Class Interaction as Domain Recalibration: A Variance-Aware Sound-Event-Detection Pipeline for BirdCLEF+ 2026

Yannan Chen[1,*]

[1]Independent Researcher, New York, USA

Abstract

We describe our solo solution to BirdCLEF+ 2026 (identifying 234 bird, mammal, insect and amphibian classes from one-minute soundscapes under a CPU-only inference budget, scored by macro-averaged ROC-AUC), which placed 2nd public / 9th private among 4,000+ teams, built on Xeno-Canto–pretrained EfficientNetV2-S and eca_nfnet sound-event detectors. Two challenges organize the work. The first is the *domain shift* from focal recordings to soundscapes, which we attack with raw-waveform MixUp, external in-domain data, and semi-supervised noisy-student distillation, plus our one architectural addition: a *self-recalibrating class-interaction* module. Probing it across multiple independently-trained models shows it converges to a reproducible *per-class domain recalibration* that systematically boosts the under-represented non-avian taxa, correcting the focal-corpus bird-bias at the output side. The second challenge is *evaluation noise*: the public leaderboard carries a ± 0.003 run-to-run seed scatter comparable to many of the gains we chase, so we treat the whole pipeline as variance reduction at the weight, ensemble, and prediction levels (weight averaging, engineered ensemble diversity, peak-preserving post-processing), reading every gain against it. Seen this way, our largest public-board climb, multi-round distillation, barely moves the private board, a gap that structural diversity in the final ensemble hedges. We also document our AI-assisted research workflow. Our code is publicly available at https://github.com/yannan90/birdclef-2026-private-9th-public-2nd-solution.

Keywords

Bioacoustics, Sound event detection, BirdCLEF, Domain shift, Semi-supervised learning, Noisy-student distillation, Ensemble diversity, Variance reduction, Class interaction, Evaluation noise

1. Introduction

The BirdCLEF+ 2026 challenge [1, 2, 3] asks participants to identify vocalizing fauna in passive acoustic monitoring (PAM) recordings from Brazil's Pantanal — one of the world's largest wetlands, where a network of $\sim$1,000 autonomous recorders captures continuous, multi-species soundscapes for biodiversity monitoring [4]. Continuing the multi-taxonomic setting introduced in 2025 [5], it spans 234 target classes (birds, mammals, insects and amphibians) and requires a probability of presence for every 5-second chunk of one-minute soundscapes, scored by a macro-averaged ROC-AUC that skips classes without positive labels, under a strict CPU-only inference deadline.

Two difficulties dominate. The first is a severe *domain shift*: training audio is overwhelmingly *focal* (directed, mostly single-species Xeno-Canto / iNaturalist clips), while the test data are *soundscapes* (many overlapping species, field conditions, long "nocall" stretches). The second, which organizes this paper, is *evaluation noise*: labeled soundscapes are so scarce — 66 files, validated four-fold on $\sim$16 each — that the per-epoch metric swings by ± 0.04, and a full configuration's leaderboard score still carries a ± 0.003 seed std, on the order of the component gains themselves, so a single-run gain is not, by itself, evidence.

Against these two challenges, our contributions are:

- **A self-recalibrating class-interaction module** (Sections 4.2, 5.3): our one architectural addition. Probed across seven independently-trained models of two architectures, it converges to

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

yannan@alumni.upenn.edu (Y. Chen)

a *reproducible per-class recalibration* (cross-architecture Spearman $\rho \approx 0.96$) that systematically boosts the under-represented non-avian taxa, correcting the focal-corpus bird-bias at the output side. The rest of the domain-shift recipe — raw-waveform MixUp (Section 4.3), external in-domain data (Section 4.4), and semi-supervised noisy-student distillation [6] (Section 4.6) — follows the established BirdCLEF+ 2025 approach.

- **Variance reduction across the pipeline**: we reduce it at the weight (Section 4.5), ensemble (Sections 4.6, 5.5), and prediction (Section 4.7) levels: weight averaging, engineered ensemble diversity (steered by a pseudo-label similarity analysis), and peak-preserving post-processing.
- **Seed-aware evaluation** (Section 5.1): we bring established random-seed variance-accounting to the leaderboard, quantifying the seed noise and re-reading every component against it, with most gains falling within a small multiple of it.

2. Related Work

Architecture and class modeling. The dominant approach treats audio as an image: a Mel-spectrogram fed to a CNN backbone [4, 7], with EfficientNetV2 [8] and NFNet [9] common backbones and the weak/strong SED head of Adavanne and Virtanen [10] (producing both clip- and frame-level predictions), underlying most BirdCLEF solutions. Most solutions score classes independently; methods that model relationships among classes exist in multi-label classification (e.g. graph-convolutional label embeddings [11]) but are rare in BirdCLEF. We add a small class-interaction module that learns a *target-domain species prior* from the clean labeled soundscapes (Section 4.2).

Domain shift and semi-supervised distillation. The recurring challenge is the shift from focal recordings to soundscapes. Since unlabeled soundscapes became available, semi-supervised pseudo-labeling and noisy-student self-training [12, 13] have dominated: the 2025 winning solution [6] established *multi-iterative noisy-student* training — MixUp between focal clips and pseudo-labeled soundscapes, iterated with sharpened pseudo-labels and an ensemble across iterations — which our pipeline (Section 4.6) follows. MixUp [14] and SpecAugment [15] are near-universal augmentations in this setting. Our class-interaction module (Section 4.2) addresses the same shift from a different angle, learning a target-domain species prior at the output side rather than through input augmentation.

Evaluation reliability. Less attention has gone to the reliability of the comparisons themselves. Past BirdCLEF editions have recurrently struggled with unstable or scarce evaluation: in-domain labeled data is limited, and above a public score of ≈ 0.9 the local-validation–to-leaderboard correlation is widely observed to break down. More broadly, accounting for random-seed variance is established practice in rigorous ML evaluation [16, 17, 18], yet competition write-ups routinely rank methods on single-run deltas. We bring this variance-accounting to the leaderboard, quantifying the seed noise and re-reading every component against it.

3. Task and Data

The 234 target classes are taxonomically lopsided — 162 birds, 35 amphibians, 28 insects, 8 mammals and a single reptile — so the macro-average is dominated by, but not reducible to, birds. Training data are overwhelmingly *focal*: 35,549 Xeno-Canto / iNaturalist clips with weak recording-level labels (a label may apply to only a few seconds of a multi-minute file), and even so only 206 of the 234 classes have any focal recording. Our only strongly-labeled, in-domain data — the *labeled* soundscapes — are strikingly scarce: 66 one-minute recordings (1,478 five-second chunks), in which only **75 of the 234 classes** ever appear and 89% of call-bearing chunks contain more than one species. A further 10,658 *unlabeled* soundscapes are available for pseudo-labeling. This mismatch — in label granularity, simultaneous-species count, and recording conditions — is the core challenge. Because the labeled soundscapes are the only trustworthy in-domain supervision, we reserve them for validation and the class-interaction module; their narrow coverage (75/234) is a limitation we revisit in Sections 5.3 and 8.

