Perch-Embedding Sequence Models and Rank Fusion for BirdCLEF+ 2026 Multi-Taxa Sound Identification*

Zhengyang Ren[1,*]

[1]Jiaying University, Meizhou, Guangdong, China

Abstract

This paper describes the system our team submitted to the BirdCLEF+ 2026 challenge, which asks for the identification of bird and other faunal species from long-duration soundscape recordings under background noise, overlapping vocalisations, severe class imbalance, and a domain shift between focal training recordings and passively monitored test soundscapes. Rather than training spectrogram classifiers from scratch, our solution follows a recipe that many high-performing public solutions converged on in 2026: it builds on the pretrained Perch v2 bioacoustic foundation model as a frozen feature extractor and combines lightweight sequence models on top of it. Concretely, the system is an ensemble of two complementary branches. The first treats each one-minute soundscape as a twelve-window sequence of Perch v2 embeddings and logits and models it with a bidirectional selective state-space network equipped with class prototypes (ProtoSSM), then refines the output with site/hour activity priors, per-class multilayer-perceptron probes, and a residual state-space correction. The second is a distilled sound-event-detection branch, the mean of five ONNX folds. The two branches are combined by class-wise percentile-rank fusion and a taxonomy- and time-aware post-processing stage, and all inference runs CPU-only through ONNX Runtime within the competition budget. Our final submission reached a **public score of 0.955 and a private score of 0.950**, placing **39th of 4,094 teams** and earning a **silver medal** on the BirdCLEF+ 2026 leaderboard. As this work adapts several publicly released 2026 notebooks, we frame our contribution as a controlled integration and empirical evaluation of these Perch-based components under the competition's CPU constraint: a cross-seed out-of-fold ablation and a controlled fusion-rule comparison quantify the Perch-embedding branch (the cheap activity priors, not the sequence model, drive most of the in-domain gain, and the fusion rule has only a small effect, with percentile-rank fusion at least as robust as averaging), the two-branch fusion is validated at the submission level, and we analyse where the system fails through the lens of Perch label coverage.

Keywords

BirdCLEF+ 2026, bioacoustic classification, soundscape recognition, foundation-model embeddings, Perch, state-space sequence models, rank fusion, ensemble of solutions, CPU-efficient inference

1. Introduction

Identifying species from passive acoustic monitoring (PAM) data is an important tool for biodiversity assessment and ecological monitoring. PAM allows continuous, low-cost observation of remote habitats, but the resulting recordings are difficult to annotate at scale. Compared with the manually collected focal recordings that dominate training corpora, long-duration soundscape recordings are far more challenging: they contain substantial background noise, multiple temporally overlapping species, heterogeneous recording conditions, and only weak temporal annotations. The BirdCLEF series [1, 2, 3, 4, 5] has framed this as a recurring benchmark, and the BirdCLEF+ 2026 edition continues to focus on recognising species from soundscape audio, requiring models trained mostly on focal recordings to generalise to unseen acoustic environments.

A central difficulty is the *domain shift* between training audio, which typically isolates a single calling species, and the test soundscapes, which capture realistic, cluttered acoustic scenes. Earlier BirdCLEF solutions addressed this by treating audio as a computer-vision problem — converting waveforms to mel-spectrograms and training convolutional or transformer backbones from scratch [7, 8],

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

*Notebook for the BirdCLEF Lab at CLEF 2026.

*Corresponding author.

d58350215@gmail.com (Z. Ren)

0009-0001-3403-8942 (Z. Ren)

then ensembling several models and applying careful inference-time post-processing. The 2026 edition, however, saw many high-performing public solutions converge on a different recipe. Large pretrained bioacoustic foundation models, in particular Google's Perch v2 [9], now provide per-window embeddings and logits that transfer remarkably well to new recording regions, so the most effective public solutions *freeze* such a model as a feature extractor and learn only a small classifier or sequence model on top of its embeddings, before ensembling several of these solutions together. Our system follows exactly this line.

Our submission is an *ensemble of solutions* that rank-fuses two complementary branches built on top of the pretrained Perch v2 backbone:

- **A Perch-embedding sequence-modelling branch.** Each one-minute soundscape is encoded by Perch v2 into a sequence of twelve 5-second-window embeddings and logits, which a bidirectional selective state-space network with class prototypes (ProtoSSM) [17] models as a temporal sequence. Its output is refined with site/hour activity priors, per-class multilayer-perceptron probes over the embeddings, and a residual state-space correction.
- **A distilled sound-event-detection branch.** A separately distilled five-fold sound-event detector [12], exported to ONNX, provides an event-level view that is less tied to the Perch label mapping.

The two branches are combined by class-wise percentile-rank fusion and a taxonomy- and time-aware post-processing stage, and the whole pipeline runs CPU-only through ONNX Runtime within the competition's inference budget.

