	ICLR24	60.5	49.4	40.9	50.1	57.2	47.2	19.7	44.0	57.4	42.1	13.8	39.9
TS-MOF	NeurIPS25	58.3	51.6	46.8	52.2	53.4	48.9	28.8	46.1	53.5	45.8	11.2	41.5
LOS	ICLR25	58.0	51.1	47.9	52.2	51.2	49.5	31.1	46.4	53.1	46.8	23.3	44.1
RBL	ICML23	58.9	51.6	47.5	52.5	53.0	49.9	33.6	47.5	52.7	46.8	23.3	44.0
SEL	ICCV25	40.3	49.8	51.5	47.4	32.0	50.7	37.5	45.0	33.8	47.0	31.9	42.6
CETE	Ours	57.5	52.8	51.9	**53.9(+1.3)**	51.4	51.2	37.9	**48.8(+0.9)**	50.4	47.4	35.8	**46.1(+1.7)**

Evaluation Metrics. Classes are grouped into Many, Medium, and Few by training recording counts. We report mean class-wise recall within each group and across all evaluated classes (ALL), considering only classes with ground-truth recordings in the respective evaluation split. Classes absent from that split are excluded from all averages. Scores are reported as percentages. Group thresholds and evaluation class coverage are detailed in Appendix B.1.

Baselines. We compare CETE with data-oriented and training-oriented methods for long-tailed recognition, using cross-entropy (CE) as a reference. All methods share data splits, audio preprocessing, and ResNet-50 backbone initialization (He et al., 2016). Focal, CB-CE, LDAM, and RBL jointly train the encoder and classifier. RBL retains its geometric constraints and inference correction. Methods based on classifier retraining retain their prescribed training stages, including the frozen-backbone stage of TS-MOF. LAS denotes the label-aware smoothing component of MiSLAS. Appendix B.2 specifies the objectives, training procedures, and inference rules used for each baseline.

Implementation Details. All methods utilize a ResNet-50 encoder with the same backbone initialization and joint-training settings. Reliability thresholds and loss coefficients are selected on the validation set and fixed for evaluation. Architecture details, audio preprocessing, optimization schedules, and hyperparameters are provided in appendix B.3.

4.2 Experimental Results and Discussions

Tables 1 and 2 show that CETE achieves the highest balanced accuracy among the compared methods across all evaluated settings. Its gains are accompanied by improved recognition of underrepresented classes, suggesting that CETE helps address the uneven performance associated with imbalanced training data. Although some baselines retain higher accuracy on frequent classes, CETE provides a more favorable overall trade-off across frequency groups. This property is particularly relevant to bioacoustic recognition, where infrequently observed species remain important recognition targets. The consistent overall improvements across controlled and naturally imbalanced benchmarks further support the applicability of CETE to different long-tailed audio settings.

4.2.1 Ablation Study

Component ablation. We remove one component at a time to examine the roles of preservation, complementarity, compact selection, reliability filtering, and classwise temporal correction. Removing preservation or complementarity deletes the corresponding term from $\mathcal{V}_i$, while removing

Table 2: Performance comparison on iNatSound and BirdCLEF under their natural class distributions.

Method	Reference	iNatSound				BirdCLEF			
		Many	Medium	Few	ALL	Many	Medium	Few	ALL
CE	–	71.3	50.4	30.6	49.0	73.4	57.0	36.0	57.1
Focal	ICCV17	72.9	54.4	30.8	51.1	72.8	58.0	36.8	57.4
CB-CE	CVPR19	66.8	53.1	37.3	51.2	72.1	59.2	37.7	57.9
LDAM	NeurIPS19	67.6	46.5	25.8	44.8	70.9	56.6	38.6	56.8
cRT	ICLR20	67.0	53.1	37.1	51.2	72.6	59.3	39.5	58.6
τ-norm	ICLR20	65.5	44.9	33.0	46.0	73.2	57.8	38.6	58.0
LWS	ICLR20	71.4	50.6	31.3	49.3	73.8	57.6	37.7	57.9
NCM	ICLR20	57.5	44.1	32.5	43.5	72.5	60.1	39.5	58.8
LogitAdjust	ICLR21	70.4	52.2	36.1	51.4	70.8	61.7	42.1	59.6
LAS	CVPR21	59.6	49.4	37.2	47.8	71.4	60.6	39.5	58.6
BS	NeurIPS20	69.7	52.5	36.4	51.4	71.6	60.8	43.0	59.8
CC-SAM	CVPR23	70.1	49.5	31.1	48.5	68.9	53.6	36.0	54.2
DODA	ICLR24	70.0	49.7	31.0	48.5	72.8	55.3	34.2	55.7
TS-MOF	NeurIPS25	72.8	53.8	34.7	52.1	72.4	57.7	33.3	56.2
LOS	ICLR25	68.8	54.6	36.3	52.0	71.8	59.1	38.6	58.0
RBL	ICML23	69.1	51.3	35.7	50.5	73.3	59.3	36.8	58.1
SEL	ICCV25	47.4	38.3	26.8	36.7	63.1	61.0	40.4	56.0
CETE	Ours	71.7	54.7	37.6	**53.2(+1.1)**	70.8	62.8	47.4	**61.4(+1.6)**

compactness deletes $s(a_i)$ from the temporal objective. Removing reliability filtering sets $r_i = 1$ for every example in both temporal-head optimization and classwise aggregation. Removing temporal adjustment sets $\alpha = 0$, recovering the count-based reweighting in Eq. (17). All other settings remain fixed. Table 3 summarizes the component ablations.