4. Method

4.1. Backbone and front-end

We use two Xeno-Canto–pretrained backbones — EfficientNetV2-S [8] and eca_nfnet_l0 [9], from the 2025 in-domain checkpoints released by V. Sydorskyi — loaded through timm [19]. This domain-matched initialization gave a very strong baseline. We also pretrained alternative architectures (hgnetv2_b4, regnety032, seresnext26t), but all underperformed: EfficientNetV2-S was the best, and we kept eca_nfnet_l0 only as a cross-architecture member for ensemble diversity (Section 4.6). The Mel front-end is coupled to the pretraining and was kept fixed ($sr = 32$k, $n_{\text{mels}} = 128$, $n_{\text{fft}} = 2048$, hop$= 512$, $f_{\min} = 20$, $f_{\max} = 16$k).

4.2. SED architecture and the class-interaction module

Each model is a standard SED stack: backbone → GeM frequency pooling [20] → attention head producing both a clip-level logit and a frame-level (frame-max) logit, trained with BCE on both heads. We operate at **10-second windows split into two 5-second sub-segments** (denoted 10s/2). Figure 1 shows the stack.

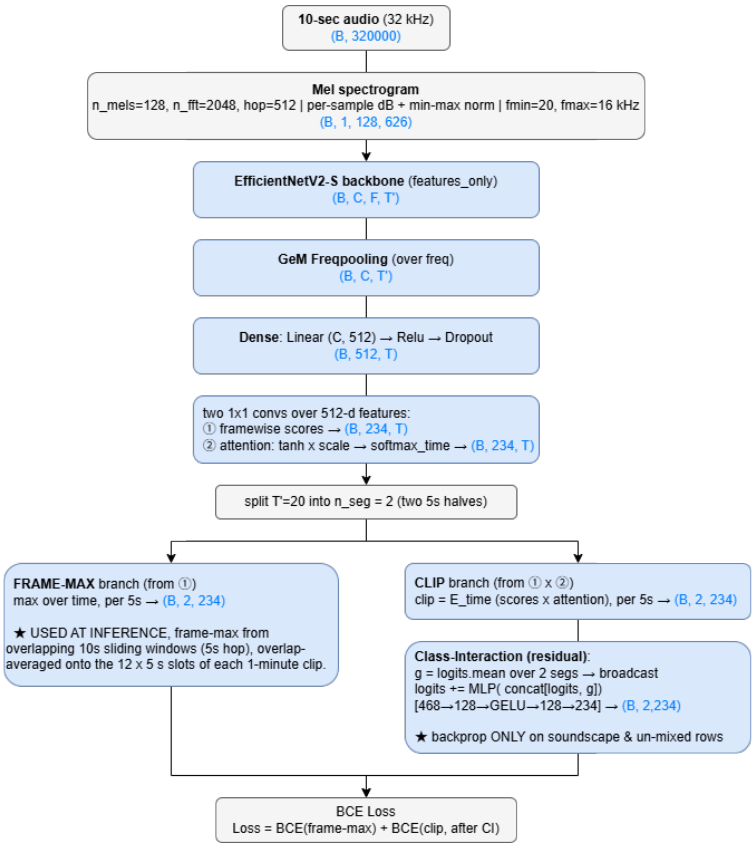


Figure 1: The SED stack with the class-interaction (CI) module. The frame branch is used for inference; the clip branch carries training gradients. The CI module residually refines the clip logits and is updated only from clean, un-mixed labeled-soundscape rows.

On top of the clip logits we add a *class-interaction* (CI) module: a small residual class×class MLP with a zero-initialized last layer, so it starts as the identity and gradually learns a correction. Because labeled soundscapes are too scarce to adapt the backbone to directly, the module is a cheap way to learn a lightweight *species-level soundscape prior* on top of the focal-trained model, pushing back on the focal→soundscape domain shift. We deliberately leave its form open (it sees the full class-logit vector and may couple any classes) and ask later (Section 5.3) what prior it actually learns. The key

design choice is *what it learns from*: it receives gradients only from labeled-soundscape rows that were not MixUp'd, so it adapts to the real target domain rather than to MixUp artifacts. At 10s/2 it uses a cross-segment "pool" variant: each 5-second segment is refined from its own logits concatenated with the 10-second-window mean, a context channel that does not exist at 5s/1. Concretely, writing $z_s \in \mathbb{R}^C$ for the clip logits of segment s and $\bar{z} = \frac{1}{S}\sum_s z_s$ for the 10-second-window mean, the module refines

$$\hat{z}_s \;=\; z_s \;+\; W_2\,\sigma\big(W_1\,[\,z_s \,\|\, \bar{z}\,]\big), \qquad W_1 \in \mathbb{R}^{128\times 2C},\;\; W_2 \in \mathbb{R}^{C\times 128}, \tag{1}$$

with σ the GELU nonlinearity and W_2 zero-initialized so the module begins as the identity, applied only on the masked rows above. Mechanically, then, the recalibration is a context-aware, per-class *additive* shift: the module reads the full logit vector $[z_s \,\|\, \bar{z}]$ and returns a correction $\hat{z}_s - z_s$ that nudges each class up or down given the others present, learned end-to-end from the in-domain rows. We defer the empirical questions — how much this module actually helps, and what it really learns — to Sections 5.2 and 5.3, where we measure its effect against the seed noise and then probe the trained weights.

4.3. MixUp as a focal→soundscape bridge

Because almost all training data are clean single-species focal clips while the target is multi-species soundscapes, we use MixUp [14] deliberately to *synthesize soundscape-like, multi-species examples* from abundant focal audio. Two choices make it work. First, we mix on the *raw waveform before the Mel transform* (each clip absmax-normalized), which preserves per-frame, per-species energy contrast so the attention head can still separate the mixed species. With probability MixProb $= 0.5$ a clip s_i is blended with a partner s_j (at most two sources) as

$$\tilde{s} = \lambda\, s_i + (1-\lambda)\, s_j, \qquad \tilde{y} = \max(y_i, y_j), \qquad \lambda \sim \mathrm{Beta}(0.4, 0.4), \tag{2}$$

where the mixture is taken on the raw waveform s (not the spectrogram) and labels are combined multi-hot. The U-shaped $\mathrm{Beta}(0.4, 0.4)$ usually yields one loud and one faint clip, mimicking a foreground bird over background calls. Second, the roles are asymmetric: only focal clips receive a partner, while the scarce labeled-soundscape rows are kept clean on the left ($\mathrm{perm}[i] = i$) so their supervision stays a pure anchor; they may, however, serve as partners blended into focal clips. Figure 2 shows the effect: blending two clean single-species focal clips produces an example whose time–frequency structure carries both species at once, far closer to a real multi-species soundscape chunk than either source.