This pipeline adapts several publicly released BirdCLEF+ 2026 notebooks [18, 19, 20]. Rather than proposing a new architecture, we frame our contribution as a *controlled integration and empirical study* of Perch-based components under the competition's CPU constraint. Concretely:

- we assemble a Perch-embedding sequence branch (ProtoSSM, activity priors, MLP probes, ResidualSSM) and a distilled-SED branch into a single CPU-only pipeline, and tune the rank-fusion weights and post-processing within the runtime budget;
- we quantify each Perch-embedding component in a cross-seed out-of-fold ablation (Section 5.4) — finding that the activity priors, not the sequence model, carry most of the in-domain gain — and compare fusion rules in a controlled study, with the two-branch fusion itself validated at the submission level;
- we analyse the system's failure modes through Perch label coverage — 28 target species have no Perch logit and must rely on the SED branch and the priors (Section 6.3).

For reference we also evaluated a publicly shared OpenVINO inference baseline [21]; our rank-fusion ensemble clearly outperformed it. This working note was prepared by one team member, documenting the team's final solution and a post-submission analysis of the runs.

2. Related Work

From spectrogram CNNs to foundation-model embeddings. The traditional strategy for BirdCLEF treats audio as a computer-vision problem: signals are converted into mel-spectrograms and classified with CNN or transformer backbones trained from scratch [7, 8], typically combined into ensembles [6]. More recently, large pretrained bioacoustic models — BirdNET [6] and especially the self-supervised Perch family [9] — have been shown to produce embeddings that transfer well across recording domains. In the 2026 edition this shifted the prevailing public recipe: instead of training image backbones, the strongest publicly released notebooks use a *frozen* Perch v2 model to embed each audio window and learn only a lightweight head or sequence model on top, which is far cheaper to train and naturally fits the CPU-only inference budget. Our system is built entirely on this paradigm.

Sequence modelling of soundscapes. Because a one-minute soundscape is a temporal sequence of windows, modelling the *ordering* of per-window features can recover continuity that a per-window classifier ignores. Selective state-space models (SSMs) such as Mamba [17] provide an efficient, attention-free way to model such sequences with linear-time recurrence. The ProtoSSM and ResidualSSM components we adopt are bidirectional selective SSMs operating over the twelve Perch windows of each recording, augmented with class prototypes and a residual correction head.

Domain shift, pseudo-labels, and distillation. To cope with the focal-to-soundscape domain shift, prior BirdCLEF solutions [10] have relied on *pseudo-labelling* of target-location soundscapes [11] and on *knowledge distillation* [12], in which soft targets from a strong model reduce the effective label noise of weakly annotated clips; cyclic semi-supervised relabelling has achieved top results in recent editions [13, 14]. The distilled sound-event-detection branch of our ensemble follows this idea: it is a compact detector distilled from a Perch teacher and trained on the labelled soundscapes.

Ensembling, rank fusion, and efficient inference. Combining heterogeneous predictors is standard in BirdCLEF, but averaging raw probabilities is sensitive to calibration mismatch between models. *Rank* (percentile) fusion sidesteps this by blending the *orderings* of the predictors rather than their raw scores, which is the fusion rule our final submission uses. The competition's strict inference budget — a CPU-only notebook with limited runtime — has long motivated efficient deployment, for example CNN ensembles accelerated with OpenVINO [15]; we instead run every component through ONNX Runtime on CPU. Finally, *stochastic weight averaging* [16], which averages weights near convergence to find flatter minima, is used to stabilise our sequence-model training.

3. Task and Data Overview

3.1. Task Description

The BirdCLEF+ 2026 challenge targets the Brazilian Pantanal — a tropical wetland in Mato Grosso do Sul, with passive recorders deployed roughly between 16.5°–21.6°S and 55.9°–57.6°W — and is multi-taxonomic: the goal is to recognise which of 234 species — spanning birds, amphibians, mammals, reptiles, and insects — are vocalising in passively recorded soundscapes. The training data combine 35,549 short, focal-style recordings of individual species (sourced from Xeno-canto and iNaturalist) with 10,658 one-minute training soundscapes, of which only 66 carry expert frame-level annotations. Crucially, only 206 of the 234 target species are represented in the focal training set, so the remaining 28 species must be recognised with no focal examples at all. The hidden test set comprises roughly 600 one-minute field soundscapes. All audio is provided at a 32 kHz sampling rate in `.ogg` format, and a prediction is required for every 5-second segment. This focal-to-soundscape format gap, compounded by strong class imbalance and a long tail of rare species, is the central challenge; to bridge it we segment audio into fixed-length windows and convert each window into a time–frequency representation for both training and inference.

3.2. Dataset Statistics

Table 1 summarises the data splits and Table 2 the taxonomic composition of the 234 target species. Counts are taken from the official metadata (`train.csv`, `taxonomy.csv`); soundscape durations follow from the fixed one-minute length, while the focal-recording duration is left blank as it is not provided in the metadata.

Sources and labels. The focal recordings come from two public repositories — Xeno-canto (64.8%) and iNaturalist (35.2%) — and were contributed by 4,017 distinct recordists, so recording devices and conditions vary widely even within the training set. About 12% of focal recordings additionally carry

Table 1
Dataset summary for BirdCLEF+ 2026. Counts are taken from the official release; soundscape durations are approximate (one minute each).