4.2.2 Robustness Analysis

Robustness to sound duration and repetition. To assess robustness across sound categories with different durations and repetition rates, we construct a synthetic evaluation dataset using target events from the DESED soundbank (Turpault et al., 2019). Following established soundscape synthesis principles (Salamon et al., 2017), we mix these events with non-target background recordings. We control two factors: target occupancy, the fraction of each recording covered by target segments, and repetition, the extent to which the same target event is reused. We vary target occupancy from low to high while keeping the repetition setting fixed, while repetition levels are compared at a fixed total target occupancy. Recording duration and background sampling are matched across the corresponding comparisons. As shown in Fig. 3, CETE consistently outperforms the baselines across the tested conditions, supporting its robustness across sound categories with different target durations and repetition levels. Detailed synthesis settings and class assignments are provided in Appendix C.1.

Table 3: Component ablations on BirdCLEF. Many, Medium, and Few denote macro-averaged accuracy within each frequency group; ALL denotes balanced accuracy over classes (%).

Variant	Many	Medium	Few	ALL
w/o Preservation	70.6	62.2	45.9	60.7
w/o Complementarity	70.9	62.2	45.7	60.8
w/o Compactness	71.1	62.5	46.1	61.0
w/o Reliability	71.0	62.0	45.1	60.5
w/o Temporal	70.5	61.3	44.7	60.0
Full (CETE)	70.8	62.8	47.4	**61.4**

Table 4: Backbone comparison on *BirdCLEF*.

	ResNet-50 (He et al., 2016)				PANNs CNN14 (Kong et al., 2020)				AST-Base (Gong et al., 2021)				BEATs-Base (Chen et al., 2023)			
Method	Many	Medium	Few	ALL	Many	Medium	Few	ALL	Many	Medium	Few	ALL	Many	Medium	Few	ALL
CE	73.4	57.0	36.0	57.1	76.8	62.1	42.6	62.0	77.2	61.4	41.9	61.7	80.1	67.3	49.8	67.1
Focal	72.8	58.0	36.8	57.4	76.2	62.8	44.1	62.4	75.9	63.0	44.0	62.4	79.3	68.1	51.0	67.4
LDAM	70.9	56.6	38.6	56.8	74.7	61.9	45.6	62.0	74.3	63.6	47.3	62.9	77.4	67.8	53.2	67.2
CETE	70.8	62.8	47.4	61.4	74.9	66.3	52.7	65.6	74.6	65.8	51.4	65.0	77.8	70.3	57.6	69.5

Different audio backbones. We compare CE, Focal, LDAM, and CETE across four audio backbones: ResNet-50 (He et al., 2016), PANNs CNN14 (Kong et al., 2020), AST-Base (Gong et al., 2021), and BEATs-Base (Chen et al., 2023). Within each backbone, comparisons share data splits, audio preprocessing, and initialization, with training procedures specified in Appendix B.2. CETE retains its temporal objective and class-wise correction across backbones, adapting the temporal head to each encoder's feature dimension. Table 4 assesses CETE across these backbones relative to CE, Focal, and LDAM. Training and inference costs are detailed in Appendix C.3. The temporal head and EMA teacher are discarded after training, leaving only the student encoder and classifier at inference.

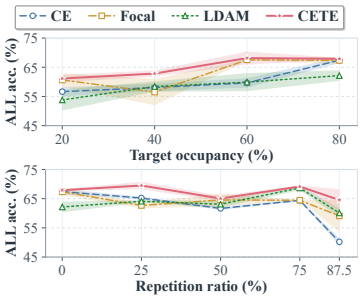

Figure 3: ALL balanced accuracy on DESED soundscapes.

4.2.3 Sensitivity Analysis

We vary temporal correction α, temporal feature count T, penalty strength η, and reliability thresholds τ_p and τ_m one at a time, keeping other settings fixed. The α sweep includes the count-based reference $\alpha = 0$, while the T sweep fixes recording duration and audio preprocessing. Figure 4 reports changes in Many, Medium, Few, and ALL accuracy relative to the default configuration, alongside temporal selection and teacher acceptance diagnostics. Accuracy varies modestly near the defaults, while larger departures tend to affect Few classes more strongly. Higher η increases selected temporal mass even after the joint validation rate largely saturates. Stricter thresholds reduce teacher acceptance, especially for Few classes, and coincide with lower Few accuracy. These patterns suggest balancing temporal validation, selection compactness, and coverage of underrepresented classes.