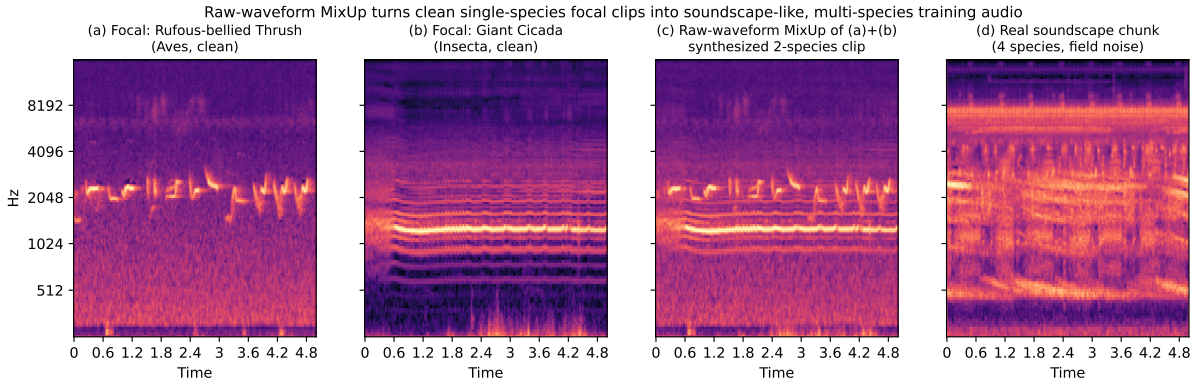


Figure 2: Raw-waveform MixUp as a focal→soundscape bridge. (a,b) two clean single-species focal clips (a bird and an insect); (c) their raw-waveform mix, in which both the bird's harmonic whistles and the insect's broadband drone survive into one example; (d) a real labeled-soundscape chunk with several overlapping species and field noise. The synthetic mix (c) is far closer in structure to the target (d) than either source clip.

4.4. External data: closing the taxonomic gap

Adding external training data was, by a wide margin, the single most effective change we made, and unlike most of our tweaks its effect is unambiguous on both the local set and the leaderboard (Section 5.2).

The reason is specific to this competition's taxonomy: the focal recordings are 98% birds, so the non-avian target taxa — amphibians, mammals, and especially insects — are barely represented, and a model trained on focal audio alone is almost blind to exactly the non-bird sounds that fill a Pantanal soundscape. We closed this gap with open, in-domain corpora — **AnuraSet** [21] for Neotropical anurans, **InsectSet459** [22] for insects, and research-grade **iNaturalist** recordings, plus additional Xeno-Canto material for birds [23] — matching every source recording to the 234 target classes, capping each class at 200, and leaving the soundscape validation set untouched.

The distribution is the point (Figure 3): the added material lands exactly where the focal corpus is thin, and for amphibians it is itself in-domain Neotropical pond audio, even closer to the target soundscapes than directed focal clips. The insect classes stay near-empty: they are unidentified sound types (Insect son01–son25) with no external counterpart, so no public data can reach them. This is the largest jump in the study (Table 2): $+0.045$ local and $+0.005/+0.006$ public/private, both clearly above the seed noise of Section 5.1.

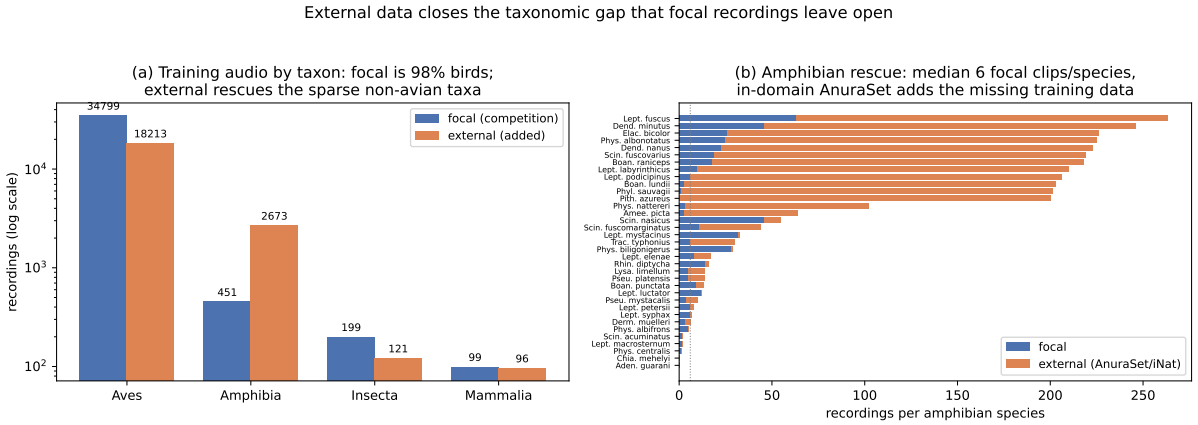


Figure 3: Why external data helped. (a) Training audio by taxon (log scale): focal recordings are 98% birds; external data rescues the sparse non-avian taxa. (b) Per-amphibian-species rescue: most species sit at or below the focal median (dotted line); in-domain AnuraSet / iNaturalist data lifts them by an order of magnitude. Insects cannot be rescued: most are unidentified sound types with no external counterpart.

The local gain far outstrips the leaderboard gain: the model sharpens its separation of the specific validation frogs and insects more than that resolution generalizes to unseen recordings, another instance of the local metric flattering a real but smaller true effect (Section 5.1).

4.5. Validation and training

We use 4 folds (focal: stratified K-fold on the primary label; soundscapes: grouped K-fold by filename) and *validate on soundscapes only*, reporting fold-mean macro-AUC as a reference metric whose leaderboard correlation is weak (Section 5.1). Training uses AdamW [24] (lr 5×10^{-4}, wd 10^{-4}, batch 64) and a flat-cosine schedule (2 warmup + 20 flat + 8 cosine). Upsampling rare focal classes to $\geq$100 segments gave a small but consistent public-LB gain.

EMA is load-bearing, not a finishing touch. With only 66 labeled soundscapes to validate on (Section 3), the per-epoch validation macro-AUC is dominated by sampling noise: on raw weights it swings by up to ±0.04 between adjacent epochs (Figure 4). This blocks development twice over: the best-epoch checkpoint is frequently a lucky spike that does not generalize, and the epoch-to-epoch wobble is as large as the single-variable effects we were trying to measure. An exponential moving average of the weights (decay 0.999, from epoch 3) [25, 26, 27] cuts the jitter by $\approx3\times$ (2.3–4.5$\times$ across folds), so that the best epoch and the onset of over-fitting become legible and component comparisons readable (Figure 4). We therefore validate and save every model in this paper on EMA weights, our first variance-reduction lever, acting at the weight level.