Split / source	Files	Duration	Notes
Training audio (focal)	35,549	—	Xeno-canto + iNaturalist; 206/234 species
Training soundscapes	10,658	∼178 h	1-min; 66 expert-annotated
Hidden test soundscapes	∼600	∼10 h	1-min PAM recordings
Target classes	234	—	multi-taxonomic

Table 2
Taxonomic composition of the 234 target species. "Focal" is the number of species in each class that have at least one focal training recording; 28 species (mostly insects) have none.

Taxonomic class	Species	Focal	Focal recordings
Aves (birds)	162	162	34,799
Amphibia	35	32	451
Insecta	28	3	199
Mammalia	8	8	99
Reptilia	1	1	1
Total	234	206	35,549

one or more secondary (background) species labels, a first hint of the dense polyphony that characterises the soundscapes (Section 3.3). The class imbalance is extreme: the most common species has 499 focal recordings and the rarest a single one (imbalance ratio ≈500), with a median of 125 recordings per species.

3.3. Class Imbalance and Domain Shift

Three properties of the data dominate the modelling difficulty (Figure 1). First, the per-species distribution is strongly *long-tailed*: the number of focal recordings per species ranges from 1 to 499 (median 125), and 25 species have fewer than ten recordings, so naive training under-fits the tail. Second, the dataset is severely imbalanced *across taxa*: birds account for 34,799 of the 35,549 focal recordings (≈98%), while amphibians, insects, mammals, and reptiles together contribute only about 750, and reptiles are represented by a single recording. Compounding this, only 206 of the 234 target species have any focal recording at all — in particular just 3 of the 28 target insect species — so a substantial fraction of classes must be learned almost entirely from soundscapes. Third, there is a pronounced *domain shift* between the clean, single-species focal recordings that dominate training and the passively recorded test soundscapes. Together these properties motivate the per-class weighting, the rare-class post-processing, and the in-domain soundscape training that our pipeline applies (Section 4).

The domain shift manifests along three concrete, measurable axes. The most visible is acoustic (Figure 2); two further quantitative axes — geographic spread and multi-label density — are documented in Appendix B.

Geographic shift. The focal recordings are sourced globally from Xeno-canto and iNaturalist and span every continent, whereas the test recorders sit in a single $\sim 5^\circ \times 5^\circ$ box in the Pantanal: only *2.4%* of the 35,549 geolocated focal recordings fall inside that box (Appendix B, Figure 5). The model must therefore generalise from a globally diverse training set to one specific acoustic environment.

Multi-label density. Focal clips are essentially single-species, but the labelled test-style soundscapes are densely polyphonic: a labelled 5-second chunk contains *4.2 species on average* and up to ten simultaneously (Appendix B, Figure 6). This dense polyphony is exactly what the sequence model and the

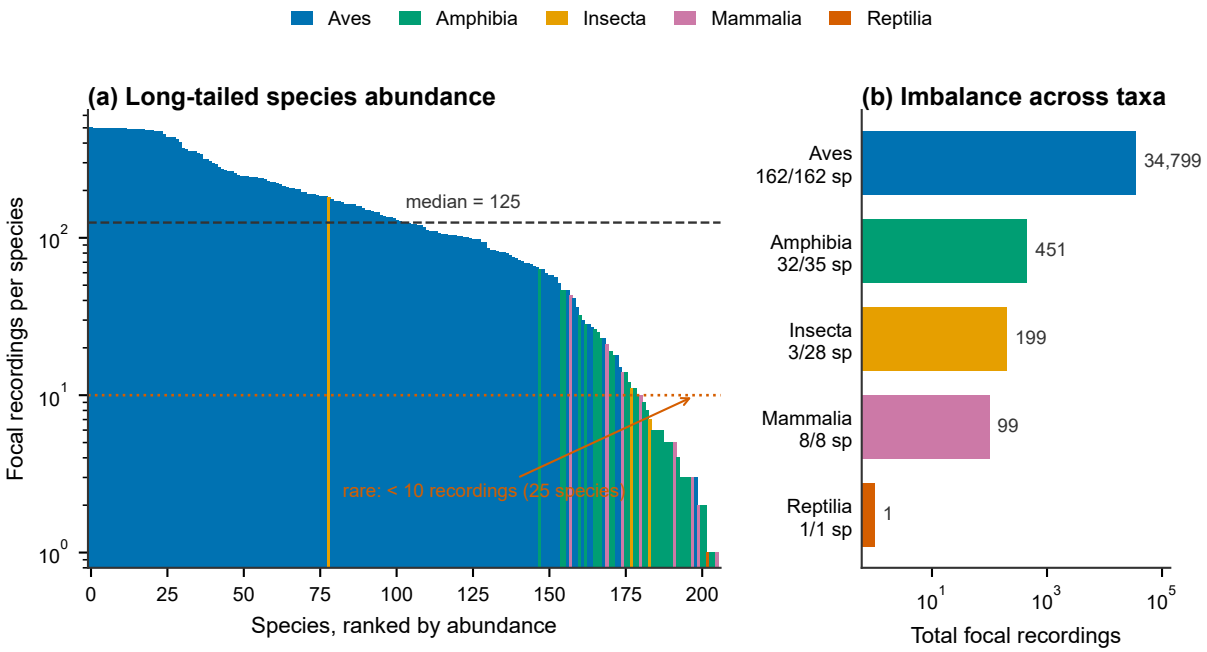

Figure 1: Data imbalance in BirdCLEF+ 2026. (a) Focal recordings per species (log scale), ranked by abundance and coloured by taxonomic class: the distribution is long-tailed (median 125; 25 species below the dotted rare-species threshold of ten recordings). (b) Total focal recordings per taxonomic class (log scale); birds dominate, and the y-axis labels report how many of each class's target species actually have focal recordings (e.g. only 3 of 28 insect species).

multi-label SED branch must resolve at test time.