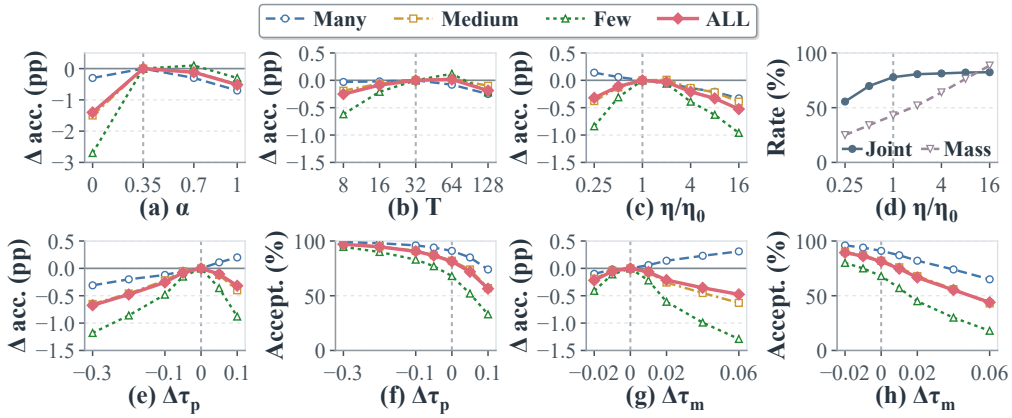

Figure 4: Hyperparameter sensitivity of CETE. Panels (a–c), (e), and (g) report accuracy changes from the default configuration in percentage points. Panel (d) shows the joint validation rate and selected temporal mass; panels (f) and (h) show teacher acceptance rates. η is normalized by its default η_0, and $\Delta\tau_p$ and $\Delta\tau_m$ denote offsets from default thresholds.

5 Conclusion

We presented constrained effective temporal evidence learning for long-tailed audio recognition, motivated by the observation that recording counts alone do not fully reflect differences in sparse target occurrence and temporal redundancy across classes. Our analysis relates classwise temporal evidence to feature mean estimation precision under a fixed-weight model, motivating the incorporation of recording content into class reweighting. CETE introduces a temporal head to identify features supporting the class decision through continuous weights, with a loss that encourages compact selection and penalizes violations of margin preservation and complementarity constraints. An EMA teacher selects recordings for temporal weight learning and classwise aggregation, and the aggregated weights provide a practical correction to recording counts. Experiments on VGGSound-LT, iNatSound, and BirdCLEF show improvements in rare-class recognition and balanced accuracy.

AI USE STATEMENT

We used generative AI tools to assist with literature retrieval, draft and revise manuscript text, and provide feedback on experimental methodology and the interpretation of results. AI assistance also supported checks of reported metrics and consistency between the main text and appendix. The authors take full responsibility for the paper, including all claims and results.

REFERENCES

Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6897–6907, 2022.

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

Mustafa Chasmai, Alexander Shepard, Subhransu Maji, and Grant Van Horn. The iNaturalist sounds dataset. In *Advances in Neural Information Processing Systems*, volume 37, pp. 132524–132544. Curran Associates, Inc., 2024.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 2023. URL `https://proceedings.mlr.press/v202/chen23ag.html`.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In *Computer Vision – ECCV 2020 Workshops*, pp. 95–110, 2020.

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

Janek Ebbers, François G. Germain, Kevin Wilkinghoff, Gordon Wichern, and Jonathan Le Roux. No class left behind: A closer look at class balancing for audio tagging. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, April 2025.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022.

Peifeng Gao, Qianqian Xu, Peisong Wen, Zhiyong Yang, Huiyang Shao, and Qingming Huang. Feature directions matter: Long-tailed learning via rotated balanced representation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27542–27563. PMLR, 2023.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech*, pp. 571–575, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Zhongquan Jian, Yanhao Chen, Yancheng Wang, Junfeng Yao, Meihong Wang, and Qingqiang Wu. Supervised exploratory learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1870–1880, October 2025.

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

Holger Klinck, Juan Sebastián Cañas, Maggie Demkin, Sohier Dane, Stefan Kahl, and Tom Denton. BirdCLEF+ 2025. https://www.kaggle.com/competitions/birdclef-2025, 2025. Kaggle.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

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

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

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4175–4186, 2020.

Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 344–348, 2017.

Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OeKp3AdiVO.

Nicolas Turpault, Romain Serizel, Ankit Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop*, pp. 253–257, 2019.

Binwu Wang, Pengkun Wang, Wei Xu, Xu Wang, Yudong Zhang, Kun Wang, and Yang Wang. Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RzY9qQHUXy.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhe Zhao, Zhiheng Gong, Pengkun Wang, HaiBin Wen, Cankun Guo, Bo Xue, Xi Lin, Zhenkun Wang, Qingfu Zhang, and Yang Wang. TS-MOF: Two-stage multi-objective fine-tuning for long-tailed recognition. In *Advances in Neural Information Processing Systems*, volume 38, pp. 5224–5252. Curran Associates, Inc., 2025.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16489–16498, 2021.

Zhipeng Zhou, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Wei Gong. Class-conditional sharpness-aware minimization for deep long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3499–3509, 2023.

A Proofs and Additional Theoretical Discussion

This section provides the arguments supporting Section 2 and the penalty-limit result in Section 3.1. Throughout the mean-estimation analysis, the temporal weights are fixed, and the feature component, marginal moments, and covariance matrix are those specified in Section 2.3.

A.1 Unbiasedness and the Variance Identity

Fix class c and write $Q_c = \boldsymbol{e}^\top \boldsymbol{q} > 0$ for its total temporal weight. Because $\boldsymbol{q}$ is fixed and $\mathbb{E}[\boldsymbol{S}] = \mu_c \boldsymbol{e}$,

$$\mathbb{E}[\widehat{\mu}_c] = \frac{\boldsymbol{q}^\top \mathbb{E}[\boldsymbol{S}]}{Q_c} = \mu_c. \tag{19}$$

Using $\boldsymbol{\Sigma}_c = \mathrm{Cov}(\boldsymbol{S})$, we obtain

$$\begin{aligned} \mathrm{Var}(\widehat{\mu}_c) &= \frac{1}{Q_c^2} \sum_{j=1}^{M} \sum_{k=1}^{M} q_j q_k \,\mathrm{Cov}(S_j, S_k) \\ &= \frac{\boldsymbol{q}^\top \boldsymbol{\Sigma}_c \boldsymbol{q}}{Q_c^2} = \frac{\sigma_c^2}{n_c^{\mathrm{e}}}. \end{aligned} \tag{20}$$

The last equality follows from Definition 2. Together with Eq. (19), it gives $\mathbb{E}[(\widehat{\mu}_c - \mu_c)^2] = \sigma_c^2 / n_c^{\mathrm{e}}$. This statement concerns estimation of the fixed feature component under the specified model.

