Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliability-aware regional calibration for fine-grained bird sound recognition with a frozen BirdNET backbone

Hao Zhang
Zhejiang A&F University
Dayu Xu
Zhejiang A&F University
Si Shi
sishi6566@163.com

Zhejiang A&F University

Research Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11304920/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Reliability-aware regional calibration for fine-grained bird sound recognition with a frozen BirdNET backbone

Hao Zhang[a,1], Dayu Xu[a,1], Si Shi[b,*]

[a] *College of Artificial Intelligence, Zhejiang A&F University, Hangzhou 311300, China,*
[b] *School of Environmental & Resource Sciences, Zhejiang A&F University, Hangzhou 311300, China,*

Abstract

Background Passive acoustic monitoring produces large volumes of bird sound recordings, but automated species recognition for a regional checklist must handle sparsely represented species and provide a reliability score for each prediction, so that limited expert review effort can be allocated to uncertain cases.

Results We present RRC-BirdNET, a regional calibration framework built on a frozen BirdNET backbone. A lightweight window-score head projects acoustic embeddings into regional species scores and aggregates them through per-class-temperature log-mean-exp pooling (LME-PC); balanced-softmax training incorporates class-frequency information; and a confidence-based reliability gate provides a development-selected operating point for automatic reporting versus expert review. On 3929 provenance-audited recordings covering 276 species, the five-head ensemble achieves 84.05% top-1 and 79.94% balanced accuracy on 815 test recordings; individual heads achieve 83.51±0.34% top-1 across five seeds. On the 787-recording subset with direct BirdNET taxon mappings, the regional ensemble achieves 84.24% top-1, compared with 64.68% for checklist-restricted frozen scores. The reliability gate reaches an error-detection AUROC of 0.921; at a development target of 5% accepted-set error it retains 80.7% of test predictions with 95.6% observed accuracy. One head occupies 1.14 MB and its forward pass takes 0.54 ms per 96-window input on a CPU workstation.

Conclusions Compact regional adaptation of a shared acoustic representation supports species-level monitoring with selective reporting: the checklist readout improves over frozen global scores, the long-tail correction improves few-shot recall, and the reliability gate makes the trade-off between automation and expert review explicit and measurable. The framework runs locally without network connectivity, supporting deployment in field monitoring workflows.

Keywords: bioacoustics, passive acoustic monitoring, fine-grained recognition, regional calibration, long-tailed learning, selective prediction, CPU inference

*Corresponding author. E-mail: sishi6566@163.com.
[1] These authors contributed equally to this work and share first authorship.

1. Introduction

Passive acoustic monitoring provides repeated observations of bird communities across habitats and seasons. Species-level information is valuable for biodiversity assessment (IPBES, 2019) and for the management of wetlands along the East Asian–Australasian Flyway (Studds et al., 2017). Advances in acoustic recorders and deep learning enable these observations to be collected and analysed at increasing scale (Stowell, 2022). Bird vocalisations complement visual surveys, including observations of species active at night or within dense vegetation. Automated recognition can therefore extend the temporal coverage of field programmes while supporting expert interpretation (Kahl et al., 2021).

Among these models, BirdNET (Kahl et al., 2021) offers a widely used acoustic representation and a vocabulary of more than six thousand output taxa. It runs on commodity hardware, and is embedded in widely used analyser toolchains (Wood and Kahl, 2024; Pérez-Granados, 2023). Its analyser supports custom-classifier training on a fixed encoder (BirdNET team, 2024a,b), and additional workflow wrappers support site-specific custom-classifier training (Schiavo, 2025). These developments motivate studying how a shared acoustic representation can serve an explicit regional checklist with a compact, separately trainable decision layer.

Regional adaptation builds on this transferable representation. Global birdsong embeddings transfer effectively to downstream bioacoustic classification (Ghani et al., 2023), and regional BirdNET deployments can complement manual monitoring workflows (Ware et al., 2023). Closest to our setting, Márquez-Rodríguez et al. (Márquez-Rodríguez et al., 2025) combined a spectrogram-based bird-song detector with classifiers trained on BirdNET features for a 34-class Doñana wetland checklist. That pipeline combines detection and classification for a regional application. The present study examines a complementary configuration: retained sliding windows are classified jointly through recording-level pooling, with the BirdNET encoder held fixed throughout regional training.

For recording-level learning, weakly supervised aggregation of bird-sound windows is well studied (Ghaffari and Devos, 2025; Schlüter, 2021), recent BirdCLEF systems use model ensembles for bird-sound recognition (Sydorskyi and Gonçalves, 2025), and balanced-softmax training offers a principled per-class logit shift for long-tailed class distributions (Ren et al., 2020). Together, these approaches motivate learning a regional readout that combines temporal aggregation with class-frequency-aware training. The contribution of each design can be examined using a common embedding cache and a fixed recording-level split.

Reliability provides a further connection between classification and monitoring practice. Selective classification associates predictions with an acceptance decision (Geifman and El-Yaniv, 2017), and posterior confidence provides a useful signal for identifying cases for review (Hendrycks and Gimpel, 2017). This study

uses development data to fit a compact gate and select its reporting threshold, then measures coverage and accepted-set accuracy on held-out test recordings. The resulting operating points make the allocation of automatic reports and expert review explicit.

The evaluation adopts the focal-recording setting: each recording has a species label, and the model produces a recording-level prediction. Public collections such as Xeno-canto (Xeno-canto Foundation, 2026) support this form of regional classifier development. Continuous soundscape analysis also involves temporal localisation and potentially overlapping species (Kahl et al., 2021; Ghaffari and Devos, 2025). Fig. 1 illustrates these complementary recording contexts. Here, "regional" specifies an East Asian working species checklist; source recordings are drawn from public archives and may originate across each species' distribution. The checklist defines the prediction vocabulary, while the documented provenance and split policy define the evaluation population.

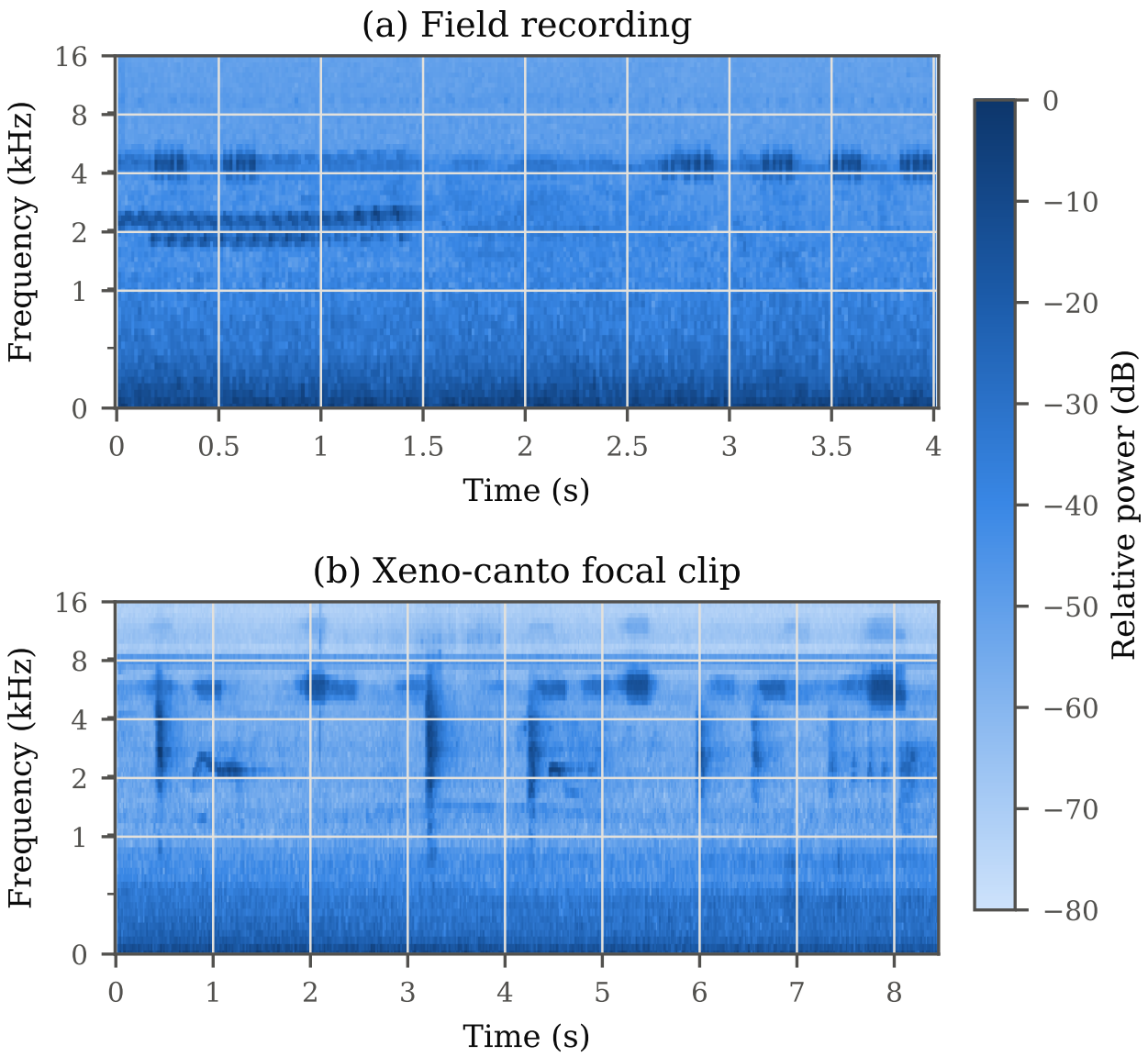


Figure 1. Illustrative field and archival recordings: (a) a locally supplied field recording, external to the evaluation corpus; (b) a focal clip recorded by Penghui Lou (XC1006321, CC BY-NC-SA 4.0). Mel spectrograms show up to 12 s of each recording, with a mel-scaled frequency axis and power relative to each clip's maximum on a shared −80 to 0 dB colour scale.

Four requirements guide the design of the regional framework:

1. **Checklist-aligned predictions.** The regional output layer should represent the target species explicitly. Frozen global scores and checklist-restricted scores provide reference points for measuring the value of regional training.
2. **Learning across recording frequencies.** Species contribute different numbers of public recordings. Training and evaluation should therefore account for both overall recognition and class-balanced performance, including species with few examples.