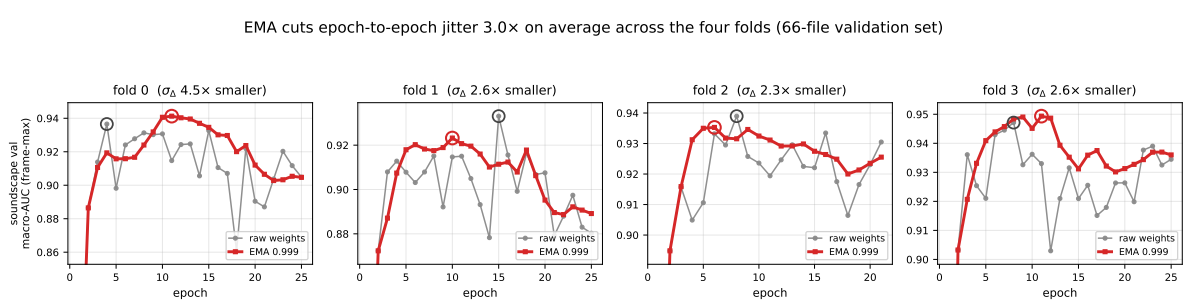


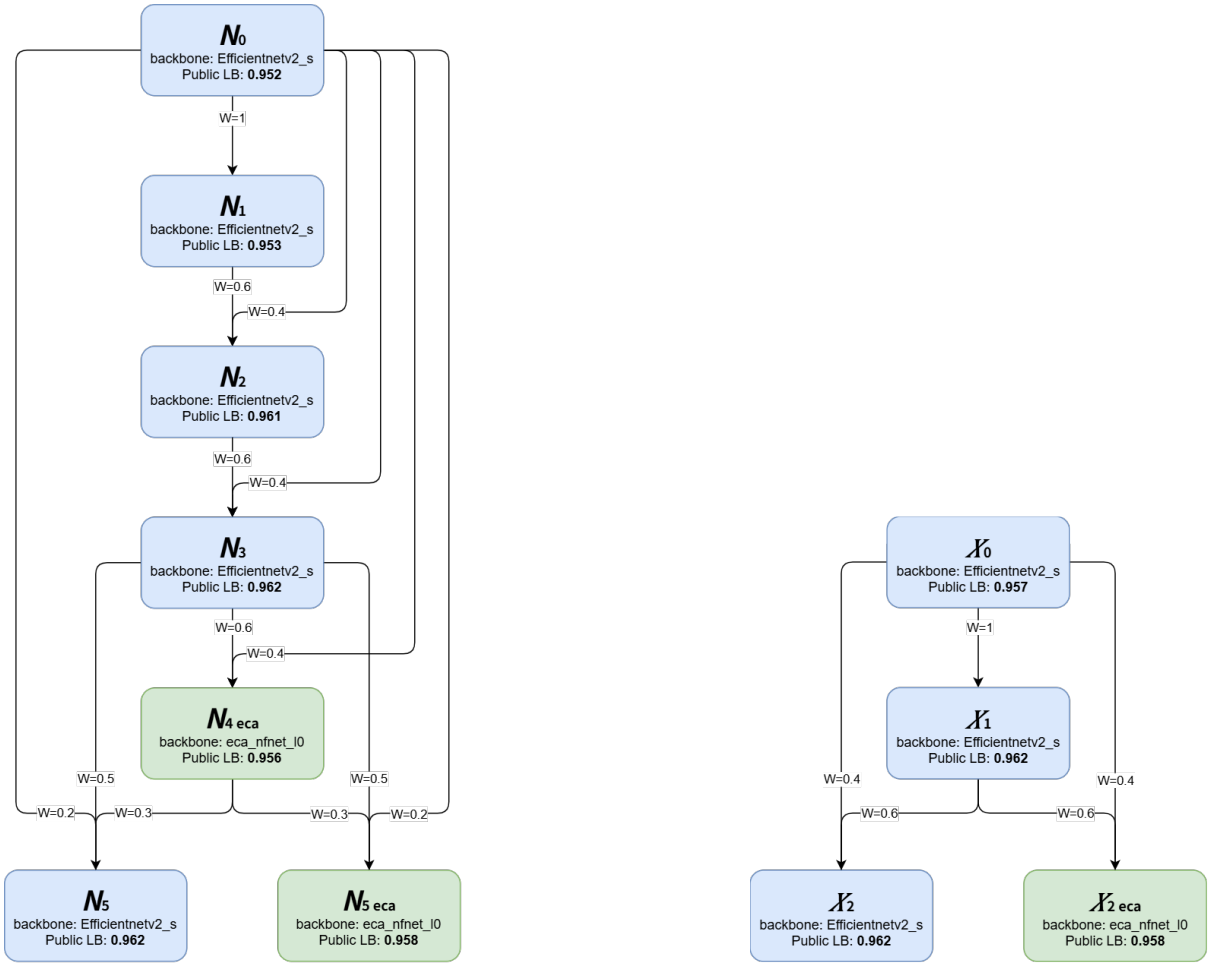
Figure 4: Why EMA is necessary. Per-epoch soundscape val macro-AUC (frame-max head) per fold, on raw weights (grey) vs. their EMA (red); circles mark each curve's best epoch. On 66 validation files the raw curve bounces epoch-to-epoch with an erratic best epoch, whereas EMA is markedly smoother with a stable best epoch.

4.6. Noisy-student distillation: engineering diversity

Our in-domain supervision is only 66 labeled soundscapes (Section 3), yet the competition also ships 10,658 *unlabeled* soundscapes from the same domain — by far our largest pool of in-domain audio, and otherwise unused. To exploit it, we mine it with iterated *noisy-student self-training* [13, 6, 28], a semi-supervised scheme (the winning BirdCLEF+ 2025 recipe [6]) that turns these unlabeled soundscapes into training signal. At each round a *weighted-ensemble teacher* of earlier models produces *soft pseudo-labels* on the ∼127k unlabeled soundscape segments; we sharpen them (raising probabilities to a power of 1.3–1.4, tuned by public-LB feedback) and distil a fresh, noised student on them as MixUp targets, with drop-path 0.15 to curb over-fitting to the pseudo-labels. Figure 5 traces the two chains we ran. The long one starts from N_0, our base single model trained *without* external data, and runs five rounds ($N_1 \ldots N_5$); the short one starts from X_0, the same model trained *with* the external in-domain data of Section 4.4, and already strong, so it reaches a comparable level after a single round; a stronger starting point needs far fewer iterations. Subscripts denote the distillation generation and a superscript "eca" marks the eca_nfnet_l0 backbone (default is EfficientNetV2-S). The per-round leaderboard trajectory is deferred to the analysis in Section 5.4.

The non-obvious ingredient is *diversity preservation.* Distillation pulls each student toward its teacher ensemble, so successive students converge: along the no-external effv2 chain the pseudo-label predictions reach a mutual Pearson correlation of ≈ 0.98 (Figure 6), and left unchecked the "ensemble" collapses toward a single effective model. Two moves keep it diverse. First, we keep the ancestor in every teacher ensemble: N_0 is by far the most decorrelated member — it correlates only 0.89–0.94 with its own descendants, against ≈ 0.98 among them — so even a small ensemble weight on it re-injects an independent view the distilled students have lost. Second, architecture matters: an eca_nfnet_l0 student sits visibly apart from the EfficientNet cluster (cross-architecture correlation ≈ 0.96 vs ≈ 0.98 within), a second diversity axis. We add the eca backbone late on purpose: it is the weaker model on this task in isolation, and only trains into a competitive student once the pseudo-labels are already high quality.

This preserved diversity does double duty. Inside the loop it makes a better teacher — a decorrelated ensemble yields sharper, less self-reinforcing soft labels than any single model — so each round's pseudo-labels improve on its own members. Across runs, the same axes (ancestor, distillation generation, and backbone) became the blueprint for our final graded submission, which blends deliberately along them rather than stacking near-duplicate students. A diverse blend of this kind scores markedly higher on the public leaderboard than any of its members; we quantify that climb (and its subtler, less favorable relationship with the private board) in Sections 5.4 and 5.5.