A.2 Independent Units and the Effect of Covariance

Proof of Proposition 1. Independence and equal marginal variances give $\boldsymbol{\Sigma}_c = \sigma_c^2 I_M$, where I_M is the M-dimensional identity matrix. Substituting into Definition 2 yields

$$n_c^{\mathrm{e}} = \frac{\left(\sum_{j=1}^{M} q_j\right)^2}{\sum_{j=1}^{M} q_j^2}. \tag{21}$$

In general, the denominator in Definition 2 can be written as

$$\boldsymbol{q}^\top \boldsymbol{\Sigma}_c \boldsymbol{q} = \sigma_c^2 \sum_{j=1}^{M} q_j^2 + 2 \sum_{j<k} q_j q_k \,\mathrm{Cov}(S_j, S_k). \tag{22}$$

Consider two valid covariance matrices that differ only by an increase $\kappa > 0$ in the entries (j, k) and (k, j), where $j \neq k$. The denominator increases by $2\kappa q_j q_k$, while the numerator remains unchanged. Thus, n_c^{e} decreases strictly when $q_j q_k > 0$.

Proof of Proposition 2. The binary independent case and the covariance comparison follow from Proposition 1. To establish that recording counts do not determine estimation precision, fix n_c, T, and $M = n_c T \geq 2$, and set all weights to one. Consider the valid covariance family

$$\boldsymbol{\Sigma}_c = \sigma_c^2 \big((1-\rho) I_M + \rho \boldsymbol{e}\boldsymbol{e}^\top\big), \qquad \rho \in [0, 1], \tag{23}$$

where ρ is the common pairwise correlation. Substitution into Definition 2 gives

$$n_c^{\mathrm{e}} = \frac{M}{1 + (M-1)\rho}. \tag{24}$$

With n_c, T, and the weights fixed, temporal evidence equals M at $\rho = 0$ and 1 at $\rho = 1$. Therefore, recording counts alone do not determine estimation precision.

A.3 Simultaneous Concentration across Classes

Proof of Theorem 1. Fix class c. Let $Q_c = \boldsymbol{e}^\top \boldsymbol{q} > 0$, $V_c = \boldsymbol{q}^\top \boldsymbol{\Sigma}_c \boldsymbol{q} > 0$, and $Z_c = \boldsymbol{q}^\top (\boldsymbol{S} - \mu_c \boldsymbol{e})$. Then $\widehat{\mu}_c - \mu_c = Z_c / Q_c$. For $t > 0$ and $\lambda > 0$, Markov's inequality and Eq. (8) give

$$\begin{aligned} \mathbb{P}(\widehat{\mu}_c - \mu_c \geq t) &\leq e^{-\lambda Q_c t} \mathbb{E}[e^{\lambda Z_c}] \\ &\leq \exp\left(-\lambda Q_c t + \frac{\lambda^2 V_c}{2} \right). \end{aligned} \tag{25}$$

Choosing $\lambda = Q_c t / V_c$ and applying the same argument to $-Z_c$ yields

$$\mathbb{P}(|\widehat{\mu}_c - \mu_c| \geq t) \leq 2 \exp\left(-\frac{Q_c^2 t^2}{2 V_c} \right) = 2 \exp\left(-\frac{n_c^{\mathrm{e}} t^2}{2 \sigma_c^2} \right). \tag{26}$$

For $\delta \in (0, 1)$, set $t_c = \sigma_c \sqrt{2 \log(2C/\delta)/n_c^{\mathrm{e}}}$. The probability of exceeding t_c is at most δ / C for each class. A union bound therefore gives

$$\mathbb{P}(\exists c : \ |\widehat{\mu}_c - \mu_c| > t_c) \leq \delta. \tag{27}$$

A.4 Connecting Temporal Weight Aggregation to Effective Evidence

This subsection relates normalized temporal weight mass to the covariance-dependent evidence in Eq. (5). The analysis retains the fixed-weight model and assumes that responses from different recordings are uncorrelated, while allowing arbitrary temporal dependence within each recording.

Proposition 4. *For class c, let $S_{i,t}$ denote the scalar response at temporal position t of recording i, with mean μ_c and variance $\sigma_c^2 > 0$. Assume $\mathrm{Cov}(S_{i,t}, S_{j,u}) = 0$ for $i \neq j$. For fixed coefficients $b_{i,t} \in [0, 1]$, define $Q = \sum_{i:y_i=c} \sum_{t=1}^{T} b_{i,t} > 0$. Let b collect these coefficients and let Σ_c be the covariance matrix of the corresponding responses. If $b^\top \Sigma_c b > 0$, then*

$$n_c^{\mathrm{e}}(b) = \frac{\sigma_c^2 Q^2}{b^\top \Sigma_c b} \geq \frac{Q}{T}. \tag{28}$$