3. **Reliability-aware reporting.** A prediction and its confidence serve different purposes. A compact review gate should allow a monitoring workflow to select an operating point according to its review capacity.
4. **Traceable evaluation data.** Each retained recording should have verifiable label provenance, an explicit canonical taxon, and a documented split assignment. These properties support repeatable comparisons across regional models.

A further design goal is local processing without network connectivity; the measured computation here is on a CPU workstation.

The framework combines a regional window-score head with balanced-softmax training and a review gate selected on development data. Provenance verification supplies the data protocol. The three model components form RRC-BirdNET (Regional Reliability-aware Calibration of BirdNET); the data protocol supports their evaluation. Regional calibration here refers to adapting the decision layer and its reporting rule to the checklist. The reported metrics assess recognition and selective reporting. This application-oriented combination permits the regional model and review policy to be updated while retaining a shared frozen encoder.

The main contributions of this paper are as follows:

- A score-domain log-mean-exp head with one learnable temperature per species (*LME-PC*) reaches 84.05% top-1 as a five-seed ensemble on a 276-species regional checklist. Its individual head checkpoint occupies 1.14 MB.
- Controlled comparisons of pooling, training objective, and head capacity establish a compact configuration for the regional task. Balanced softmax raises few-shot recall from 73.4% to 79.3% on the scalar-LME head; the final LME-PC configuration uses a development-selected shift strength.
- A confidence-based review gate identifies misclassifications at AUROC 0.921. Its development-set-selected threshold gives an observed 4.4% test error among the 80.7% accepted predictions.

The remainder of this paper is organised as follows. Section 2 describes the model, dataset, experimental setup, and evaluation metrics. Section 3 reports comparative experiments, CPU component costs, visual results, ablation studies, and model-state analyses. Section 4 discusses the findings and their application to regional monitoring, and Section 5 concludes the paper.

2. Materials and Methods

2.1. Model overview

RRC-BirdNET adapts a frozen BirdNET representation to a regional species checklist through a compact classification head and a reliability gate (Fig. 2). The encoder retains its pretrained parameters, following the head-training customisation supported by the BirdNET ecosystem (BirdNET team, 2024a,b). Keeping this representation fixed makes regional heads inexpensive to retrain and provides a common basis for comparing aggregation and training objectives. All head comparisons use the same cached window embeddings and split; the raw-audio extraction protocol is specified in Section 2.5. Here, regional calibration denotes adaptation of the decision rule and review threshold to the checklist; probability calibration is a separate evaluation objective.

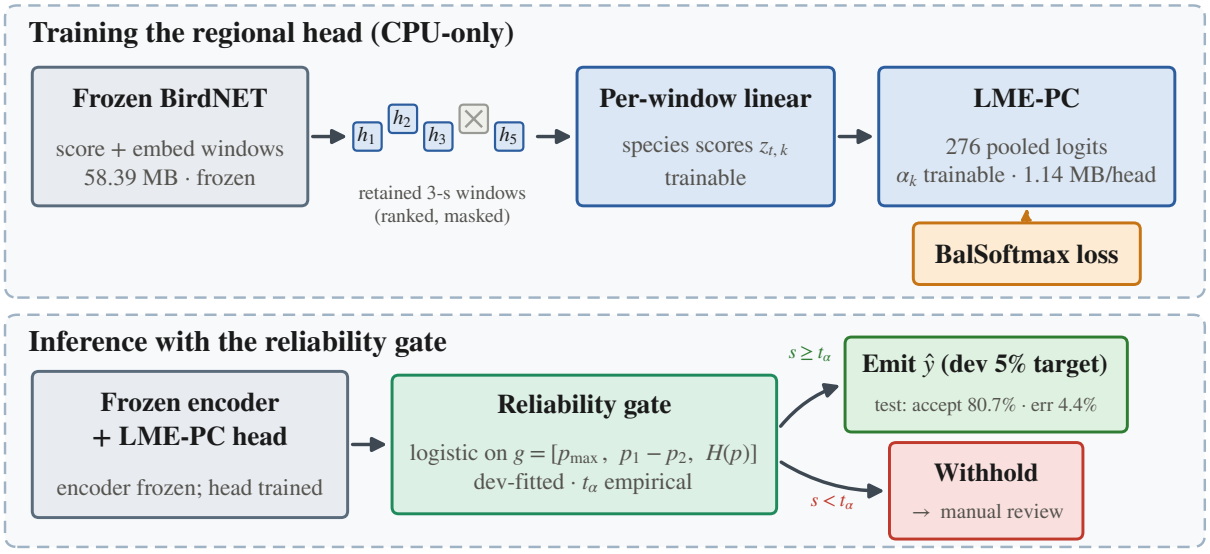


Figure 2. The overall architecture of RRC-BirdNET: frozen BirdNET embeds sliding windows, the LME-PC head pools their regional species scores under the balanced-softmax training objective, and a confidence gate flags low-scoring predictions for review.

The framework comprises three stages. First, BirdNET embeds overlapping 3-s windows with a 1.5-s hop and retains at most 96 windows per recording. Each retained vector has 1024 dimensions. Feature-wise means and standard deviations are estimated from the masked mean embedding of each training recording. The same affine transform, using a 1×10^{-6} stabiliser in the denominator, is then applied to every window in all splits. This gives recordings equal weight when estimating the normalisation statistics while preserving their individual windows. Second, a linear projection produces window scores for the 276 checklist species. Per-class-temperature log-mean-exp (LME-PC) pooling combines these into recording-level logits, and balanced-softmax training incorporates the training class frequencies. Third, a logistic reliability gate reads three

confidence features from the recording-level prediction and routes predictions according to a development-selected threshold. Cached embeddings connect these stages: the head can be retrained on recordings already represented in the cache, and the gate can be refitted from the corresponding head outputs.

The encoder and window-extraction protocol are fixed across regional-head experiments; the learned components are the regional head and the gate. For the five-seed ensemble, the recording-level logits of the separately trained heads are averaged before softmax. The predicted species is the largest ensemble logit, and the resulting softmax vector supplies the gate features.

2.2. Per-class-temperature window-score head (LME-PC)

A recording-level prediction must aggregate a variable number of overlapping windows. The extractor evaluates at most the first 200 windows of a recording and, when there are more than 96, retains the 96 windows with the highest maximum BirdNET output logit. Otherwise all valid windows are retained. Ranking by this encoder-derived score bounds the input size and prioritises windows with stronger global-model responses. Every recording proceeds to classification without a separate detection threshold. Let $\mathbf{h}_t \in \mathbb{R}^{1024}$ denote a normalised retained embedding and T the number of valid windows. A projection with $\mathbf{W} \in \mathbb{R}^{K\times 1024}$ and $\mathbf{b} \in \mathbb{R}^K$, where $K = 276$, produces species scores:

$$\mathbf{z}_t = \mathbf{W}\mathbf{h}_t + \mathbf{b}, \qquad t = 1, \ldots, T, \tag{1}$$

$$z_k = \frac{1}{\alpha_k}\left(\log \frac{1}{T}\sum_{t=1}^{T} e^{\alpha_k\, z_{t,k}}\right), \qquad \alpha_k = \mathrm{softplus}(\rho_k) > 0, \tag{2}$$

where one unconstrained parameter ρ_k is learned per species. Initialising $\rho_k = 0$ gives $\alpha_k = \log 2$. The implementation uses numerically stable log-sum-exp; padded entries receive a large negative score and the normalising count includes only valid windows. For positive α_k, the pooled score lies between the mean and maximum window scores. Its limiting forms are mean pooling as $\alpha_k \to 0$ and max pooling as $\alpha_k \to \infty$. The gradient makes the aggregation explicit:

$$\frac{\partial z_k}{\partial z_{t,k}} = \frac{\exp(\alpha_k z_{t,k})}{\sum_{u=1}^{T}\exp(\alpha_k z_{u,k})}. \tag{3}$$

Thus, each species has its own smooth weighting of the retained windows, with larger scores receiving greater influence. The temperatures describe aggregation sharpness; connecting them to vocalisation density would require window-level annotations. This formulation builds on learnable log-mean-exp pooling (Schlüter, 2021). Related audio-pooling comparisons (Wang et al., 2019) and attention aggregation (Ghaffari and

Devos, 2025) motivate the alternatives evaluated in Section 3.4. Fig. 3 contrasts LME-PC with mean pooling, and Fig. 4 details the module. The projection and temperatures are trained jointly with the objective below. Their parameter count is $K(1024+1)+K = 283176$; one head checkpoint occupies 1.14 MB.

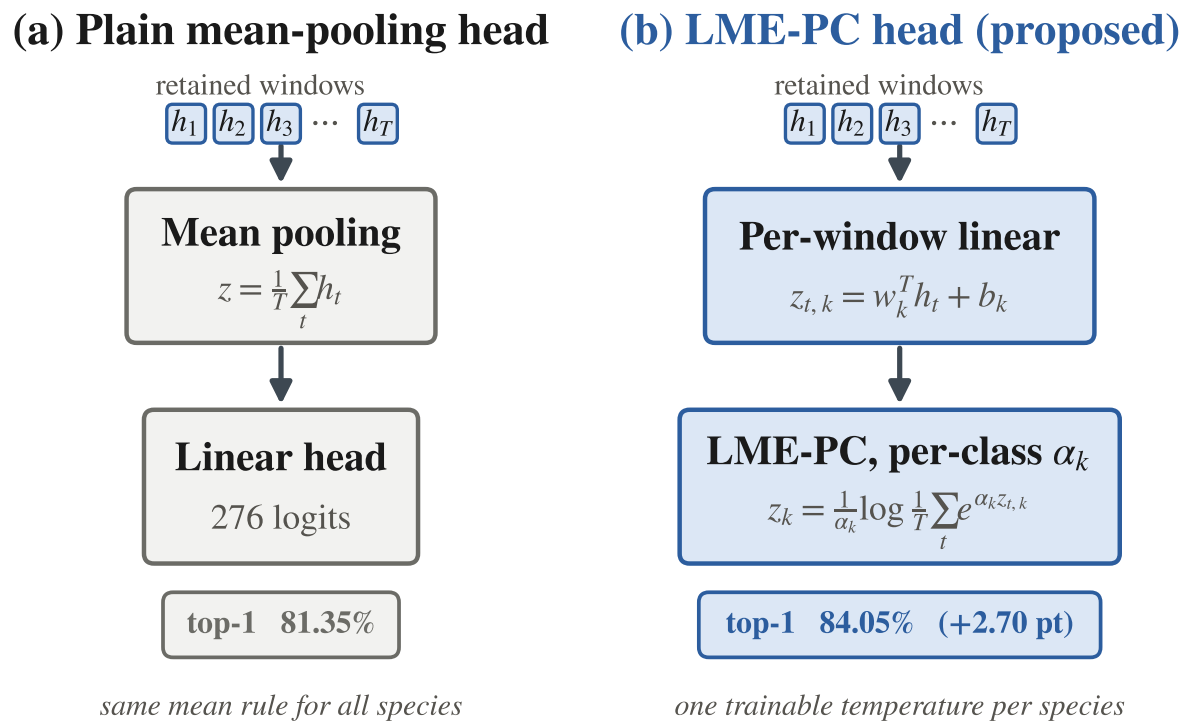


Figure 3. The head scheme comparison: (a) a plain mean-pooling head aggregates window embeddings equally before classification; (b) LME-PC projects window embeddings into species scores before log-mean-exp pooling, with one trainable temperature per species. The two designs differ in both pooling domain and temperature parameterisation. Displayed accuracies are from five-head ensembles.