Acoustic shift. The recording conditions differ sharply. Figure 2 contrasts focal and soundscape log-mel spectrograms of the same species: the focal calls are loud and largely isolated, while the soundscape chunks embed the same call in continuous background energy and overlapping vocalisations from other species. Closing this acoustic gap is why the sequence models are trained directly on the in-domain soundscapes (Section 4.4).

3.4. Audio Representation

All audio is processed at the competition's native 32 kHz sampling rate. Each one-minute recording is divided into twelve non-overlapping 5-second windows — the unit scored by the competition — and the two branches of our system (Section 4) consume these windows through different front ends.

Perch branch (raw waveform). The Perch v2 backbone ingests each 5-second window directly as a raw 32 kHz waveform and returns a 1536-dimensional embedding together with its own classification logits; no hand-designed spectrogram is computed on this path. These per-window embeddings and logits are the inputs to the sequence models of Section 4.4.

SED branch (log-mel spectrogram). The distilled sound-event detector instead consumes a log-mel spectrogram of each window, computed with 256 mel bins, a 2048-point FFT and a 512-sample hop over 20 Hz–16 kHz, converted to decibels (top 80 dB) and standardised per instance. Table 10 in Appendix A lists the exact settings.

Handling class imbalance. Because the metric weights all classes equally, the long tail (Section 3.3) is addressed throughout the pipeline rather than by a single sampler: the per-class probes use class-frequency balancing, the first-pass blend up-weights the ProtoSSM branch for species that Perch can

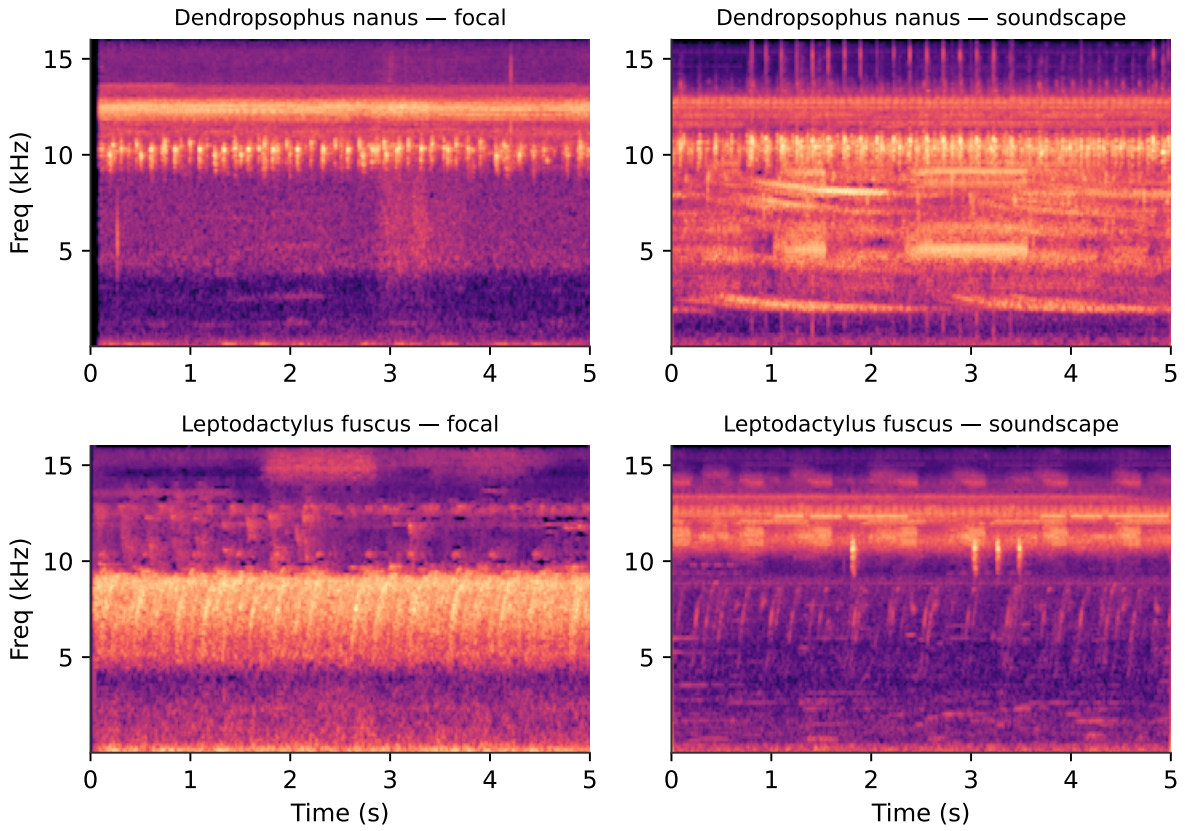

Figure 2: Acoustic domain shift. Log-mel spectrograms (5 s) of two target amphibian species in a focal recording (left) and in a passively recorded soundscape (right). The focal calls are clean and prominent, whereas the soundscape versions are embedded in continuous background energy and overlapping vocalisations — the conditions the model must handle at test time.