Proof. Let $Q_i = \sum_{t=1}^{T} b_{i,t}$, so that $0 \leq Q_i \leq T$ and $Q = \sum_{i:y_i=c} Q_i$. Cauchy–Schwarz gives $|\mathrm{Cov}(S_{i,t}, S_{i,u})| \leq \sigma_c^2$. Together with the absence of cross-recording covariance and the nonnegativity of the coefficients, this yields

$$\begin{aligned} b^\top \Sigma_c b &= \sum_{i:y_i=c} \sum_{t,u} b_{i,t} b_{i,u} \mathrm{Cov}(S_{i,t}, S_{i,u}) \\ &\leq \sigma_c^2 \sum_{i:y_i=c} Q_i^2 \leq \sigma_c^2 T Q. \end{aligned} \tag{29}$$

Substitution into the definition of $n_c^{\mathrm{e}}(b)$ gives $n_c^{\mathrm{e}}(b) \geq Q/T$. □

If the fixed coefficients also satisfy the tail condition in Eq. (8), Theorem 1 immediately gives

$$|\widehat{\mu}_c(b) - \mu_c| \leq \sigma_c \sqrt{\frac{2 \log(2C/\delta)}{Q/T}} \qquad \text{for all } c \tag{30}$$

with probability at least $1 - \delta$, where $\widehat{\mu}_c(b) = Q^{-1} \sum_{i:y_i=c} \sum_t b_{i,t} S_{i,t}$. This step requires the stated tail condition in addition to the covariance assumptions.

A.5 Limit of the Temporal Penalty Objective

Proof of Proposition 3. Fix recording i. Continuity of the margins and the positive-part function ensures that $\mathcal{V}_i$ is continuous. Hence, the feasible set $\mathcal{F}_i = \{a \in [0, 1]^T : \mathcal{V}_i(a) = 0\}$ is a closed subset of $[0, 1]^T$ and therefore compact. Since $\mathcal{F}_i$ is nonempty and s is continuous, there exists

$a^\star \in \mathcal{F}_i$ minimizing s over $\mathcal{F}_i$. Let $s_i^\star = s(a^\star) \in [0,1]$. For each $\eta > 0$, the continuous penalized objective also attains its minimum on $[0,1]^T$. Global optimality of a_η and feasibility of $a^\star$ imply

$$s(a_\eta) + \eta \mathcal{V}_i(a_\eta) \le s(a^\star) + \eta \mathcal{V}_i(a^\star) = s_i^\star. \tag{31}$$

Nonnegativity of both terms on the left yields

$$s(a_\eta) \le s_i^\star, \qquad 0 \le \mathcal{V}_i(a_\eta) \le \frac{s_i^\star}{\eta} \le \frac{1}{\eta}. \tag{32}$$

Let $\bar{a}$ be any accumulation point of a_η as $\eta \to \infty$, and choose a sequence $\eta_k \to \infty$ such that $a_{\eta_k} \to \bar{a}$. By Eq. (32) and continuity of $\mathcal{V}_i$, we have $\mathcal{V}_i(\bar{a}) = 0$, so $\bar{a} \in \mathcal{F}_i$. Continuity of s gives $s(\bar{a}) \le s_i^\star$, while feasibility of $\bar{a}$ implies $s(\bar{a}) \ge s_i^\star$. Thus, $s(\bar{a}) = s_i^\star$, proving constrained optimality.

The violation bound in Eq. (32) applies to globally optimized per-recording weight vectors at finite η. It does not guarantee exact feasibility at finite η, convergence to a unique vector, or global minimization by CETE's shared temporal head g_ψ, which is trained with finite η.

B Additional Experimental Details

This section supplements the experimental settings in the main text with details on dataset construction, baseline methods, and implementation. We describe the data partitions and evaluation protocols, followed by the baseline objectives, model architecture, audio preprocessing, and training procedures, including gradient updates and classwise statistics.

B.1 Datasets

- **VGGSound-LT.** We construct three long-tailed VGGSound (Chen et al., 2020) variants, retaining all 310 classes. Training clips are sampled without replacement using exponentially decreasing class frequencies and seed 42. Imbalance ratios of 10, 50, and 100 yield 121,333, 77,923, and 66,936 training clips, respectively. All variants share 3,100 validation and 12,363 test clips, with 10 clips per class held out from the source test partition for validation and the remainder for testing.
- **iNatSound.** iNaturalist Sounds (Chasmai et al., 2024), retaining all 5,569 training categories and the official splits of 137,012 training, 45,698 validation, and 49,527 test recordings. Validation and test each cover 1,212 classes, over which the respective macro-averaged metrics are computed. Each recording uses its official category annotation as the target, and window-level probabilities are averaged to obtain recording-level predictions.
- **BirdCLEF.** We use the public training collection of BirdCLEF (Klinck et al., 2025), containing 206 primary-label classes. A recording-level, class-stratified split with seed 42 yields 22,815 training, 2,882 validation, and 2,867 test recordings. The test set covers all 206 classes, comprising 74 Many, 75 Medium, and 57 Few classes. Each recording is supervised by its primary label. Secondary labels are excluded from supervision, while recordings containing them are retained.

Table 5: Class counts by frequency group and data split.