(a) no-external chain (N_0 ancestor) (b) external chain (X_0 ancestor)

Figure 5: The two noisy-student distillation chains (node = a model generation with its public LB; edge weight = that model's weight in the next round's teacher ensemble). The ancestor (N_0 / X_0) is fed back into every round as a diversity anchor, and an eca_nfnet branch (green) adds a cross-architecture view late.

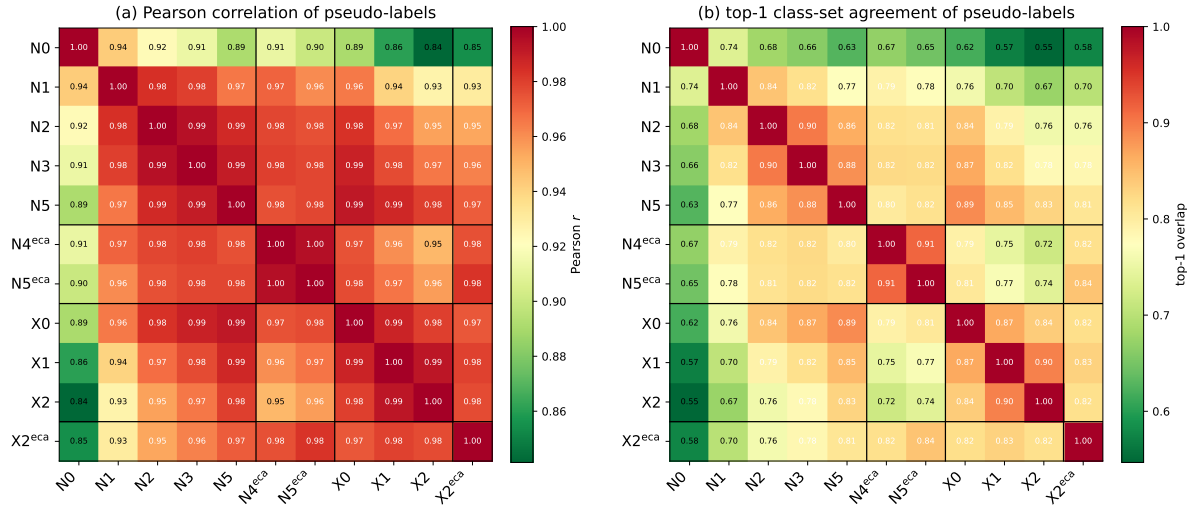


Figure 6: Pseudo-label similarity across the distillation chain: **(a)** Pearson correlation of the probability surface and **(b)** top-1 class-set agreement (how often two models pick the same top species per 5s clip). Both show distilled effv2 students converging (≈ 0.98), while the ancestors N_0/X_0 stay most decorrelated and the eca models form a separate cluster: the diversity sources we preserve.

4.7. Inference post-processing

Post-processing is the last of our variance-reduction stages, acting on the prediction timeline; we use two steps, both at the file level.

(i) Taxon-aware temporal smoothing. Each one-minute file yields a timeline of twelve 5-second predictions per class, and neighboring windows are correlated (a call often spills across a 5-second boundary), so some temporal smoothing helps. *How* one smooths matters far more than *whether*, because a strong model already produces sharp, confident peaks at the true call times, and the obvious schemes erode them. We compared three (Table 1):

- **Gaussian**: convolve every class with a fixed $(.1, .2, .4, .2, .1)$ 5-tap kernel. Blunt and taxon-blind; being a symmetric average it pulls confident peaks down toward their neighbors.
- **Texture (taxon-split)**: a lighter symmetric 3-tap, with separate kernels for event taxa ($.2/.6/.2$, mostly self) and texture taxa ($.35/.3/.35$). Respects the taxon distinction but is still an average, so it still erodes peaks.
- **Taxon-aware (ours)**: match the operator to the call type. *Texture* taxa (Insecta, Amphibia: continuous choruses whose true signal is temporally smooth) get mean diffusion; *event* taxa (Aves, Mammalia, Reptilia: transient calls) get a *one-directional* local-max, $(1-\alpha)x + \alpha \max(x, x_{\text{prev}}, x_{\text{next}})$, which only *raises* an under-confident window toward a neighboring peak and never lowers the peak itself. A confidence-adaptive term ($\alpha \propto 1 - \max_c p_c$) further spares windows the model is already sure about.

On a held-out 6-model out-of-fold ensemble the difference is decisive on **Aves**, the taxon with the sharpest peaks: the gaussian average *lowers* its AUC by 0.0021 while the taxon-aware scheme *raises* it by 0.0011, a 0.003 swing (Table 1). The naive smoothers are net-negative overall; only the peak-preserving scheme helps. A layer-by-layer probe shows the lift comes entirely from the texture-mean and event-local-max operators; the confidence-adaptive term is redundant on OOF ($+0.0001$) and could be dropped. The gains are small (post-processing is near saturation here), but the lesson is not: for a high-performance model, smoothing must be peak-preserving, or it costs more than it gives.

Table 1
Temporal smoothing schemes on a 6-model out-of-fold ensemble (frame-max, 75 active classes). Δ is macro-AUC vs. no smoothing; the Aves column (transient calls) isolates the peak-preservation mechanism.

Scheme	Operator	Aves AUC	Δmacro
none	—	0.9746	—
Gaussian 5-tap	uniform average, all taxa	0.9725	-0.0002
Texture 3-tap	taxon-split average	0.9744	$+0.0003$
Taxon-aware	texture-mean + event local-max	**0.9757**	$\mathbf{+0.0009}$

(ii) File-post [28]: each 5-second prediction is reinforced by the per-species top probability over the whole one-minute file,

$$\hat{p}_{a,s,c} = p_{a,s,c} \cdot \max_{s'} p_{a,s',c}, \tag{3}$$

where a indexes the one-minute audio file, s the 5-second segment and c the class, boosting species that are confidently present somewhere in the clip.

5. Results and Analysis

5.1. How noisy is the metric?

The most important caveat in this work is that the public leaderboard is noisy. The public score is computed on a hidden set of roughly 200 one-minute soundscapes, and re-running the same configura-

tion under a different random seed — which reshuffles the focal cross-validation split and the training randomness, with the soundscape validation split held fixed — shifts it by about ± 0.003, comparable to the gaps between configurations. Our *local* validation is even noisier: the only 66 labeled soundscapes cover just 75 of the 234 classes, and validating per fold leaves each of the 4 folds scored on only $\sim$16–17 files, so the 4-fold-mean macro-AUC wobbles by ± 0.005 across seeds and barely tracks the leaderboard. A ± 0.005 local swing thus cannot resolve a ± 0.003 leaderboard difference, and neither a single submission nor local validation can rank configurations reliably.

We therefore lean on the leaderboard itself: we submit heavily and judge each strategy by its *best-of-n-seed* public score (also how we pick the seed and weights we keep), while reporting the across-seed spread throughout (the variance-accounting standard in rigorous ML evaluation [16, 17, 18], but rarely applied to competition leaderboards), so the reader can separate signal from seed luck.