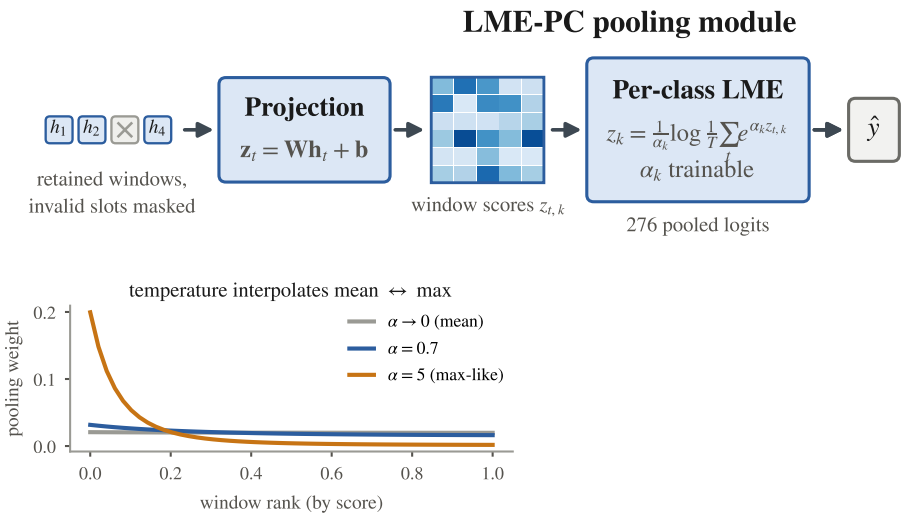


Figure 4. Detail of the LME-PC module: per-window species scores are pooled with a per-class trainable temperature; invalid window slots are excluded, and the temperature interpolates between mean and max pooling. The inset illustrates the aggregation rule using example score values.

2.3. Balanced-softmax long-tail training

To accommodate the class-frequency distribution of the audited corpus, we train the head with a strength-controlled balanced-softmax objective (Ren et al., 2020). Let n_k denote the number of training recordings for species k and $\pi_k = n_k / \sum_j n_j$ its training prior. Each class has training support. For a recording $\mathbf{x}$ with label y, the loss is

$$\ell(y \mid \mathbf{x}) = -\log \frac{e^{z_y + \tau \log \pi_y}}{\sum_{k=1}^{K} e^{z_k + \tau \log \pi_k}}, \tag{4}$$

where τ controls the log-prior shift. Setting $\tau = 0$ recovers cross-entropy and $\tau = 1$ gives the standard balanced-softmax form; development accuracy selects $\tau = 0.5$ for the final LME-PC head (Section 3.4). The shift appears only inside the training loss. Inference uses the unshifted logits $\mathbf{z}$, without a further post-hoc prior adjustment. During training, a frequent class receives a larger relative prior term, so the learned raw scores must account for this frequency contribution. The strength parameter controls this compensation while preserving the original training-recording sampling. The correction is computed once from the training counts and introduces no additional learned parameters or inference operations. The loss comparison in Section 3.4 examines its effect alongside cross-entropy and class-balanced weighting on a common head. Fig. 5 shows the underlying training distribution.

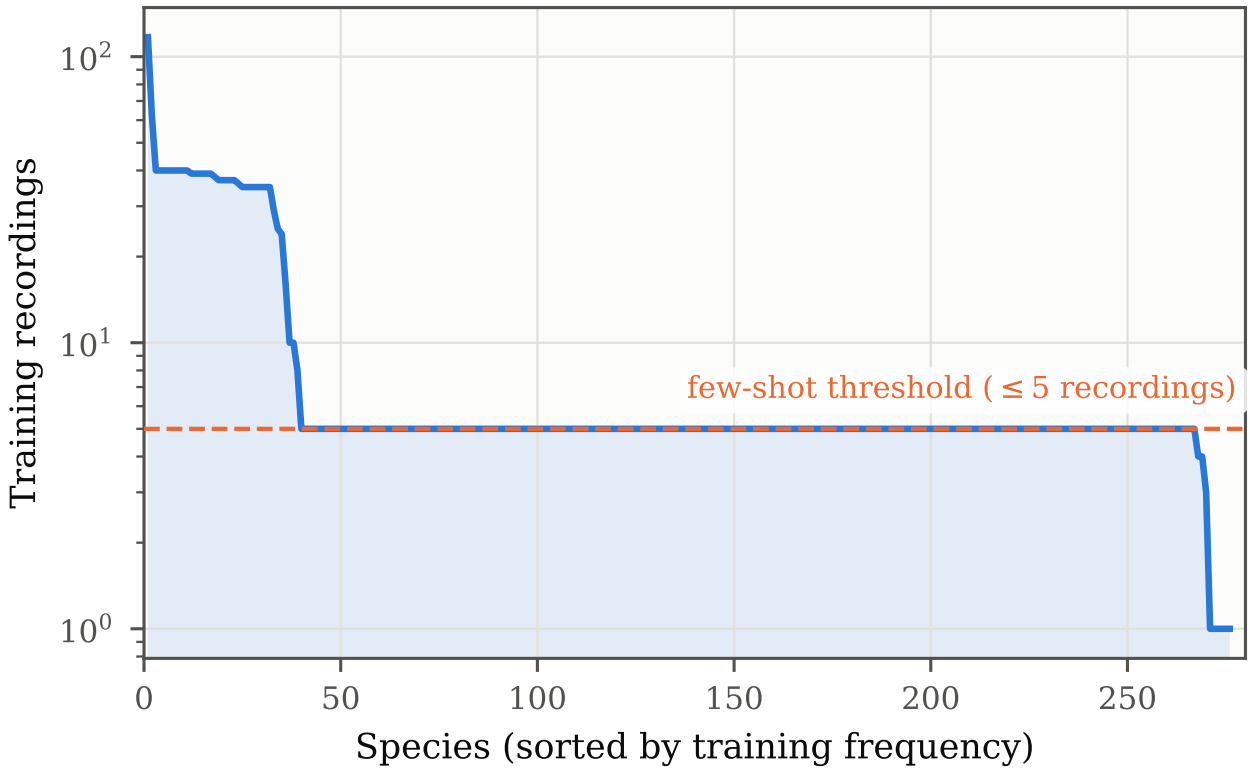


Figure 5. The class-frequency distribution of the audited training set on a logarithmic scale: 237 species have at most 5 training recordings.

2.4. Single-modality reliability gate

The reliability gate assigns each regional prediction to automatic acceptance or review. For the ensemble softmax vector $\mathbf{p}$, let $p_1 \geq p_2$ be its two largest entries and $p_{\max} = p_1$. The gate uses the peak probability, the top-two margin, and entropy:

$$\mathbf{g} = \left[\, p_{\max},\ p_1 - p_2,\ -\sum\nolimits_k p_k \log p_k \,\right]^\top, \tag{5}$$

These features summarise confidence level, separation between candidate species, and the spread of the class scores. Their means and standard deviations are fitted on the development set $\mathcal{D}$. Logistic regression then maps the standardised features to a decision score $s_i = \mathbf{a}^\top \tilde{\mathbf{g}}_i + b_g$, using correctness $\mathbf{1}[\hat{y}_i = y_i]$ as the binary training label (Hendrycks and Gimpel, 2017; Geifman and El-Yaniv, 2017). The implementation uses L_2 regularisation with inverse strength $C = 1$ and up to 1000 solver iterations. All inputs derive from the audio head's output, which keeps the gate inexpensive at inference.

The descriptive fixed-threshold analysis accepts $s_i \geq 0$, equivalent to a logistic model output of at least 0.5. For selective prediction, the same development records used to fit the gate also select its threshold. Given a target empirical accepted-set error α, the selected threshold is the lowest observed development

score satisfying

$$t_\alpha = \min\left\{ t \in \{s_i : i \in \mathcal{D}\} : \frac{\sum_{i\in\mathcal{D}} \mathbf{1}[s_i \geq t,\, \hat{y}_i \neq y_i]}{\sum_{i\in\mathcal{D}} \mathbf{1}[s_i \geq t]} \leq \alpha \right\}, \tag{6}$$

where thresholds with a nonempty accepted set are considered. This rule maximises development coverage among the candidate thresholds satisfying the empirical criterion. The fitted gate and selected threshold are then fixed before evaluating the held-out test set. Because fitting and threshold selection share development records, α specifies an empirical operating target; its performance on new recordings is assessed through the observed test coverage and accepted-set error in Section 3.5. Predictions below the threshold are withheld from the accepted subset and form a review queue. This makes coverage directly interpretable as the fraction processed automatically, with the remaining fraction specifying the review workload. Fig. 6 summarises the module.