	Training			Validation			Test		
Dataset	Many	Medium	Few	Many	Medium	Few	Many	Medium	Few
VGGSound-LT, IR=10	93	123	94	93	123	94	93	123	94
VGGSound-LT, IR=50	55	199	56	55	199	56	55	199	56
VGGSound-LT, IR=100	47	214	49	47	214	49	47	214	49
iNatSound	301	548	4,720	299	511	402	299	511	402
BirdCLEF	74	75	57	74	75	57	74	75	57

Classes are grouped by their training recording counts n_c. For iNatSound and BirdCLEF, we adopt frequency thresholds of 100 and 20 training recordings following prior work (Kang et al., 2020), defining Many as $n_c > 100$, Medium as $20 < n_c \le 100$, and Few as $n_c \le 20$. VGGSound-LT has larger class counts at lower IRs, where these cutoffs would leave the Few group empty or nearly empty. We therefore define Many by $n_c > 500$ and Few by $n_c \le 200$, $n_c \le 40$, and $n_c \le 20$ for IR=10, 50, and 100, respectively; Medium contains the remaining classes. The Few thresholds

equal twice the minimum class count at each imbalance ratio. Table 5 reports the number of classes represented in each split and frequency group. For iNatSound, validation and test each omit two Many, 37 Medium, and 4,318 Few classes. BirdCLEF retains all 206 classes in each split. Frequency groups are determined from training counts and retained when evaluating each split.

B.2 BASELINES

The baselines in Tables 1 and 2 share data splits, audio preprocessing, and backbone initialization, with CE as the reference. Method-specific settings are detailed below.

Data-oriented methods. DODA selects classwise time masking, frequency masking, time shifting, their composition, or identity (Wang et al., 2024). Augmentation probability is 0.5, with preferences updated each epoch from changes in correct predictions. Maximum mask widths and absolute shifts are 96 frames, 16 bins, and 32 frames on VGGSound-LT, and 64 frames, 16 bins, and 16 frames on BirdCLEF and iNatSound. SEL fine-tunes the validation-selected CE linear classifier on frozen, unnormalized, temporally mean-pooled features using natural sampling (Jian et al., 2025). Each class-c example generates $k_c = \lceil (n_{\max} - n_c)/n_c \rceil$ features using training counts n_c and $n_{\max} = \max_c n_c$. Among the $\min(10, C_B - 1)$ nearest other classes by Euclidean distance between minibatch feature means, one is drawn uniformly, where C_B counts represented classes. Candidates are sampled with replacement in proportion to correct-class probabilities. Interpolation uses $z_e = (1 - r)z + rz'$, $r \sim \mathcal{U}(0, 1)$, retaining the target label. Real and synthetic CE receive weights 0.1 and 0.9. Synthetic CE is averaged within target classes, then across them. Without eligible pairs, SEL uses real CE.

Training-oriented methods. Focal Loss uses $\gamma = 2$ without class weighting (Lin et al., 2017). CB-CE uses $\beta = 0.9999$ and mean-one effective-number weights (Cui et al., 2019). LDAM uses maximum margin 0.5 and logit scale 30 without deferred reweighting (Cao et al., 2019). All three jointly train the encoder and classifier. Balanced Softmax (BS) adjusts training logits by class counts (Ren et al., 2020), while LogitAdjust uses prior adjustment with $\tau = 1$ (Menon et al., 2021). Both use the learned classifier logits directly at inference.

cRT retrains a new cosine classifier with balanced sampling. τ-norm rescales classifier weights with $\tau = 0.5$, preserving biases. LWS learns classwise logit scales initialized to one, fixing weights and biases (Kang et al., 2020). NCM averages unnormalized pooled features by class and compares normalized prototypes and queries by cosine similarity. LAS applies MiSLAS's concave label-aware smoothing from 0 for tail classes to 0.3 for head classes, freezing the encoder and batch-normalization statistics (Zhong et al., 2021). LOS uses $q_y = 1 - s + s(1/C + 0.05)$ and $q_{c \neq y} = (1 - q_y)/(C - 1)$ for C classes, with $s = 1$ on BirdCLEF and 0.5 otherwise (Sun et al., 2025). LWS, LAS, and LOS use balanced sampling, with cosine classifiers for LAS and LOS.

RBL jointly trains with CE and natural sampling (Gao et al., 2023). For feature dimension d, training class count C, and $m = \max(d, C)$, fixed prototypes E are the first C columns of $I_m - \mathbf{1}\mathbf{1}^\top/m$. Following the original Appendix E, Q comprises the first d rows of the matrix exponential $\exp(A - A^\top)$, with A a zero-initialized square matrix of order m. Unnormalized pooled features yield unit-scale, bias-free logits through QE. Matrix A uses the classifier learning rate. Validation and test subtract $\log(n_c / \sum_j n_j)$ from class-c logits, using training counts.

CC-SAM perturbs classifier parameters during joint training, with noise ratio 0.05, flat-loss coefficient 0.5, and class-weight divisor 1 (Zhou et al., 2023). TS-MOF freezes the CE encoder and trains four linear heads with CE, BS, LDAM, and LOS (Zhao et al., 2025). R-PLA uses the nonnegative cube root of each task's performance-vector cosine similarity to the LOS reference. RD-PCGrad follows the published gradient projection and aggregation rules. Inference weights head probabilities by a unit-temperature softmax of classwise validation accuracies. TS-MOF follows the paper equations where supplementary code differs. Both methods use natural sampling.

Shared training settings. All training stages use AdamW, batch size 64, weight decay 10^{-4}, and 100 epochs per stage. Learning rates are 10^{-4} for the encoder and 10^{-3} for classifier parameters. Classifier retraining freezes the encoder. Runs use seed 42 and validation balanced accuracy for checkpoint selection. τ-norm and NCM require no additional optimization.

Table 6: Temporal configurations by data split and class group.