recognise directly and the acoustic evidence for the rest, and a dedicated post-processing gate suppresses spurious activations of the rarest taxa (Section 4.6).

3.5. External Data and Pretrained Components

Pretrained backbone. The system is built on *Perch v2*, a large bioacoustic foundation model pretrained by Google on a broad corpus of recordings [9]. We use it strictly as a frozen feature extractor (ONNX export, no fine-tuning); only the small sequence models and probes on top of it are trained, on the organiser-provided labelled soundscapes. The distilled sound-event detector is a public five-fold model [20] distilled from a Perch teacher.

External data. Beyond Perch's own pretraining corpus we add no external recordings: the sequence models and probes are fitted only on the official BirdCLEF+ 2026 labelled `train_soundscapes`. To stay within the CPU budget we reuse a publicly shared cache of pre-computed Perch embeddings for the training soundscapes, which is equivalent to running the frozen backbone ourselves.

Open-source components. Our pipeline adapts several public BirdCLEF+ 2026 notebooks — the rank-fusion inference notebook [18], the Perch-embedding sequence-model reproduction [19], and the distilled SED [20]. A separate public OpenVINO inference notebook [21] was evaluated only as a reference baseline and is not part of the submitted system. All of these are acknowledged accordingly (Section 7.2).

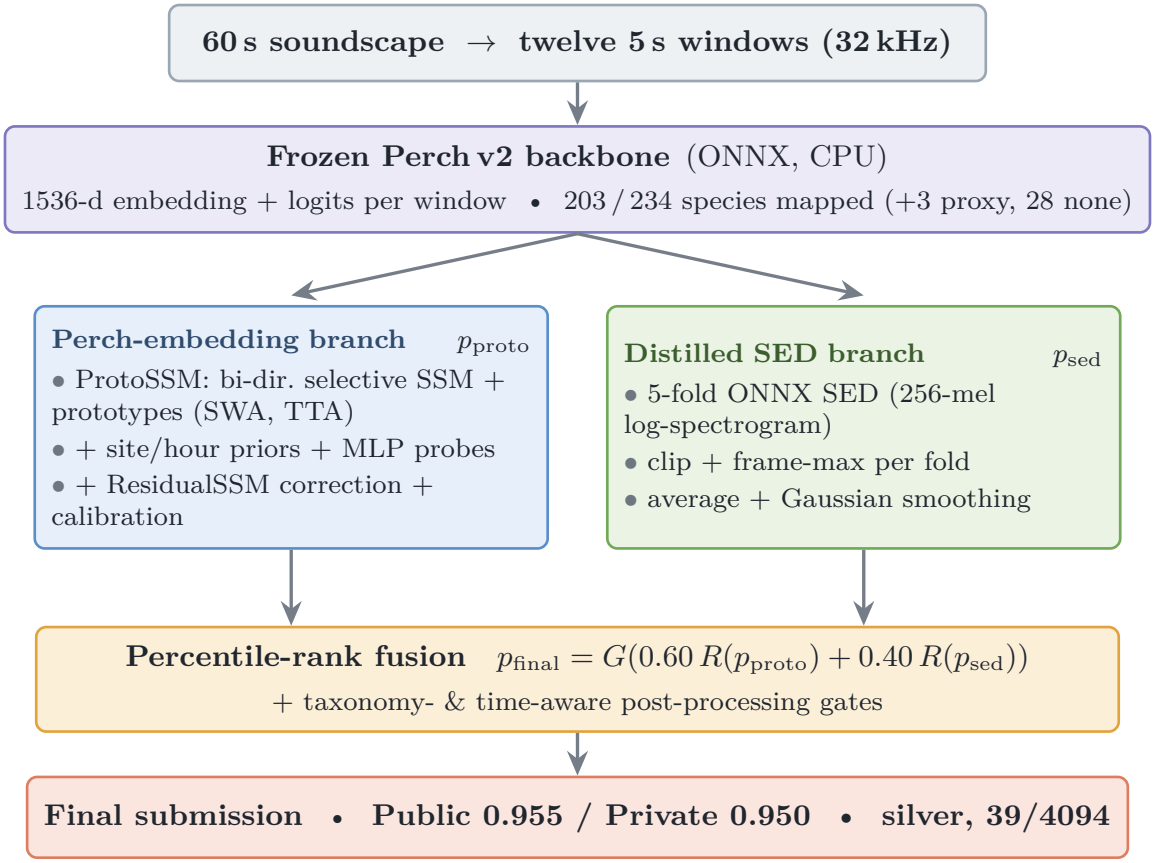

Figure 3: Overview of the BirdCLEF+ 2026 pipeline, adapted from public 2026 notebooks [18, 19]. A frozen Perch v2 backbone encodes each one-minute soundscape into a length-12 sequence of embeddings and logits. The Perch-embedding branch models this sequence with a bidirectional selective state-space network with class prototypes (ProtoSSM), refined by site/hour priors, per-class MLP probes and a residual state-space correction; the distilled-SED branch averages five ONNX folds. The two branches are combined by class-wise percentile-rank fusion (0.60/0.40) and a taxonomy- and time-aware post-processing stage, and all inference runs CPU-only via ONNX Runtime.

4. Method

4.1. Pipeline Overview

Our system is an *ensemble of solutions* built on a frozen Perch v2 backbone. It produces a prediction for every 5-second window of every test soundscape through two complementary branches that are combined by rank fusion:

$$p_{\text{final}} \;=\; G\big(\, w_{\text{proto}}\, R(p_{\text{proto}}) \;+\; w_{\text{sed}}\, R(p_{\text{sed}})\,\big)\,, \qquad (w_{\text{proto}}, w_{\text{sed}}) = (0.60,\, 0.40), \tag{1}$$

where $R(\cdot)$ is a class-wise percentile-rank transform, p_{proto} is the output of a Perch-embedding sequence-modelling branch (Section 4.4), p_{sed} is the output of a distilled sound-event-detection branch (Section 4.5), and $G(\cdot)$ is a taxonomy- and time-aware post-processing operator (Section 4.6). Both branches read the same twelve 5-second Perch windows per recording (Section 4.3); only the small models on top of Perch are trained, on the labelled training soundscapes. The whole pipeline runs CPU-only through ONNX Runtime. Figure 3 gives an overview.