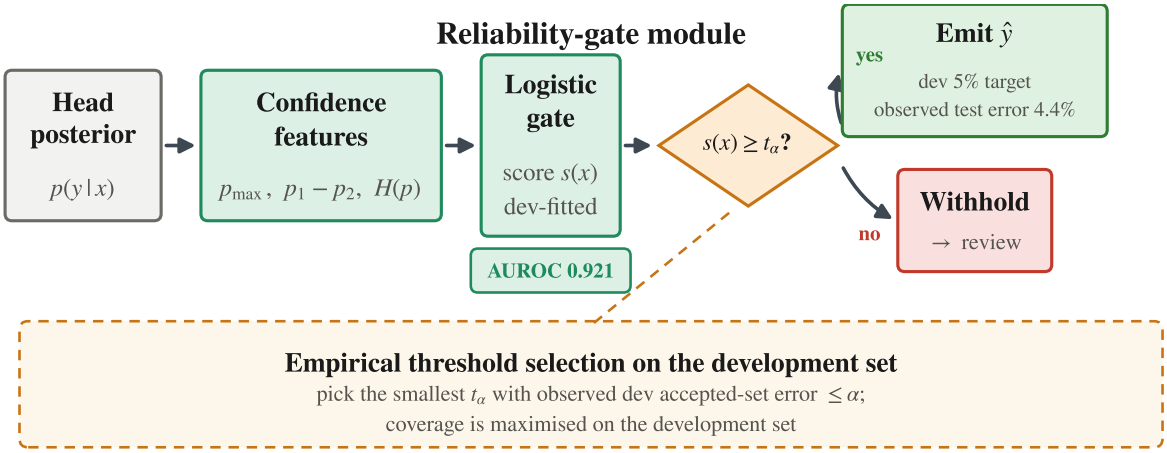


Figure 6. Detail of the reliability-gate module: three confidence features are scored by a logistic gate, and a threshold t_α selected on the development set decides between accepting and flagging a prediction.

2.5. Dataset

The corpus was assembled from public archives (Xeno-canto (Xeno-canto Foundation, 2026) and a regional birdsong library), originally 6642 recordings with 287 observed labels from a 295-entry East Asian working checklist. Because the assembly ingested multi-source benchmark material through English common-name keyword mapping, we audited every file with two auditable provenance signals rather than a ground-truth re-labelling exercise: filename provenance (Xeno-canto and library files carry the uploader's Latin species name, while benchmark-sourced files carry English common-name text) and BirdNET-logit consistency (cached BirdNET species logits compared against the file's own name, used neither as reference labels nor as a filtering criterion). The audit classified 2499 of the 6642 files as benchmark-sourced, including 1529 whose SSW filenames contain no species name; another 70 files lacked a parseable name under the audit parser.

BirdNET-score consistency provided a diagnostic view for the provenance review; retention followed the documented source and taxonomy rules.

We therefore adopted a single documented cleaning policy with four rules: a recording is kept only when (i) its filename provenance is verifiable (Xeno-canto and library files carry the uploader's Latin species name; benchmark-sourced files are admitted only when their embedded English name maps to the class-majority Latin), (ii) its parsed scientific name equals the project-local canonical taxon of its stated species, (iii) its SHA-256 content is unique and label-consistent, and (iv) it is assigned to exactly one split by its SHA-256 group. The policy retains 3929 recordings over 276 species (train/dev/test = 2591/523/815). The ledger assigns the remaining 2713 files to explicit provenance, taxonomy, duplicate, or label-consistency categories (Table 1, Fig. 7). The split unit is a SHA-256 file-hash group: all retained windows from a file belong to one split, following the recording-level evaluation principle discussed by Rauch et al. (2025). Window extraction follows a fixed rule with recording-level labels. SHA-256 identifies byte-identical files; source identifiers provide an additional level of identity across differently encoded archive copies.

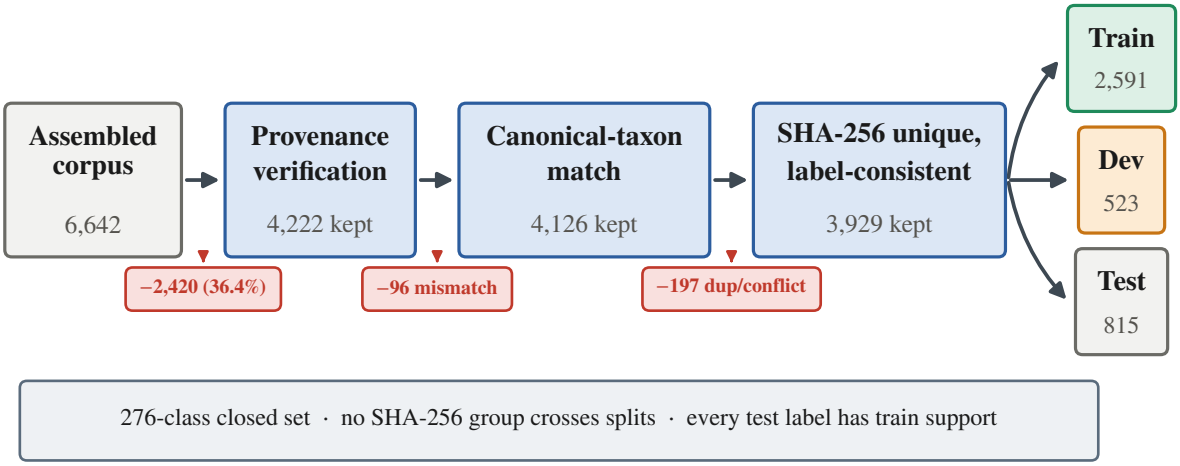


Figure 7. The curation pipeline: provenance verification (6642→4222), exact canonical-taxon matching, SHA-256 deduplication, and the resulting 276-class split. Each retained file-hash group belongs to one split.

Table 1. Audio corpus audit and cleaning policy. The single documented cleaning policy comprises: provenance verification (Latin-name files plus benchmark files whose embedded English name maps to the class-majority Latin), exact canonical-taxon match, SHA-256 deduplication, and a class-stratified file-hash-group split and train support for every test label.

Cleaning step	Recordings	Note
Assembled corpus	6642	287 labels; 295 listed
Cleaning policy retained	3929	276 classes; SHA-256-distinct files
provenance review	2420	Outside documented retention rules
taxonomy mismatch	96	canonical taxon differs
duplicate SHA-256	162	exact content duplicates
SHA group label conflict	35	same audio, conflicting labels

The retained corpus covers 276 of the 295 regional checklist species with an average of 9.4 training recordings per species, and its class distribution is markedly long-tailed (Fig. 5): 237 species have at most five training recordings, yet these few-shot species still contribute 463 of the 815 test recordings, so the evaluation probes the tail directly and therefore gives substantial weight to sparsely represented species. Filename and canonical-taxon checks supply a traceable label provenance for every retained file. These checks document source-label consistency; the acoustic labels remain those supplied by the source archives.

A supplementary source-identity sensitivity check sets aside two test records designated as candidate repeated-source recordings in the local audit. Applying the existing five-head ensemble to the resulting 813-recording set gives 84.01% top-1 and 79.94% balanced accuracy, compared with 84.05% and 79.94% on the full test set. This is a test-set sensitivity analysis with the saved checkpoints; their training and development selection remain those of the original split. The main tables use the fixed 815-recording test population to preserve comparability across all configurations.

2.6. Experimental setup

Table 2 summarises the training and evaluation configuration. All experiments run on a single CPU workstation (no discrete GPU). The frozen encoder is BirdNET V2.4 (Kahl et al., 2021; BirdNET team, 2024a), applied to 3-s windows with a 1.5-s hop and at most 96 retained windows per recording under the cap described above. Window features are z-scored with pooled-recording statistics fitted on the training split. The head projects each window to species scores with Eq. 1 and pools them with the per-class-temperature LME of Eq. 2. We train the head with the balanced-softmax objective of Eq. 4 (τ selected on the development set) using AdamW (learning rate 10^{-3}, weight decay $3 \cdot 10^{-4}$), batch size 256, up to 150 epochs with early stopping (patience 20), and five fixed seeds (1–5); the balanced-softmax objective alone carries the long-tail correction, with no mixup, class sampling, or logit adjustment. The reliability gate is a logistic regression on the three confidence features of Eq. 5, fitted on the development split, with the empirical threshold of Eq. 6 selected on that same split. The same embedding cache, labels, and recording-level splits are used for the regional-head comparisons. Checkpoint selection follows development top-1 accuracy, and the five selected checkpoints contribute equally to the final logit average. The frozen-output references use a separate global-score cache with at most 16 windows per recording; regional heads use up to 96 cached embeddings. The comparison with frozen scores therefore measures the complete evaluated configurations on common recordings. Comparisons among regional heads hold the embedding input fixed. The final linear LME-PC head has no active dropout. Mean, scalar-LME, and gated-attention heads apply dropout with probability 0.3 to pooled embeddings; the deeper LME-PC variants apply dropout within their trunks. Pooling comparisons therefore include these regularisation differences.

Table 2. Training and evaluation configuration. Development data determine checkpoint selection and the gate operating point; test data provide the reported evaluation.

Item	Setting
Encoder and embedding dimension	Frozen BirdNET V2.4; 1024
Window length / hop	3 s / 1.5 s
Regional-head input	At most 96 valid windows; padded slots masked
Normalisation	Training-only pooled-embedding mean and standard deviation
Optimiser	AdamW; learning rate 10^{-3}; weight decay 3×10^{-4}
Training schedule	Batch 256; at most 150 epochs; patience 20
Final loss	Balanced softmax, $\tau = 0.5$
Repeated training	Seeds 1–5; equal logit averaging
Reliability model	Standardised confidence features; logistic regression
Gate fitting and threshold selection	Same 523 development recordings

The classical classifiers use the same mean-pooled, standardised embeddings. Logistic regression uses an iteration limit of 2000; the linear SVM is fitted by stochastic gradient descent with hinge loss and an iteration limit of 3000. The nearest-neighbour classifier uses $k = 5$, and the random forest uses 500 trees. These reference configurations complement the five-seed neural-head comparisons by characterising alternative readouts of the shared representation.