Split	Class group	Condition	Occupancy (%)	Repetition (%)
Training	Affected	Occupancy sweep	20, 40, 60, 80	0
		Repetition sweep	80	0, 25, 50, 75, 87.5
	Reference	Fixed	80	0
Validation	All classes	Fixed	80	75
Test	All classes	Fixed	80	75

B.3 Implementation Details

The main comparisons share data splits, audio preprocessing, and joint-training settings. Methods requiring additional classifier training follow the protocol in Section B.2. This subsection describes the model architecture, audio preprocessing, optimization, and classwise statistics.

Architecture and Hyperparameters. In the reference configuration, CETE uses an ImageNet-pretrained ResNet-50, with the input convolution adapted to single-channel spectrograms by averaging its RGB kernels. Frequency averaging and temporal resampling of the final convolutional features produce 32 temporal features, each with 2,048 dimensions. Temporal mean pooling and a linear classifier yield recording-level predictions. The temporal head applies layer normalization, two linear layers with hidden dimension 128, GELU activation, and a sigmoid output, with parameters shared across temporal positions. It produces A_i, from which the target-class weights are extracted as $a_i = A_{i,:,y_i}$. Batch-normalization statistics and affine parameters remain fixed. The preservation tolerance and complementarity margin are $\varepsilon = \Delta = 0.02$. The teacher and classwise-statistic EMA momenta are $\mu = 0.99$ and $\beta = 0.97$, respectively. Temporal correction and class reweighting use $\alpha = \gamma = 0.35$. The remaining settings are $\tau_p = 0.50$, $\tau_m = 0.20$, $\eta = 10$, $\lambda_{\text{tw}} = 1$, and $\epsilon_w = 10^{-6}$; τ_p operates on class probabilities and τ_m on logit margins. Hyperparameter selection uses validation data only, with settings fixed for evaluation.

Audio Preprocessing. The reference pipeline uses mono audio sampled at 32 kHz. Evaluation divides each recording into non-overlapping 10-s windows, pads the final short window, and averages class probabilities across windows to obtain one recording-level prediction. For ResNet-50, the reference input is a 64-band log-mel spectrogram computed with a 1,024-sample Hann window, a 320-sample hop, and a frequency range of 50–14,000 Hz. Inputs are standardized using training-set statistics. Random gain is included in the common augmentation pipeline, with method-specific augmentation retained where required. The temporal feature count sensitivity analysis varies T while keeping recording duration and preprocessing fixed; resampling the feature sequence does not create additional independent observations.

Optimization and Training Procedure. CETE uses AdamW for 100 epochs with batch size 64 and weight decay 10^{-4}. Learning rates are 10^{-4} for the encoder, 10^{-3} for the classifier, and 10^{-4} for the temporal head. Backbone batch-normalization statistics remain fixed. Checkpoints are selected by validation balanced accuracy, with test labels excluded from model selection.

Teacher selection follows Eq. (14) and controls temporal-head training and classwise aggregation. All recordings retain their observed labels and contribute to the classification loss. The classification loss updates only the student encoder and classifier parameters θ, while the temporal loss updates only the temporal head parameters ψ. In the temporal branch, $a_i = g_\psi(\text{sg}(H_i))_{:,y_i}$, where sg denotes stop-gradient. The complementary representations also use detached encoder features. Gradients to the encoder and classifier parameters are blocked, while gradients with respect to a_i are retained through the classifier evaluations of $H_i^+(a_i)$ and $H_i^-(a_i)$. The teacher, v_c, and w_c receive no gradients. We initialize the classwise statistic as $v_c = n_c$ and use EMA momentum $\beta = 0.97$ for its updates. The temporal correction coefficient and class reweighting strength are $\alpha = 0.35$ and $\gamma = 0.35$, respectively. Class weights are normalized to mean one according to Eq. (17).

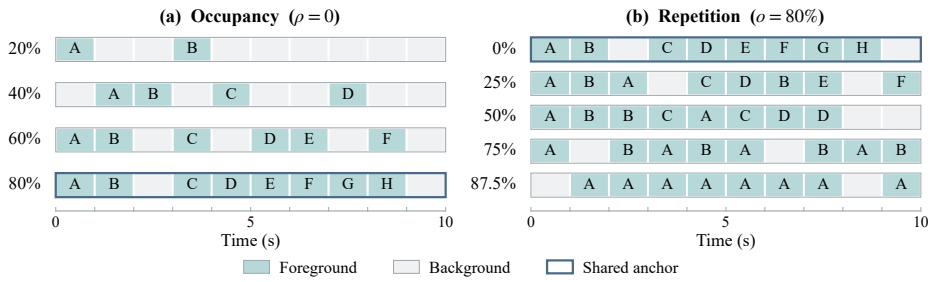

Figure 5: Controlled soundscape construction. Rows represent 10-s recordings with 1-s foreground slots over a continuous background. Repeated letters denote source reuse, and outlines mark the shared anchor. Panel (a) varies occupancy with no repetition, while panel (b) varies repetition at 80% occupancy. Slot positions are schematic, and occupancy includes pauses within inserted segments.

C Further Analysis

This section provides the controlled soundscape synthesis protocol and visualizations of the learned temporal weights. It also analyzes the computational and memory costs of training and inference.