We adapt the public BirdCLEF+ 2026 rank-fusion notebook [18] and the Perch-sequence reproduction it builds on [19]; we describe the method in enough detail to be self-contained and credit those sources throughout.

4.2. Validation Strategy

Reliable local validation is difficult here: the hidden test set is passively monitored soundscapes, and the training soundscapes carry only partial frame-level annotations. In total 66 files contain 739 distinct labelled 5-second chunks (Appendix B), of which 59 files are *fully* labelled — all twelve windows — giving the 708 windows (covering 71 active species) that we use for out-of-fold (OOF) validation. We treat the OOF macro-AUC on these windows and the public-leaderboard (LB) score as complementary, partially unreliable signals.

Cross-validation design. The sequence models and probes are fitted with grouped k-fold splits over the labelled soundscapes, grouped by recording file so that no file contributes windows to both sides of a fold. Because the labelled set is small and dominated by a handful of common species, the OOF macro-AUC is high-variance and well below the leaderboard (around 0.82 for the full Perch-embedding branch; Section 5.4); we use it mainly to compare the *relative stability* of choices rather than as an absolute score.

CV–LB consistency. As in the 2025 edition, the correlation between local CV and the public LB weakens once scores become high [14]. We therefore used OOF AUC to select robust hyper-parameters (for example the residual-correction strength, Section 4.4) and consulted the public LB cautiously for the final fusion weights, while remaining aware that the public LB is itself only a subset of the hidden test set (we revisit the resulting public–private behaviour in Section 6.2).

4.3. Perch v2 Backbone and Species Mapping

Frozen feature extractor. Each 5-second window is passed through Perch v2 [9], exported to ONNX and run on CPU, which returns a 1536-dimensional embedding and a vector of Perch class logits. A one-minute recording thus becomes a length-12 sequence of (embedding, logit) pairs together with its site identifier and UTC hour, both parsed from the file name. To stay within the runtime budget the embeddings for the training soundscapes are read from a pre-computed cache (Section 3.5); the backbone itself is never fine-tuned.

Mapping Perch labels to the target taxa. Perch was not trained on the exact 234-species Pantanal label set, so its logits must be projected onto the target classes. We align species by scientific name: **203 of the 234 target species map directly to a Perch logit**. Of the remaining 31, three are given a *genus proxy* — the maximum logit over Perch classes of the same genus — and **28 species have no Perch signal at all** (almost exactly the 28 focal-free species of Table 2, mostly insects). These 28 are the hardest classes and must be recovered, if at all, from the SED branch and the activity priors. We also apply per-taxon temperature scaling to the Perch logits (0.95 for the "texture" taxa Amphibia and Insecta, 1.10 otherwise) to balance their calibration before fusion.

4.4. Perch-Embedding Sequence Branch

This branch (p_{proto}) turns the per-window Perch features into refined per-window probabilities through a sequence model and several light corrections.