2.7. Evaluation metrics

Recognition performance is measured with three standard metrics: (1) *Top-1 accuracy*, the fraction of test recordings whose predicted class equals the ground-truth species:

$$\text{Top-1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i], \tag{7}$$

where N is the number of test recordings and $\hat{y}_i$ the predicted species of recording i; (2) *Balanced accuracy*, the mean of per-class accuracies over the species represented in the test set:

$$\text{Balanced} = \frac{1}{|\mathcal{K}_{\text{test}}|} \sum_{k \in \mathcal{K}_{\text{test}}} \frac{\sum_{i:y_i=k} \mathbf{1}[\hat{y}_i = y_i]}{|\{i : y_i = k\}|}, \tag{8}$$

where $|\mathcal{K}_{\text{test}}|$=273 of the 276 output species have test support; (3) *Macro-F1*, the unweighted mean of F1 over the same supported species. The vocabulary-aligned subset contains 787 recordings and 261 test-supported species. For the classical classifiers, stored macro-F1 values averaged over all 276 outputs are converted to the common 273-species support by multiplying by 276/273; the three unsupported classes have zero F1. For the reliability gate we additionally report *coverage* (the fraction of predictions accepted), *accepted-set error* (the error fraction within the accepted subset), *error recall* (the fraction of all errors flagged), and the threshold-free AUROC and average precision of the gate score treated as an error detector. Together these quantities describe both the operating point a practitioner selects and the gate's ranking quality across all

thresholds. Efficiency is reported with separate units for the frozen encoder and the trained heads: on-disk weight size, single-thread CPU latency per 3-s BirdNET window or per 96-slot head input, and the linear projection FLOPs of the head. These component measurements exclude audio decoding, resampling, window construction, normalisation, and gate latency. For a selected threshold t, coverage and accepted-set error are

$$C(t) = \frac{\sum_i \mathbf{1}[s_i \geq t]}{N}, \qquad R(t) = \frac{\sum_i \mathbf{1}[s_i \geq t, \hat{y}_i \neq y_i]}{\sum_i \mathbf{1}[s_i \geq t]}. \tag{9}$$

Accepted-set accuracy is $1 - R(t)$, and the review fraction is $1 - C(t)$. All operating points are selected on development data before evaluation on the test set. Seed summaries use the mean and population standard deviation of the five runs. The pooling and training-objective sensitivity plots display these seed means with one-standard-deviation error bars; ensemble metrics are reported separately in the tables.

3. Results

3.1. RRC-BirdNET for regional bird sound recognition

Table 3 compares the evaluated configurations on the same 815 test recordings. The global BirdNET reference averages its cached window scores and selects an output from 6522 taxa. It achieves 52.6% top-1; 34.2% of predictions have no unique mapping to the regional checklist. Restricting the same scores to mapped checklist taxa yields 62.5%, a 9.8-percentage-point increase under an otherwise unchanged frozen-score protocol. This reference quantifies the contribution of an explicit output vocabulary.

Table 3. Regional calibration on the audited test set (276 output species, 273 with test support; recording-level split, sliding windows). *BirdNET, global argmax* uses the frozen model's 6522-taxon output with no trained head; *BirdNET, checklist-restricted* uses the same frozen scores restricted to the candidate taxa with its pretrained output layer. Unless marked as a per-seed mean, trained-head rows are five-seed logit ensembles with balanced-softmax (BalSoftmax) training at τ=0.5. Balanced accuracy and macro-F1 average over the 273 test-supported species. Frozen-score and regional-head inputs use the respective window caches described in Section 2.6. Bold type identifies the selected configuration.

Method	Top-1 (%)	Balanced (%)	Macro-F1 (%)
BirdNET, global argmax	52.6	49.1	51.7
BirdNET, checklist-restricted	62.5	60.2	55.7
Mean head + BalSoftmax	81.35	76.15	75.47
Scalar-LME head + BalSoftmax	82.94	78.24	77.69
Gated-attention head + BalSoftmax	84.66	80.80	80.48
LME-PC, per-seed mean ± std	83.51 ± 0.34	79.31 ± 0.48	79.02 ± 0.70
LME-PC, five-head ensemble	84.05	79.94	79.92
Multi-pool stacking	84.79	80.43	80.09

The regional heads learn a readout of the frozen embeddings. The mean-pooling ensemble achieves 81.35% top-1, and the LME-PC ensemble achieves 84.05% top-1, 79.94% balanced accuracy, and 79.92% macro-F1. The 2.70-point difference from mean pooling is measured on a common embedding cache. The

21.60-point difference from checklist-restricted BirdNET describes the full evaluated pipeline, including its distinct window cache. Gated attention and multi-pool stacking attain 84.66% and 84.79%, respectively. These configurations provide complementary accuracy references; LME-PC provides a compact score-domain readout with 276 pooling parameters beyond its linear species projection.

Vocabulary coverage is examined separately in Table 4. The directly mapped subset contains 787 recordings whose reference species has a BirdNET output taxon. RRC-BirdNET achieves 84.24% top-1 on this subset, compared with 64.68% for the checklist-restricted reference. Each system retains its original output space and window protocol; subsetting aligns the reference species that can be represented by BirdNET. The result supports regional adaptation across both the full checklist evaluation and the directly mapped subset. Fig. 8 complements the top-1 results with a micro-averaged precision–recall comparison.

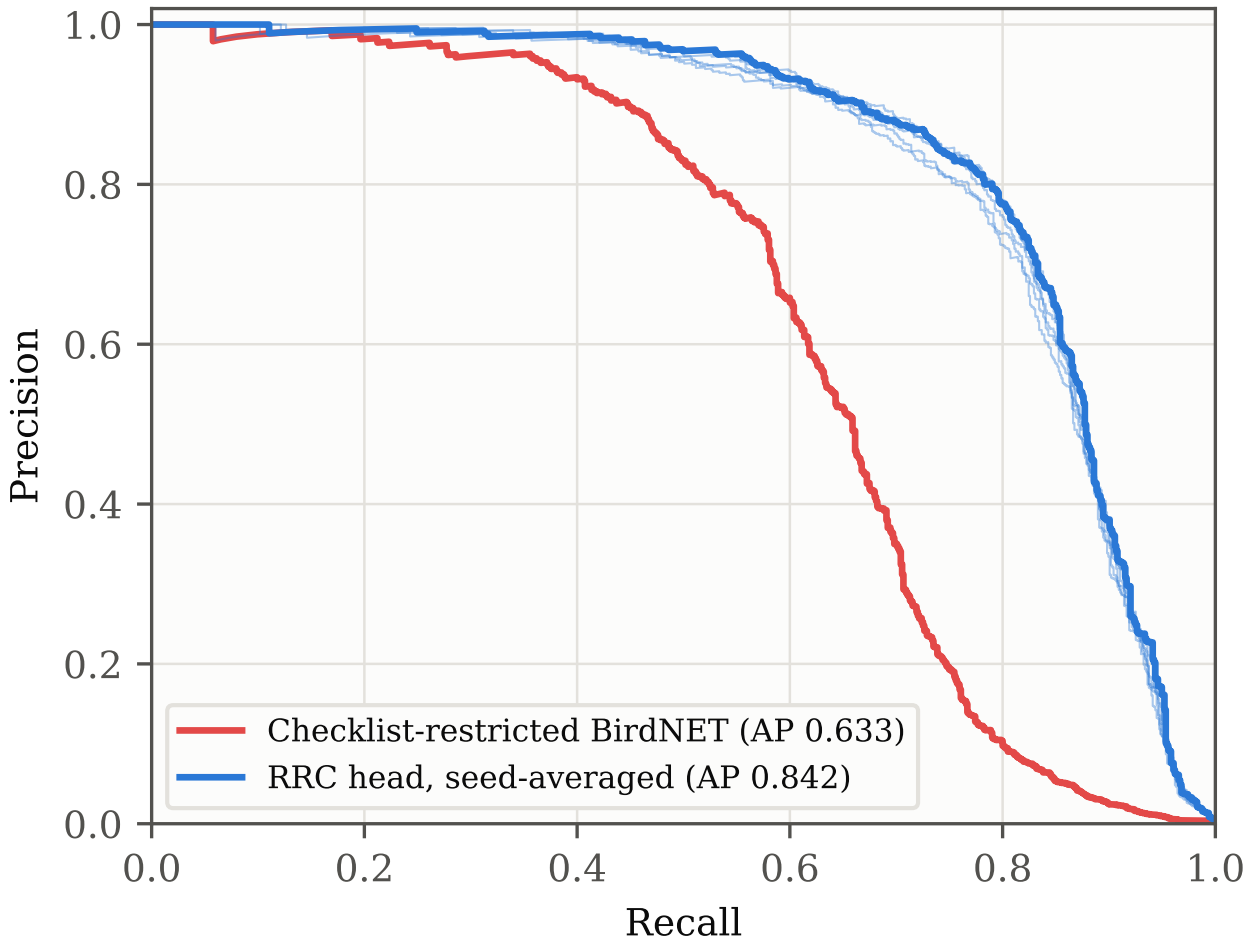


Figure 8. The PR curves of the regional head and the checklist-restricted frozen BirdNET on the audited test set (micro-averaged; thin lines = the five individual seeds, bold = seed-averaged).

Table 4. Evaluation on the 787-recording directly mapped reference-label subset (261 test-supported species). Output vocabularies and window protocols remain those of the corresponding full-set configurations.

Method	Top-1 (%)	Balanced (%)	Macro-F1 (%)
BirdNET, global argmax	54.51	51.31	54.53
BirdNET, checklist-restricted	64.68	62.96	59.27
RRC-BirdNET, five-head ensemble	84.24	80.59	80.65

Table 5 reports every LME-PC training run. Single-head top-1 accuracy ranges from 82.94% to 83.93%, with a mean of 83.51% and a population standard deviation of 0.34 percentage points. The five-head ensemble correctly classifies one more test recording than the best individual seed. Reporting the individual results alongside the ensemble provides a direct view of the accuracy available at each storage and inference budget (Lakshminarayanan et al., 2017).

Table 5. Per-seed LME-PC performance and the five-seed logit ensemble on the audited test set. Standard deviations describe the five observed training runs; the top-1 range is 0.99 percentage points.

Seed	Top-1 (%)	Balanced (%)	Macro-F1 (%)
1	83.31	79.18	78.85
2	83.68	79.54	79.59
3	83.93	79.96	79.78
4	82.94	78.49	77.78
5	83.68	79.36	79.10
Mean $\pm$ std	83.51 ± 0.34	79.31 ± 0.48	79.02 ± 0.70
Ensemble	84.05	79.94	79.92

The classical readouts in Table 6 further characterise the frozen representation. Logistic regression achieves 79.51% top-1, close to the 80.12% of the cross-entropy mean-head ensemble. Together, these results support a linear regional readout as a useful starting point. The SVM, nearest-neighbour, and random-forest references contextualise the chosen readout under their stated training configurations. The subsequent ablations examine temporal aggregation, class-frequency treatment, and additional head capacity.

Table 6. Classical classifier baselines trained on the audited training split and evaluated on the same mean-pooled, z-scored test embeddings. A trained linear head with plain cross-entropy reaches 80.12% top-1 on the same pooled features. Macro-F1 is harmonised to the 273 test-supported species.