C.1 Controlled Soundscape Synthesis

Each soundscape combines 1-s foreground segments from one of ten DESED categories with an FSD50K background excluding target-related labels (Fonseca et al., 2022). Recordings are 10-s mono audio at 16 kHz. Foreground segments are taken from the start of source recordings and placed in non-overlapping slots, with boundary fades and peak scaling. Each mixture uses a nominal SNR sampled uniformly from 0 to 10 dB. For K insertions from U distinct source segments, occupancy is $o = K/10$ and repetition is $\rho = 1 - U/K$. Occupancy includes pauses within segments.

Figure 5 and Table 6 summarize the temporal conditions. Blender, Dishes, Electric shaver/toothbrush, Frying, and Running water initially form the affected group, with the remaining classes serving as references. The groups exchange roles in a second assignment. Occupancy varies with $U = K$, whereas repetition varies at $K = 8$. Reference classes use $(o, \rho) = (0.8, 0)$, and the shared validation and test sets use $(0.8, 0.75)$. Corresponding conditions share backgrounds and nominal SNRs, with source exposure matched across repetition conditions. The training, validation, and test sets contain 560, 80, and 160 mixtures per class, drawn from 8, 2, and 2 foreground source groups, respectively.

CE, Focal, LDAM, and CETE use ResNet-50 under eight conditions and two class assignments with seed 9001. Inputs use 64-band log-mel spectrograms over 50–8,000 Hz, a 1,024-sample FFT and window, and a 320-sample hop. All models are trained for twenty epochs using AdamW, with a peak learning rate and weight decay of 10^{-4}. Training and evaluation batch sizes are 16 and 32. Cosine decay follows one warm-up epoch and ends at 0.01 times the peak learning rate. Batch-normalization statistics remain fixed. Focal uses $\gamma = 2$, and LDAM uses maximum margin 0.5 and logit scale 30.

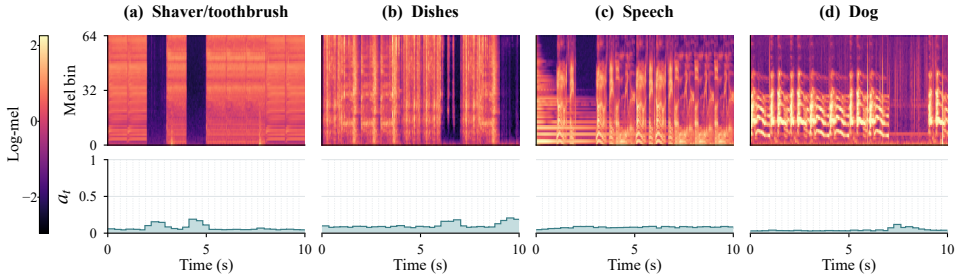

Figure 6: Temporal weights for randomly sampled test recordings, one per displayed sound class. Each column shows the recording's log-mel spectrogram and its class-conditional temporal weights.

Each reported result averages class recall across all ten sound categories and both assignments of the affected and reference groups.

C.2 Temporal Weight Visualization

Figure 6 presents temporal weights for held-out soundscapes from Section C.1. We randomly select one test recording from each of the four displayed sound classes, independently of classification correctness, margin values, and constraint satisfaction. Each panel pairs the selected recording's log-mel spectrogram with the temporal weights predicted for its labeled class.

The weights are generated by the validation-selected CETE checkpoint trained at the shared reference condition $(o, \rho) = (0.8, 0)$ under the first class assignment. Through the complementary interventions in Eq. (2), these weights act on the encoder features of each sampled recording.

C.3 Training and Inference Cost

CETE adds a temporal head and an exponential moving average (EMA) teacher during training. The analysis accounts for these auxiliary components and the student retained at inference.

Model Components and Parameter Storage. Let P_θ denote the number of parameters in the student encoder and classifier, and let P_ψ denote the number in the temporal head. The student and temporal head are trainable, while the EMA teacher stores an additional copy of the student parameters. CETE therefore has $P_\theta + P_\psi$ trainable parameters and stores $2P_\theta + P_\psi$ model parameters during training. These counts exclude optimizer states, activations, and non-parameter buffers. The class counts, running class statistics, and class weights occupy an additional $O(C)$ storage for C classes, independent of the training set size.

Extra training computation comes from teacher prediction, temporal-head evaluation and optimization, and classification of the two complementary representations. Both representations reuse the encoder output. The temporal objective updates only the head, propagating gradients through the classifier with its parameters and the encoder features fixed. Overhead depends on the backbone, teacher acceptance, and the implementation of these auxiliary operations.

Computational Overhead. At inference, CETE discards the teacher, temporal head, and classwise correction, retaining the student encoder and classifier with P_θ parameters. For the same student architecture and input shape, its inference computation matches that of CE, giving zero additional inference parameters and FLOPs. Latency and memory remain to be measured under matched conditions. Table 7 summarizes CETE's computational overhead relative to CE. For each backbone, training time and peak GPU memory are normalized by the corresponding CE values. Across the evaluated backbones, the relative increase in training time exceeds that in peak GPU memory.

Table 7: Computational overhead of CETE relative to CE with the same student architecture. Ratios are computed as CETE/CE. Zero additional inference parameters and FLOPs follow from retaining the student architecture.

Backbone	Training time ratio	Peak GPU memory ratio	Extra inference parameters	Extra inference FLOPs
ResNet-50	$1.40\times$	$1.17\times$	0	0
PANNs CNN14	$1.96\times$	$1.19\times$	0	0
AST-Base	$1.70\times$	$1.12\times$	0	0
BEATs-Base	$1.64\times$	$1.15\times$	0	0