ProtoSSM. The core model treats each recording as a 12-step sequence and processes it with a bidirectional *selective state-space* network in the style of Mamba [17]: two forward and two backward selective-SSM layers (model width 128, state size 16) with a cross-attention block, fed by a linear projection of the 1536-d embedding plus learned positional, site and hour embeddings. Classification is *prototype*-based: the hidden state of each window is compared (cosine similarity with a learned temperature) against one learned prototype per class, and a per-class sigmoid gate blends this prototype score with Perch's own logit; the prototypes are initialised from class-mean embeddings. Training uses

a positive-weighted binary cross-entropy plus a 0.15-weighted MSE term that keeps outputs close to the Perch logits, OneCycle cosine scheduling, and stochastic weight averaging [16] over the last 35% of training; at inference we average predictions over temporal shifts $\{0, \pm 1, \pm 2\}$ and a time-reversed pass (test-time augmentation).

Activity priors. Vocal activity is strongly structured by location and time of day. We build global, per-site, per-hour and joint site-hour occurrence priors from the labelled soundscapes, smooth the hourly tables with a circular Gaussian ($\sigma = 1.5$ h, wrapping at midnight), and add them in logit space with a shrinkage that down-weights sparsely observed buckets.

MLP probes. For every sufficiently frequent class we fit a small multilayer-perceptron probe on a 64-dimensional PCA of the Perch embeddings (retaining ≈81% of the variance) together with sequential summary features (previous/next window and per-file mean, max and standard deviation). The probes use class-frequency balancing and a wider hidden layer for common classes; their output is blended with the prior-adjusted evidence.

First-pass blend and residual correction. The ProtoSSM output and the prior/probe-adjusted Perch evidence — both logit-space scores — are blended per class in logit space (a sigmoid is applied to the corrected scores further below),

$$\text{first-pass}_c = \alpha_c \, z_c^{\text{proto}} + (1 - \alpha_c) \, z_c^{\text{adj}}, \qquad \alpha_c = \begin{cases} 0.60 & \text{Perch-mapped species,} \\ 0.35 & \text{unmapped species,} \end{cases} \tag{2}$$

so that species Perch knows trust the sequence model more and the rest trust the adjusted acoustic evidence more. A second selective-SSM model (*ResidualSSM*) then predicts a residual correction Δ to this first pass; its strength is chosen on OOF macro-AUC from the grid $\{0.10, 0.15, \ldots, 0.40\}$, which selected 0.40. Finally the corrected logits are temperature-scaled and passed through a sigmoid, then refined by per-class isotonic calibration and threshold reshaping, a file-level top-2 confidence scaling, a rank-aware power scaling, and an adaptive temporal smoothing, yielding p_{proto}.

4.5. Distilled Sound-Event-Detection Branch

The second branch (p_{sed}) provides an event-level view that does not depend on the Perch label mapping. It is a publicly released sound-event detector distilled from a Perch teacher [20, 12], supplied as five ONNX folds. Each 5-second window is converted to a 256-bin log-mel spectrogram (Section 3.4); each fold returns clip-level and frame-level logits, which we combine as $\frac{1}{2}\sigma(\text{clip}) + \frac{1}{2}\sigma(\max_t \text{frame})$ and average over the five folds, followed by a light Gaussian smoothing ($\sigma = 0.65$ window) along the time axis. Because the SED scores the 234 target classes directly rather than through the Perch label mapping, it can fire for species that Perch does not represent, which is exactly where it complements the ProtoSSM branch.

4.6. Rank Fusion, Post-processing, and Inference

Percentile-rank fusion. The two branches are calibrated very differently, so we fuse their *orderings* rather than their raw scores: each branch is mapped to class-wise percentile ranks $R(\cdot)$ and combined with weights $0.60/0.40$ in favour of the ProtoSSM branch (Eq. 1). Rank fusion was consistently more robust than averaging raw probabilities.

Taxonomy- and time-aware gates (G). A sequence of conservative post-processing gates then adjusts the blend: (i) a *noise-suppression* gate that trusts ProtoSSM where it is confident but the SED strongly disagrees; (ii) a *temporal-continuity* gate that protects calls spanning several windows, using a fat-tailed (t-distribution) kernel over a 35-second context; (iii) a *spike-preservation* gate that keeps brief,

high-confidence SED detections the sequence model missed; (iv) *sonotype mirroring*, which max-pools probabilities across *sonotype* groups — small sets of target labels whose vocalisations are acoustically similar and hard to separate in the spectrogram domain; and (v) an *adaptive rare-class* gate that mildly suppresses low-confidence predictions for the 44 species in the rare taxa (Amphibia, Mammalia, Reptilia). Predictions are finally clipped to $[0, 1]$ and written in the per-window `row_id` format expected by the competition.

CPU-only inference. Every learned component — the Perch backbone and the five SED folds — is executed through ONNX Runtime on CPU (four intra-op threads), and the lightweight sequence models and probes are trained on the fly from the cached training-soundscape embeddings in well under two minutes (Section 5.5). No GPU or internet access is used at submission time, matching the competition's constraints.

5. Experiments

5.1. Experimental Setup

We evaluate the system with the out-of-fold (OOF) macro-AUC on the labelled soundscapes together with leaderboard feedback; the validation design and its reliability are discussed in Section 4.2. The official metric is the macro-averaged ROC-AUC — the one-vs-rest ROC-AUC computed per species and averaged across classes, excluding species with no positive label in the test set. The components we study are the Perch-to-target label mapping, the ProtoSSM sequence model and its residual correction, the activity priors and MLP probes, the distilled-SED branch, and the percentile-rank fusion weights and post-processing gates.