Classifier	Top-1 (%)	Balanced (%)	Macro-F1 (%)
Logistic regression	79.51	72.87	73.18
Linear SVM (SGD)	73.50	64.59	66.41
k-NN (k=5)	56.93	47.26	47.80
Random forest (500 trees)	53.13	35.92	38.09

3.2. CPU component measurements

The components run on a CPU workstation without a discrete GPU. Table 7 reports measurements under explicitly different units: BirdNET processes pre-built 3-s windows in batches of 16, while a head processes one recording represented by 96 window slots. The frozen BirdNET SavedModel occupies 58.39 MB

excluding labels; the five-head ensemble corresponding to the 84.05% result adds 5.69 MB including shared preprocessing statistics. A single head occupies 1.14 MB and has a five-seed mean top-1 of 83.51%. The measured 46.22 ms per BirdNET window is batch-16 amortised throughput on a single CPU thread, which characterises batched archival processing. The batch-one measurement is 626.77 ms per window. Reporting both settings distinguishes individual-window latency from amortised batch throughput. The measured stages are neural forward passes on pre-built inputs; audio decoding, resampling, window construction, normalisation, and gate evaluation have separate runtime costs.

Table 7. Component CPU costs on an Intel Core Ultra 5 225, one thread, median after warm-up. BirdNET timing includes both embedding and basic output signatures on pre-built 3-s windows. Head timing uses five saved checkpoints on a real 96-window test recording and excludes preprocessing; its FLOPs count only linear projections. Units differ by row, and no complete-pipeline latency is inferred.

Component	Unit	Size (MB)	Linear GFLOPs	Latency (ms)
BirdNET V2.4	3-s window (batch 1)	58.39	–	626.77
BirdNET V2.4	3-s window (batch 16)	58.39	–	46.22
LME-PC, one head	recording (96 slots)	1.14	0.054	0.54
LME-PC, five heads	recording (96 slots)	5.69	0.271	2.85

3.3. Visual results

Fig. 9 compares the predictions of RRC-BirdNET and the frozen BirdNET global classifier on the audited test set. Each panel displays the cached per-window BirdNET probabilities for its three highest-ranked global taxa; the window index enumerates the cached score windows retained for the recording. This diagnostic uses the global-score cache's at most 16 windows; the regional head uses a separate up-to-96-window embedding cache. Species names emitted by the two systems are coloured green for agreement with the reference label, red for a different species, and grey for an output without a regional mapping. These global-score displays provide visual context for the two systems' predictions on shared input, complementing the regional head's decision on the embedding cache. Panels (a) and (b) show reference-label agreement for both systems. Panels (c) and (d) show regional-head predictions that agree with the reference label alongside global outputs outside the mapped checklist. Panel (e) illustrates a within-genus alternative, and panel (f) illustrates selection for expert review using the fixed 0.5-probability gate threshold. These examples connect the aggregate results to individual recordings and their temporal score patterns.

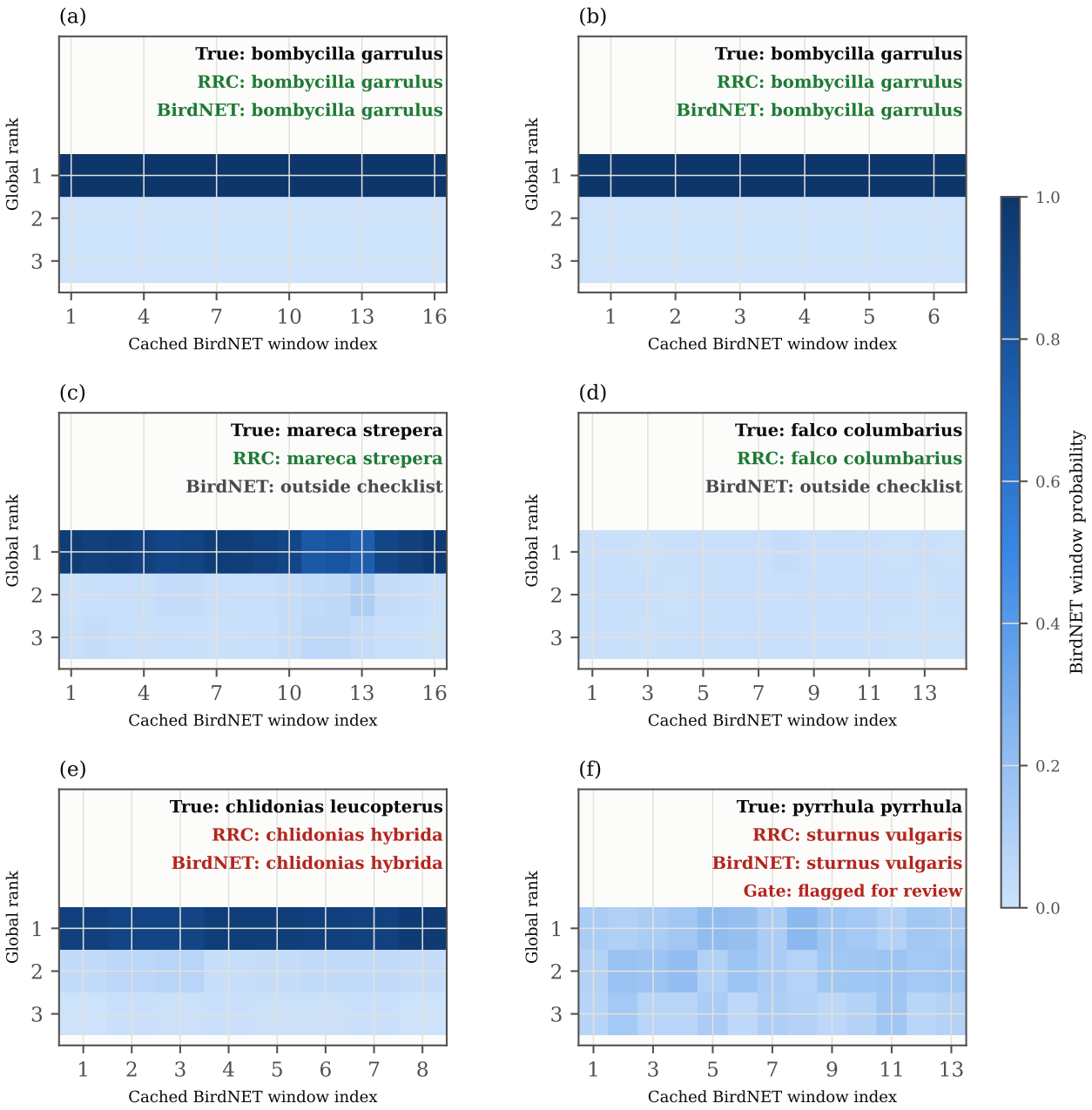


Figure 9. Visualisation of RRC-BirdNET versus the frozen BirdNET global classifier on audited test recordings. Heatmap rows show the top three global BirdNET taxa by mean logit; colour gives their softmax probability in each cached score window. Green: reference-label agreement; red: alternative species; grey: outside the mapped checklist. Panel (f) illustrates a prediction selected for review.

The visual analysis serves two purposes. It shows how a fixed global representation can support a different regional decision, and it makes the review operation visible at the recording level. The displayed global probabilities and the regional head have distinct roles: the heatmaps describe the global-score cache, while regional decisions use the embedding cache. Thus, the panels provide prediction context rather than learned attention maps of LME-PC. Their illustrative selection complements the exhaustive test-set metrics

in the comparison tables.

3.4. Ablation study

Table 8 shows a sequence of regional-head configurations on the frozen backbone. The mean-pooling cross-entropy and balanced-softmax rows isolate the loss change. The later rows compare pooling designs but change more than one architectural detail; the gate is evaluated separately below.

Table 8. Regional-head configuration sequence on the audited test set (five-seed ensembles). The mean and scalar-LME checkpoints contain genus/family projection parameters, with hierarchical loss disabled for every run so that all rows share the same single prediction path. LME-PC changes pooling domain as well as temperature parameterisation.

Pooling	Training loss	Top-1 (%)	Balanced (%)
Mean embeddings	Cross-entropy	80.12	73.90
Mean embeddings	BalSoftmax τ=0.5	81.35	76.15
Scalar LME on embeddings	BalSoftmax τ=0.5	82.94	78.24
LME-PC on class scores	BalSoftmax τ=0.5	84.05	79.94

With balanced-softmax strength fixed at τ=0.5, Table 10 and Fig. 10 compare five pooling designs. LME-PC improves on mean and embedding-domain scalar LME, while gated attention (Ghaffari and Devos, 2025) is 0.61 pt higher in top-1. The LME-PC pooling parameterisation adds 276 scalars to the linear species projection, yielding a compact implementation. Raw-score weighting achieves 63.31% top-1 and 57.77% balanced accuracy under the evaluated rule. That rule normalises signed BirdNET logits as weights; LME-PC uses positive exponential contributions to regional class scores. These results favour learned mean, attention, and LME aggregation within the present comparison. The class-wise LME temperatures are examined separately in Section 3.5. Table 9 and Fig. 11 compare loss strategies on the *scalar-LME* head: τ=1 improves balanced accuracy by +4.6 pt over cross-entropy on that head. For the final LME-PC head, development accuracy selects τ=0.5 (87.6% versus 87.0% for τ=1). On the scalar-LME head, both tested balanced-softmax strengths improve top-1 and balanced accuracy relative to cross-entropy. On LME-PC, development selection favours the half-strength shift. This distinction keeps the loss comparison tied to its pooling configuration.