5.2. Incremental ablation

Read against that seed noise, Table 2 reports an incremental build-up over our 5-second baseline: we add the class-interaction (CI) module (Section 4.2), then a 10-second window, then both together, then external data (each at several seeds; see the table caption for the columns). The best single model, exp_D (10s/2+CI), reached 0.952/0.949; adding external data (exp_E) reached 0.957/0.955, with a $+0.045$ local jump on the unchanged validation set. Figure 7 shows why: the local 4-fold-mean barely tracks the public board (Pearson $r = -0.09$), and each configuration's per-seed spread is a sizable fraction of the gaps between configurations, so every row is one draw from a distribution rather than a fixed score.

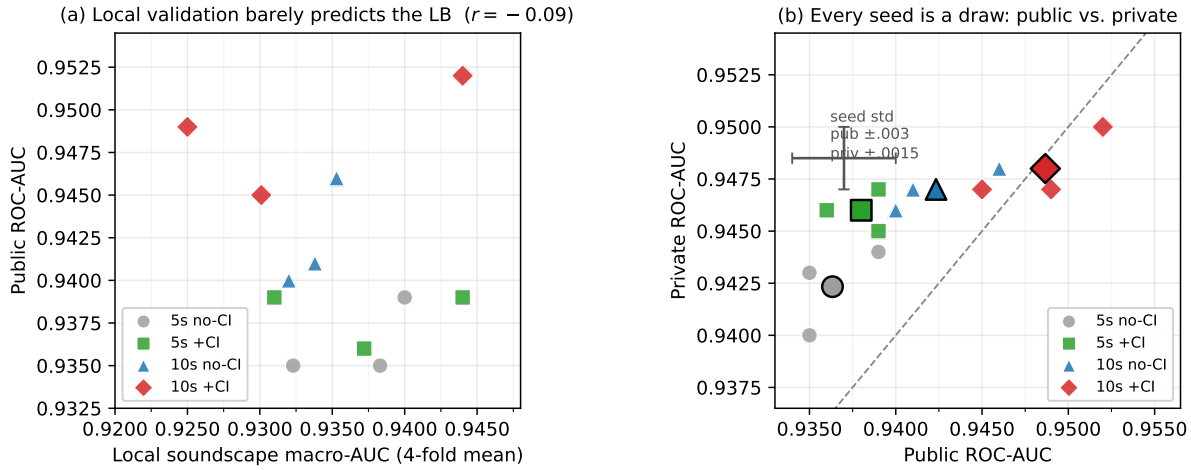


Figure 7: Seed noise rivals component gains. (a) Local 4-fold-mean macro-AUC vs. public leaderboard for every seed run ($r = -0.09$). (b) Public vs. private leaderboard per seed run, colored by configuration; large outlined markers are per-configuration centroids, the dashed line is $y = x$, and the cross shows the seed std (± 0.003 public, ± 0.0015 private).

Table 2
Incremental ablation. Public/Private are the best seed per configuration; local macro-AUC is the across-seed mean$\pm$std over n seeds (4 folds).

Config	dur/seg	CI	Public	Private	Local (mean$\pm$std)	n
exp_A	5s/1	—	0.939	0.944	0.937 ± 0.004	3
exp_B	5s/1	yes	0.939	0.947	0.937 ± 0.007	3
exp_C	10s/2	—	0.946	0.948	0.934 ± 0.002	3
exp_D	10s/2	yes	**0.952**	**0.949**	0.933 ± 0.010	3
exp_E	10s/2	yes	**0.957**	**0.955**	**0.979**†	1

†exp_E's validation set is unchanged, so its higher local AUC is comparable: a $+0.045$ jump.

5.3. The class-interaction module: a per-class domain recalibration

The CI module is an auxiliary, training-only component. Its class×class residual refines the attention-pooled *clip* logits and enters the loss only on clean, un-mixed labeled-soundscape rows, whereas inference reads the frame head (fm), which bypasses the residual. Its effect on the deployed model is therefore the gradient it contributes to the shared frame maps during training.

To see what prior it learned, we probe the trained module on the real labeled soundscapes across seven models — five EfficientNetV2-S spanning both distillation chains (N_0, N_2, N_5, X_0, X_2) and two eca_nfnet cross-architecture members (N_5^e, X_2^e) — over all four folds, i.e. 28 model-folds. Three findings hold across them (Figure 8). **(i) It is per-class, not scene-dependent**: 82–97% of the correction's variance is between classes and only a small fraction varies segment-to-segment, so the module applies a near-fixed per-class shift rather than reacting to which other species are present. **(ii) It is reproducible, even across architectures**: the six mature models agree at Spearman $\rho = 0.94$–0.99 (mean 0.97), and crucially the eca_nfnet members agree with the EfficientNetV2 ones just as strongly (cross-architecture $\rho = 0.96$ on average), so the prior is a property of the task, not of one backbone. Only the clean ancestor N_0, whose CI module is the least distilled, is an outlier ($\rho \approx 0.6$, with a larger overall shift). **(iii) It is a domain recalibration**: it systematically boosts the under-represented non-avian taxa — insects and amphibians by $\approx$+1–2 logits versus $\approx$+0.5 for birds, an effect that survives controlling for each class's confidence level — and shrinks over-confident classes. The module thus learns a lightweight *per-class domain-shift recalibration* [29] from the scarce in-domain labels, attacking the same focal-corpus bird-bias as the external data (Section 4.4), but at the output side.

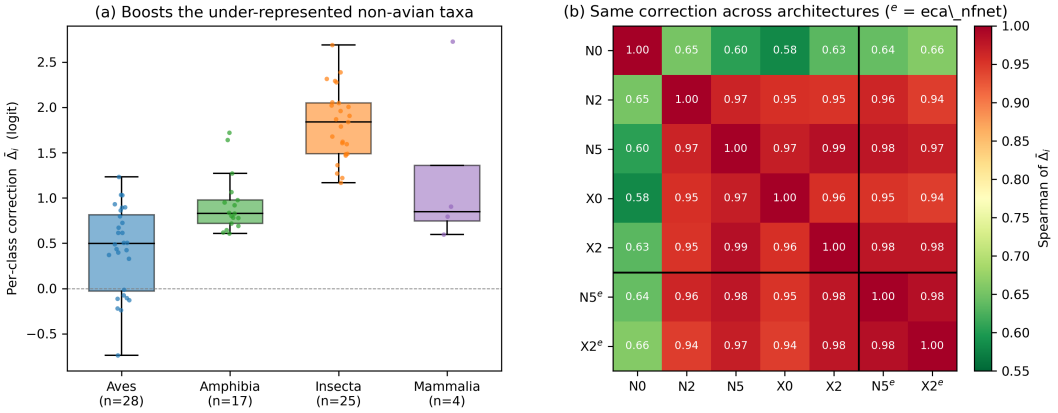


Figure 8: What the class-interaction module learns, probed on the real labeled soundscapes across seven models (five EfficientNetV2-S, two eca_nfnet marked ^{e}) × four folds. **(a)** Per-class correction $\bar{\Delta}_i$ by taxon (six mature models): it boosts the under-represented non-avian taxa, insects most of all. **(b)** The correction is highly consistent across models and architectures (Spearman 0.94–0.99); only the least-distilled ancestor N_0 differs.