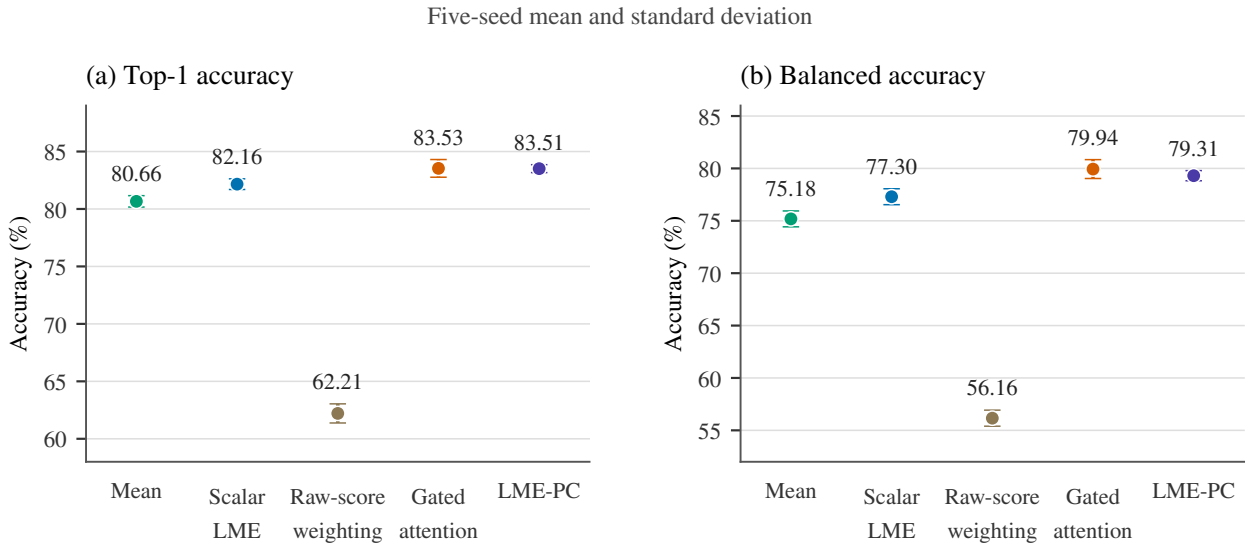


Figure 10. Pooling-strategy comparison on the audited test set. Points are five-seed means and error bars are population standard deviations; Table 10 reports ensemble metrics. Embedding-domain scalar LME and score-domain LME-PC differ in pooling domain and parameterisation.

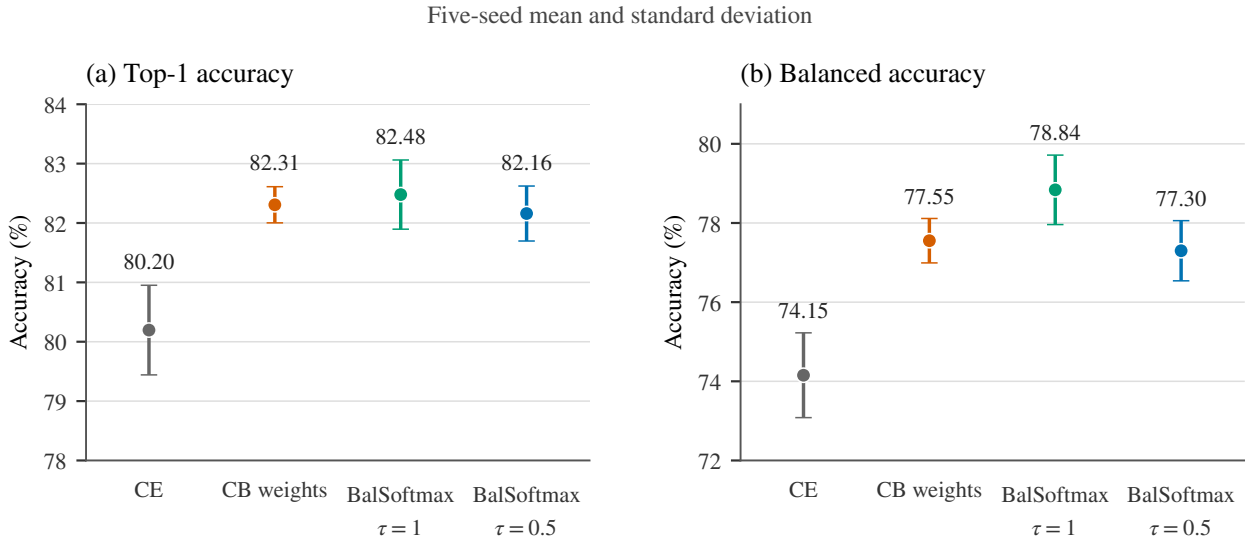


Figure 11. Training-objective comparison on the scalar-LME head. Points are five-seed means and error bars are population standard deviations. The final LME-PC head separately selects τ=0.5 by development accuracy.

Table 9. Comparison of long-tail training strategies on the scalar-LME head (audited corpus, five-seed ensembles). "CB" = class-balanced cross-entropy weights; "BalSoftmax" = balanced softmax, whose shift strength τ multiplies the log class prior in Eq. 4.

Training strategy	Top-1 (%)	Balanced (%)
Cross-entropy	81.23	75.39
CB weights	82.82	77.92
BalSoftmax, τ=0.5	82.94	78.24
BalSoftmax, τ=1	83.44	79.98

Table 10. Pooling-strategy comparison with the balanced-softmax objective fixed at τ=0.5 (audited corpus, five-seed ensembles).

Pooling	Top-1 (%)	Balanced (%)
Mean pooling	81.35	76.15
Raw-score weighting	63.31	57.77
Scalar LME	82.94	78.24
Gated attention	84.66	80.80
LME-PC (selected compact head)	84.05	79.94

On the long tail, 237 of the 276 species have at most 5 training recordings and contribute 463 test samples. On these few-shot species, balanced softmax raises recording-level recall from 73.4% under cross-entropy to 79.3% on the same scalar-LME head. This 5.8-point gain, computed before rounding, supports frequency-aware training for the largest group of checklist species. An auxiliary prototype experiment uses the scalar-LME head trained with balanced softmax at $\tau = 1$ and blends its logits with class-centroid cosine scores in pooled embedding space. Development selection chooses a blend coefficient of zero for that configuration. This finding concerns the scalar-LME reference; prototype blending was not evaluated for the final LME-PC head.

To test whether a deeper head helps, the final linear head was compared with two-layer trunks of width 256 and 512 (Table 11). All three configurations achieve top-1 within 0.12 percentage points and balanced accuracy within 0.24 points. The linear configuration uses 0.28 M trainable parameters, compared with 0.33 M and 0.67 M for the tested trunks. It is retained as the compact choice among these configurations.

Table 11. Impact of head depth on the final LME-PC head (audited corpus, five-seed ensembles). All configs share the per-class LME pooling and balanced-softmax τ=0.5 objective.

Head	Trainable	Top-1 (%)	Balanced (%)
Linear (used)	**0.28 M**	84.05	79.94
Trunk 256	0.33 M	83.93	79.97
Trunk 512	0.67 M	84.05	80.18

3.5. Analysis of model internal states

Sensitivity to per-class temperatures. The learned α_k of Eq. 2 cluster near 0.70, spanning approximately 0.66–0.74 across the five checkpoints. Their role is to control the concentration of each class's window contributions jointly with the scale of its projected scores. An inference-only intervention replaces all 276 temperatures in each checkpoint with that checkpoint's mean while preserving the class projection. Before the intervention, the saved checkpoints reproduce their archived test logits exactly. After equalisation, the ensemble retains 84.05% top-1 and 79.94% balanced accuracy; 813 of the 815 predicted labels remain unchanged. This result characterises the stability of the trained model within the observed temperature range. A scalar temperature trained from the outset is a separate experimental condition; the present intervention

evaluates the existing trained checkpoints.

Taxonomic agreement. Mapping the predicted species to its genus and family yields 86.6% genus agreement and 89.6% family agreement on the test set. These levels complement species-level top-1 accuracy and describe the taxonomic proximity of alternative predictions. For example, panel (e) of Fig. 9 shows a within-genus alternative. Species, genus, and family summaries can therefore provide different levels of context for expert review while the primary recognition target remains the species label.

Where the gate concentrates its flags. The gate's confidence features are computed from the head posterior, and its flags concentrate where the head is least certain (Fig. 12). On the audited test set the fixed-threshold gate flags 12.3% of predictions with 72.0% precision, a 4.5× enrichment over the base error rate, recalls more than half of all errors, and lifts the accepted subset to 91.9% accuracy (AUROC 0.921, AP 0.68). With the development-set empirical target α=0.05 in Eq. 6, the test-set gate accepts 80.7% of predictions and the observed accepted-set error is 4.4%, measured on the test split. At the α=0.10 development target the observed test error is 12.8% with 95.1% coverage; the operating points together trace the coverage–error trade-off available to a practitioner (Table 12).

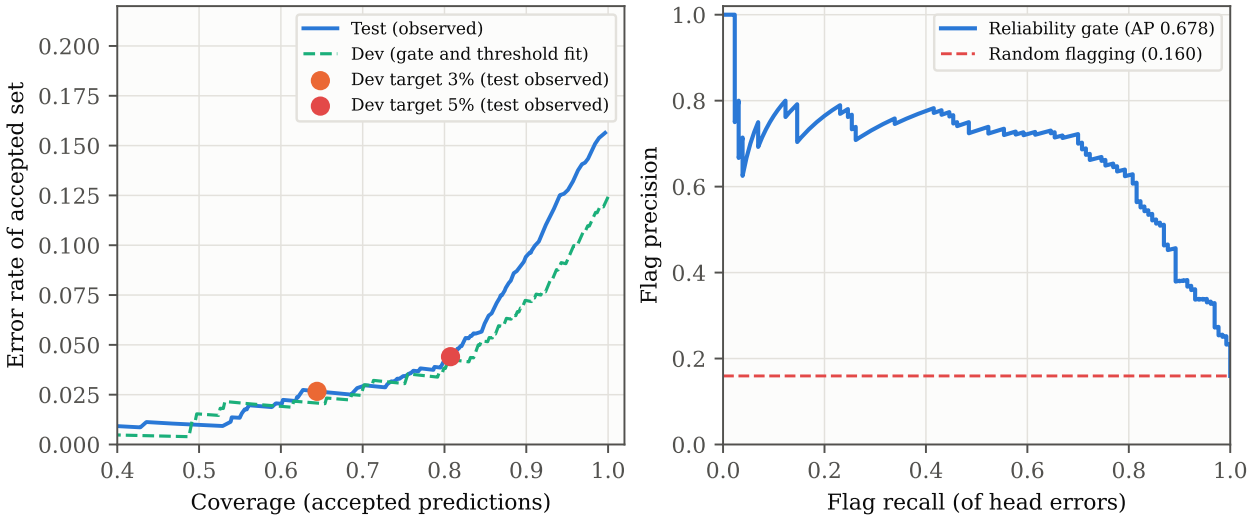


Figure 12. Reliability-gate analysis on the audited test set: (a) the risk-coverage curve with development-selected empirical operating points $\alpha \in \{0.03, 0.05\}$; (b) the PR curve of the gate's error flags against random flagging.

Table 12. Development-selected operating points of the reliability gate, evaluated on the held-out test set. Coverage is the fraction of predictions accepted; accepted error is the error fraction within the accepted subset; error recall is the fraction of all errors flagged.