5.4. The noisy-student distillation outcome: a public-board phenomenon

The engineered diversity of Section 4.6 drove public almost monotonically up, to 0.962 on both chains (Table 3), but private did not follow: it peaked at the ancestor and the first rounds ($X_0 = 0.955$, $N_1/N_3 = 0.953$) and then drifted down to 0.947–0.951 as distillation continued. The strongest private models are the early, barely-distilled ones; the chain was almost entirely a public-LB phenomenon and, on the board that decides the ranking, it slightly hurt: this within-chain private decline ($0.955 \to 0.947$) is large enough to outrun the seed noise of Section 5.1 — volatility that, as Section 5.5 shows, structural diversity in the final blend was able to absorb.

Table 3
Both distillation chains. Public climbs with every round, but private peaks at the ancestor / first rounds (bold) and then declines: the best private models are the least-distilled ones.

Model	Backbone	Public	Private
N_0 (exp_D)	EffV2-S	0.952	0.949
N_1	EffV2-S	0.953	**0.953**
N_2	EffV2-S	0.961	0.952
N_3	EffV2-S	**0.962**	**0.953**
N_4^{eca}	eca_nfnet	0.956	0.949
N_5	EffV2-S	**0.962**	0.951
N_5^{eca}	eca_nfnet	0.958	0.947
X_0 (exp_E)	EffV2-S	0.957	**0.955**
X_1	EffV2-S	**0.962**	0.954
X_2	EffV2-S	**0.962**	0.950
X_2^{eca}	eca_nfnet	0.958	0.948

5.5. Final submissions: structural diversity as a private-board hedge

Our two graded submissions blend the chain endpoints of Section 4.6 along three diversity axes — distillation generation, backbone (EfficientNetV2-S / eca_nfnet_l0), and input duration (the last via a 15s-window variant of the second-round external model X_2, denoted $X_2^{15\text{s}}$) — shaped by the CPU inference budget (Table 4). To fit more models into the time limit we average only 3 of the 4 CV folds (2 for the ancestor N_0). Both more models and overlapping windows average more passes per window and make the blend more robust, but the budget cannot afford both, and their effect on private leaderboard is untested — so v1 favors breadth (more models) and v2 depth (overlapping windows).

The blend weights follow each member's individual public-LB score (Table 3): a stronger model takes a larger share (e.g. 0.6 vs. 0.4), which a few full-ensemble submissions then calibrate and confirm. The ancestor (N_0) proves useful despite its low LB score and is kept at a small weight. Because the public leaderboard is noisy (Section 5.1), we keep these round weights rather than tuning them exhaustively.

The diversity is deliberate. The distillation rounds that lifted public also perturbed private (Section 5.4); blending across these axes damps that per-chain volatility, holding both recipes at 0.956 private, above the best single model (0.955), while public reaches $0.967/0.966$ (Table 5). The selected submission placed **2nd public / 9th private** among 4,000+ teams.

Table 4
The two graded ensemble recipes. Models use the codes of Section 4.6 (N/X chain, $^{\text{eca}}$ backbone, $^{15\text{s}}$ duration); *folds* is how many of the 4 CV folds were averaged, w the relative blend weight (renormalized to sum 1), and *window* whether overlapping-window inference was used.

Recipe	Model	folds	w	window
v1	N_5	3	0.3	no
	X_2	3	0.3	no
	N_5^{eca}	3	0.4	no
	X_2^{eca}	3	0.2	no
	$X_2^{15\text{s}}$	3	0.2	no
	N_0	2	0.2	no
v2	N_5	3	0.6	yes
	N_5^{eca}	3	0.4	yes
	$X_2^{15\text{s}}$	4	0.3	no
	N_0	2	0.2	no

Table 5
Our two graded submissions on the public and private leaderboards.

Submission	Public	Private
v1 (diverse pseudo-label ensemble)	**0.967**	0.956
v2 (diverse pseudo-label ensemble)	0.966	0.956

6. Inference Engineering: Fitting the Blend into the CPU Budget

Submissions are scored under a strict CPU-only limit (four cores, ≈1.5 h for the full hidden test set), so inference cost, not only accuracy, decides how large an ensemble is affordable, and hence the fold/window trade-offs of Table 4. Three choices made the six-model blend fit comfortably.

ONNX Runtime with graph optimization. Every model is exported to ONNX and executed under ONNX Runtime [30] with its graph optimizations (operator fusion, constant folding) enabled. A single OpenVINO [31] session was in fact the fastest per model (≈1.4× a single ONNX Runtime session), but its internal multi-threading oversubscribed the four cores the moment several models ran together; ONNX Runtime parallelizes far more cleanly at the process level.

Four parallel single-threaded workers. Rather than one internally multi-threaded session, we shard the test files across *four independent worker processes*, each capped to a single intra-op thread and running its own feature extraction and inference. This saturates the four cores with near-linear speed-up and no thread contention, the key to running a multi-model blend inside the limit.

Shared Mel front-end. Because the whole system fixes a single Mel front-end (Section 4.1), each worker computes the log-Mel of an audio window once and reuses it across every model that shares it: all 10s/2 backbones, EfficientNet and eca_nfnet alike, read the same spectrogram, and only the 15s variant pays for its own. Feature extraction is thus amortized over the whole blend instead of repeated per model.

Together these turn a multi-model CPU ensemble from a timeout into a comfortable fit; the headroom they free is exactly what the extra members of Table 4 spend.

7. AI-Assisted Research Workflow

As a solo participant with a full-time job outside machine learning, we leaned heavily on an agentic coding assistant (Claude Code).

- **Experiment automation**: the agent ran a complete pipeline — launching and configuring training runs, validating, logging results, exporting ONNX with numerical sanity checks, and preparing submissions — with autonomous control of the cloud GPU instances and of the local file system, plus an overnight monitor loop that kept the experiments iterating and reporting while we were at work.
- **Data analysis and idea testing**: a large volume of analyses and quick probes were implemented and executed by the agent, including the pseudo-label and model-similarity analysis that drove ensemble diversity (which backbones were genuinely independent, Section 4.6), the training-data and domain-similarity studies (Sections 4.3, 4.4), the incremental ablations (Section 5.2), the mechanistic probe of the class-interaction module (Section 5.3), and the figures of this note.
- **Offline search without spending submissions**: the agent built an external macro-AUC validation set from public neotropical soundscapes of earlier CLEF editions and used it to sweep post-processing and ensemble hyperparameters offline.
- **Inference engineering**: when our CPU kernel began timing out, the agent rebuilt and benchmarked the inference path, driving the four-worker ONNX Runtime design of Section 6 that kept submissions within the deadline.

Across all of this, the ideas, direction, and judgment remained the author's; the agent supplied fast, reliable execution that let us try far more ideas than we otherwise could.

8. Limitations and Conclusion

8.1. Limitations

Our evaluation and the class-interaction module see only the 75 of 234 classes present in the labeled soundscapes, which are geographically narrow and correlate weakly with the leaderboard, so coverage of the remaining taxa and transfer to other regions are untested. That same narrow coverage bounds what we can say about the module per class: we probe what it learns (Section 5.3), but firming up its *per-class* performance gains would need labeled soundscapes spanning a broader, more balanced set of taxa. The insect classes are effectively unsolved: 25 of 28 are unidentified sound types (Insect son01–son25) with no focal or external data, an open target (e.g. unsupervised clustering of the unlabeled soundscapes). In hindsight we should also have tried domain-specific embeddings (e.g. Perch [32]), which were strong on the private board. Finally, this is a fast-paced competition study, carried out under tight time and compute budgets: with a hard submission deadline and limited GPU hours, we prioritized *selecting and combining* high-scoring components under heavy noise over running the controlled, repeated experiments that would isolate why each works. Several reported effects are therefore observational and warrant fuller, controlled follow-up.

8.2. Conclusion

We presented a Xeno-Canto-pretrained SED system combining a focal→soundscape MixUp bridge, external in-domain data, and semi-supervised noisy-student distillation. The leaderboard is dominated by seed noise of the same order as the component gains, so our central effort was systematic noise management: reducing variance at the weight, ensemble, and prediction levels and reading every effect against that noise rather than trusting any single run.

Two contributions came out of this. First, by analyzing pseudo-label and model similarity we engineered a structurally diverse ensemble whose blend climbed strongly on the public board; where multi-round distillation gave only a limited private gain, that same engineered diversity absorbed the volatility and held the private score steady. Second, our one architectural addition — the class-interaction module, probed across multiple models — proved *self-recalibrating*: a reproducible per-class domain recalibration that boosts the under-represented non-avian taxa.

Acknowledgments

We thank the BirdCLEF and LifeCLEF organizers, the Cornell Lab of Ornithology, and Kaggle for hosting the challenge and releasing the data, and the wider Kaggle community for the public notebooks and discussions that taught the author this field.

Declaration on Generative AI

During the preparation of this work, the author used an agentic coding assistant (Claude / Claude Code) and other AI-assisted tools to aid in programming, debugging, experiment automation and monitoring, data analysis and visualization, literature review, and grammar/spelling and LaTeX formatting, as described in Section 7. All experiments were designed and configured by the author, all results were manually verified, and all scientific interpretations and conclusions reflect the author's own judgment. The author takes full responsibility for the accuracy, integrity, and originality of the work presented in this paper.

References

[1] S. Kahl, T. Denton, L. Sugai, L. Piatti, W. H. Arruda Moreno, M. M. Barbosa, M. Eibl, C. Z. Fieker, C. M. Garcia, D. L. Hokama Sousa, J. E. d. A. Júnior, R. C. Kridler, M. Lasseck, A. V. d. Melo,

H. Müller, M. d. O. Neves, M. G. d. Reis, L. K. Sampaio Teles, K.-L. Schuchmann, J. L. M. M. Sugai, K. T. P. Azevedo, P. d. N. Lopes, M. I. Marques, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2026: Acoustic species identification in the Pantanal, South America, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.

[2] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: AI challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2026.

[3] S. Kahl, T. Denton, L. Sugai, L. Piatti, R. Holbrook, H. Klinck, A. Oldacre, BirdCLEF+ 2026, https://kaggle.com/competitions/birdclef-2026, 2026. Kaggle.

[4] D. Stowell, Computational bioacoustics with deep learning: a review and roadmap, PeerJ 10 (2022) e13152.

[5] J. S. Cañas, S. Kahl, T. Denton, M. P. Toro-Gómez, S. Rodriguez-Buritica, J. L. Benavides-Lopez, J. S. Ulloa, P. Caycedo-Rosales, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the middle magdalena valley, colombia, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, 2025.

[6] N. Babych, 1st place solution: Multi-iterative noisy student is all you need, Kaggle BirdCLEF+ 2025 competition writeup, 2025. https://www.kaggle.com/competitions/birdclef-2025/writeups/nikita-babych-1st-place-solution-multi-iterative-n.

[7] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, BirdNET: A deep learning solution for avian diversity monitoring, Ecological Informatics 61 (2021) 101236.

[8] M. Tan, Q. Le, EfficientNetV2: Smaller models and faster training, in: International Conference on Machine Learning (ICML), PMLR, 2021, pp. 10096–10106.

[9] A. Brock, S. De, S. L. Smith, K. Simonyan, High-performance large-scale image recognition without normalization, in: International Conference on Machine Learning (ICML), PMLR, 2021, pp. 1059–1071.

[10] S. Adavanne, T. Virtanen, Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network, arXiv preprint arXiv:1710.02998 (2017).

[11] Z.-M. Chen, X.-S. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5177–5186.

[12] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, ICML, volume 3, 2013.

[13] Q. Xie, M.-T. Luong, E. Hovy, Q. V. Le, Self-training with noisy student improves ImageNet classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10687–10698.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations (ICLR), 2018.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, arXiv preprint arXiv:1904.08779 (2019).

[16] N. Reimers, I. Gurevych, Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017.

[17] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, et al., Accounting for variance in machine learning benchmarks, in: Proceedings of Machine Learning and Systems (MLSys), 2021.

[18] D. Picard, torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision, 2021. ArXiv:2109.08203.

[19] R. Wightman, PyTorch image models, https://github.com/rwightman/pytorch-image-models, 2019.

doi:10.5281/zenodo.4414861.

[20] F. Radenović, G. Tolias, O. Chum, Fine-tuning CNN image retrieval with no human annotation, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2018) 1655–1668.

[21] J. S. Cañas, M. P. Toro-Gómez, L. S. M. Sugai, et al., AnuraSet: A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring, Scientific Data 10 (2023) 771.

[22] M. Faiß, B. Ghani, D. Stowell, InsectSet459: an open dataset of insect sounds for bioacoustic machine learning, arXiv preprint arXiv:2503.15074 (2025).

[23] L. Rauch, R. Schwinger, M. Wirth, R. Heinrich, et al., BirdSet: A large-scale dataset for audio classification in avian bioacoustics, arXiv preprint arXiv:2403.10380 (2024).

[24] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[25] B. T. Polyak, A. B. Juditsky, Acceleration of stochastic approximation by averaging, SIAM Journal on Control and Optimization 30 (1992) 838–855.

[26] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.

[27] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, A. G. Wilson, Averaging weights leads to wider optima and better generalization, in: Conference on Uncertainty in Artificial Intelligence (UAI), 2018.

[28] V. Sydorskyi, F. 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 – Conference and Labs of the Evaluation Forum, 2025.

[29] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning (ICML), PMLR, 2017, pp. 1321–1330.

[30] ONNX Runtime developers, ONNX Runtime: cross-platform, high performance ML inferencing, https://onnxruntime.ai, 2021.

[31] OpenVINO Toolkit Contributors, OpenVINO toolkit, https://github.com/openvinotoolkit/openvino, 2024.

[32] B. Ghani, T. Denton, S. Kahl, H. Klinck, Global birdsong embeddings enable superior transfer learning for bioacoustic classification, Scientific Reports 13 (2023) 22876.