Target α	Threshold	Coverage (%)	Accepted error (%)	Error recall (%)
0.03	2.629	64.4	2.7	89.2
0.05	0.824	80.7	4.4	77.7
0.10	−0.594	95.1	12.8	23.8

Table 12 lists the operating points that the development-set rule of Eq. 6 selects for three empirical targets. The selected thresholds move monotonically with the target, and the observed test behaviour tracks the development criterion at distinct coverage levels: the α=0.03 point recalls 89.2% of all errors while still covering 64.4% of predictions, and the α=0.10 point covers 95.1% of predictions for workflows that favour throughput. A deployment can therefore pick a target from this curve according to its review budget, using development data only. For the 5% target, 658 of 815 test predictions are accepted and 157 enter the review set; 629 of the accepted predictions agree with the reference label. The observed operating point therefore corresponds to 95.6% accepted-set accuracy. Fitting the gate and choosing its threshold on the same development split defines an empirical selection procedure; the reported test values quantify its performance on the held-out population.

The gate's three features each carry a useful reliability signal (Table 13): every single-feature variant separates errors from correct predictions with an AUROC of at least 0.896. The top-two margin is the strongest single signal, and the full three-feature set attains the best AUROC of 0.9206 together with a competitive average precision, so the compact set is retained. It uses three confidence summaries of an already available posterior. The near-equal AUROC of the margin-only and three-feature variants also provides a compact reference for choosing a gate in a subsequent deployment study.

Table 13. Feature ablation of the reliability gate on the held-out test set. AUROC and AP measure how well each feature set separates errors from correct predictions.

Gate features	AUROC	AP
$p_{\max}$	0.9166	0.6803
$p_1 - p_2$	0.9205	0.6742
entropy	0.8958	0.6390
all three (selected)	0.9206	0.6782

4. Discussion

4.1. Interpretation of window-score aggregation

LME-PC combines regional projection with a smooth class-conditional aggregation rule. For a fixed class, windows with larger projected scores contribute more strongly to the pooled value. This mathematical property explains how the model represents concentrated evidence within a recording. Table 10 shows a 2.70-point top-1 difference between the complete LME-PC and mean-pooling configurations. The configurations also differ in pooling domain and regularisation path, so the comparison supports their evaluated combination. The temperature intervention adds a complementary observation: this combination maintains its aggregate accuracy when the small learned class-wise variation is equalised.

4.2. Computational implications for regional monitoring

Two observations follow from Table 7. First, the regional head contributes a small measured forward-pass cost: 0.54 ms for one head and 2.85 ms for the five-head ensemble on a 96-slot input. The latter corresponds to the ensemble accuracy reported in Table 3. Second, the combined on-disk encoder, head, and preprocessing artifacts occupy approximately 64.08 MB. This is a model-storage figure; runtime libraries and working memory are separate resources. The batch-one encoder timing is below the 1.5-s extraction hop on the measured workstation, providing a useful component-level reference for subsequent streaming integration.

4.3. Prospective ecological evaluation

The current evidence concerns focal archival recordings under a file-hash-based split. Different encodings of a source recording can cross those split boundaries; the saved-checkpoint sensitivity analysis quantifies the observed effect of identified candidates without replacing training under a source-disjoint protocol. Likewise, the gate's empirical development target is an operating choice, not a guaranteed error bound under changes in site, season, recording conditions, or species prevalence. These distinctions define the scope of the reported recognition and review-workload estimates.

The controlled regional-head comparisons, temperature sensitivity analysis, and component measurements support a modular approach to checklist-specific recognition. The next stage is prospective evaluation across independent sites and seasons, with source-identity grouping, soundscape annotations, and review decisions incorporated into the monitoring protocol. This progression connects reproducible archival experiments with regional ecological observation.

5. Conclusion

This paper presented RRC-BirdNET for regional bird-sound recognition using a frozen acoustic encoder, a window-score LME-PC head, balanced-softmax training, and a confidence-based review gate. The five-head ensemble achieved 84.05% top-1 and 79.94% balanced accuracy on 815 audited archival test recordings. On the directly mapped reference-label subset, it achieved 84.24% top-1 compared with 64.68% for checklist-restricted frozen scores. Individual heads averaged 83.51% top-1 with a 1.14 MB checkpoint, providing a compact alternative to the 5.69 MB ensemble package. The gate achieved AUROC 0.921; the development-selected 5% empirical target yielded 80.7% coverage and 95.6% observed accepted-set accuracy.

Declarations

Ethics approval and consent to participate

Not applicable. The study analyses archived audio recordings from public sources (Xeno-canto and a regional birdsong library) together with one illustrative field recording; it involves no capture, handling, or experimental intervention on animals, and no human participants.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional file(s). Training and evaluation code, per-configuration result files, and figure-generation scripts are provided as Additional file 1 (`RRC_BirdNET_Supporting_Information.zip`), whose README documents the verification and recomputation commands. The audited split manifest, cached BirdNET features (approximately 517 MB), and trained head checkpoints are available from the corresponding author upon reasonable request and will be deposited in a public repository before publication. Raw audio and the BirdNET SavedModel remain with their original providers (Xeno-canto, a regional birdsong library, and the K. Lisa Yang Center for Conservation Bioacoustics) and are governed by their terms. Project name: RRC-BirdNET; Programming language: Python; Operating system: platform independent; License: available from the authors upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the National Natural Science Foundation of China, grant number 72001190; the Humanities and Social Science project of the China Ministry of Education, grant number 20YJC6173; the Collaborative Education Project of the China Ministry of Education, grant number 220802518101712; and Zhejiang A&F University, grant number 2022LFR(2).

Authors' contributions

Hao Zhang: Data curation, Methodology, Software, Validation, Visualization, Writing – original draft. **Dayu Xu:** Conceptualization, Formal analysis, Funding acquisition, Project administration, Supervision, Writing – review & editing. **Si Shi:** Investigation, Resources, Validation, Writing – review & editing. All authors read and approved the final manuscript.

Acknowledgements

OpenAI Codex assisted with manuscript restructuring, language editing, LaTeX formatting, correction of figure-generation code, and consistency and citation checks. Numerical findings were checked against locally saved experimental outputs during this revision.

References

BirdNET team. BirdNET-Analyzer: a deep learning solution for bird species identification. `https://github.com/birdnet-team/BirdNET-Analyzer`, 2024a. Custom-classifier training; accessed 2026-09-30.

BirdNET team. Custom classifier on top of all existing species in BirdNET (discussion #275). `https://github.com/birdnet-team/BirdNET-Analyzer/discussions/275`, 2024b. Discussion opened 28 February 2024; accessed 2026-09-30.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4878–4887, 2017. URL `https://papers.nips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html`. arXiv:1705.08500.

Houtan Ghaffari and Paul Devos. Robust weakly supervised bird species detection via peak aggregation and PIE. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1427–1439, 2025. doi: 10.1109/TASLPRO.2025.3552983.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023. doi: 10.1038/s41598-023-49989-z.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. URL `https://arxiv.org/abs/1610.02136`. arXiv:1610.02136.

IPBES. Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services. Technical report, IPBES secretariat, Bonn, Germany, 2019.

Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021. doi: 10.1016/j.ecoinf.2021.101236.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncer-

tainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6402–6413. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Alba Márquez-Rodríguez, Miguel Ángel Mohedano-Muñoz, Manuel J. Marín-Jiménez, Eduardo Santamaría-García, Giulia Bastianelli, Pedro Jordano, and Irene Mendoza. A bird song detector for improving bird identification through deep learning: A case study from Doñana. *Ecological Informatics*, 90:103254, 2025. doi: 10.1016/j.ecoinf.2025.103254.

Cristian Pérez-Granados. BirdNET: applications, performance, pitfalls and future opportunities. *Ibis*, 165 (3):1068–1075, 2023. doi: 10.1111/ibi.13193.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. BirdSet: A large-scale dataset for audio classification in avian bioacoustics. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2403.10380. Accepted at ICLR 2025; arXiv:2403.10380.

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 (NeurIPS)*, volume 33, pages 4175–4186. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/2ba61cc3a8f44143e1f2f13b2b729ab3-Abstract.html.

Giacomo Schiavo. Finetuning BirdNET. https://github.com/giacomoschiavo/finetuning-BirdNET, 2025. Software repository; accessed 2026-09-30.

Jan Schlüter. Learning to monitor birdcalls from weakly-labeled focused recordings. In *Working Notes of CLEF 2021 — Conference and Labs of the Evaluation Forum*, volume 2936 of *CEUR Workshop Proceedings*, pages 1627–1638. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-2936/paper-139.pdf.

Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022. doi: 10.7717/peerj.13152.

Colin E. Studds, Bruce E. Kendall, Nicholas J. Murray, Howard B. Wilson, Danny I. Rogers, Robert S. Clemens, Ken Gosbell, Chris J. Hassell, Rosalind Jessop, David S. Melville, David A. Milton, Clive D. T. Minton, Hugh P. Possingham, Adrian C. Riegen, Phil Straw, Eric J. Woehler, and Richard A. Fuller. Rapid population decline in migratory shorebirds relying on Yellow Sea tidal mudflats as stopover sites. *Nature Communications*, 8:14895, 2017. doi: 10.1038/ncomms14895.

Volodymyr Sydorskyi and Fernando Gonçalves. Tackling domain shift in bird audio classification via transfer learning and semi-supervised distillation: A case study on BirdCLEF+ 2025. In *Working Notes of CLEF 2025*, volume 4038 of *CEUR Workshop Proceedings*, pages 3208–3228. CEUR-WS.org, 2025. URL `https://ceur-ws.org/Vol-4038/paper_256.pdf`.

Yun Wang, Juncheng Li, and Florian Metze. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *ICASSP 2019 — 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 31–35, 2019. doi: 10.1109/ICASSP.2019.8682847.

Lena Ware, C. Lisa Mahon, Logan McLeod, and Jean-François Jetté. Artificial intelligence (BirdNET) supplements manual methods to maximize bird species richness from acoustic data sets generated from regional monitoring. *Canadian Journal of Zoology*, 101(12):1031–1051, 2023. doi: 10.1139/cjz-2023-0044.

Connor M. Wood and Stefan Kahl. Guidelines for appropriate use of BirdNET scores and other detector outputs. *Journal of Ornithology*, 165(3):777–782, 2024. doi: 10.1007/s10336-024-02144-5.

Xeno-canto Foundation. Xeno-canto: Sharing wildlife sounds from around the world. `https://xeno-canto.org`, 2026. Accessed 2026.