5.2. Training Configuration

Table 3 lists the main hyper-parameters of the small models trained on top of the frozen Perch v2 backbone, and Figure 4 shows the training and weighting procedure end to end. The Perch backbone and the distilled SED are used as-is (no training on our side); everything else is fitted on the labelled training soundscapes in well under two minutes on CPU.

5.3. Ensemble Composition

Table 4 summarises the two branches combined in the final submission and their percentile-rank weights. The ProtoSSM branch dominates (0.60) because the Perch v2 embeddings carry most of the discriminative signal for the 203 Perch-mapped species; the distilled SED (0.40) contributes an event-level, Perch-independent view that is essential for species Perch does not represent. Because the competition did not expose per-component scores on the hidden test, we report results at the submission level (Section 5.6) and complement them with the OOF behaviour of individual choices (Sections 6.3–6.4).

5.4. Ablation of the Perch-Embedding Branch

To isolate the contribution of each component — and to separate real gains from noise — we run a controlled out-of-fold (OOF) ablation on the 708 fully-labelled soundscape windows, repeated over five seeds, each with a freshly shuffled 5-fold GroupKFold (grouped by file). Table 5 reports the mean and standard deviation of the macro ROC-AUC over the 71 active species. These OOF scores are computed on a small in-domain set, are high-variance, and are not directly comparable to the leaderboard; only their *relative* ordering is informative. Three points stand out. First, the cheap deterministic corrections do most of the work: raw Perch logits reach 0.739, the *activity priors* add the largest single increment ($+0.046$, well above the seed noise), and the per-class MLP probes add a further $+0.029$. Second, the ProtoSSM sequence model on its own (0.794) is actually *weaker* than this prior/probe path (0.814); only when the two views are blended (the first-pass, 0.817) does the sequence model help, and even then

Table 3
Configuration of the learned components on top of the frozen Perch v2 backbone. The backbone and the distilled SED are pretrained and frozen; only the sequence models, probes and priors are fitted, on the labelled soundscapes.

Component / item	Value
Perch v2 backbone (frozen)	
Output	1536-d embedding + Perch class logits per 5 s window
Execution	ONNX Runtime, CPU, 4 intra-op threads
ProtoSSM (Perch-embedding sequence model)	
Architecture	bidirectional selective SSM, width 128, state 16, $2{+}2$ layers, cross-attn (2 heads)
Head	234 class prototypes (cosine sim.), per-class Perch-logit gate
Loss	pos-weighted BCE $+\,0.15$ MSE-to-Perch
Optimiser	AdamW ($wd{=}10^{-3}$), OneCycle cosine, lr 10^{-3}, 40 epochs
Stabilisation	SWA over last 35%; TTA shifts $\{0, \pm1, \pm2\}$ + time-reversed pass
ResidualSSM correction	
Architecture	bidirectional selective SSM, width 64, state 8
Target / weight	residual to first pass; strength 0.40 (grid 0.10–0.40, by OOF AUC)
MLP probes / activity priors	
Probe input	PCA-64 of Perch embeddings ($\approx$81% var.) + sequential features
Probe net	MLP $(256, 128)$ or $(128, 64)$, class-frequency balanced
Priors	global / site / hour / site-hour; circular-Gaussian hour smoothing ($\sigma{=}1.5$ h)
Distilled SED branch (public, frozen)	
Folds / input	5 ONNX folds; 256-mel log-spectrogram (2048 / 512, 20–16 kHz)
Per-window	$\frac{1}{2}\sigma(\text{clip}){+}\frac{1}{2}\sigma(\max_t \text{frame})$, Gaussian $\sigma{=}0.65$
Fusion	
Rule	class-wise percentile-rank blend, 0.60 proto / 0.40 SED
First-pass α_c	0.60 mapped / 0.35 unmapped species

Table 4
Composition of the final rank-fusion ensemble. The two branches are combined by class-wise percentile rank; the Perch-embedding branch itself aggregates several sub-components (Section 4.4).

Branch	Role	Weight
Perch-embedding branch (ProtoSSM, ResidualSSM, priors, MLP probes)	sequence model over Perch features	0.60
Distilled SED (5-fold ONNX)	event-level, Perch-independent	0.40

the gain over priors-plus-probes is within noise. Third, the ResidualSSM does not improve the branch on this set ($0.817 \rightarrow 0.815$, inside the ±0.02 seed std), confirming that its effect is not statistically distinguishable here. We kept the SSM components because they are part of the public pipeline that produced our best leaderboard score, even though this small OOF set does not demonstrate their value — a gap that underlines how unreliable the in-domain validation is. The distilled-SED branch and the actual two-branch rank fusion cannot be placed on this OOF axis — the SED has no cached training-soundscape audio — so they are evaluated at the submission level (Section 5.6, Table 7); we examine the fusion *rule* on a controlled proxy below.

Which fusion rule? Our final submission combines the two branches by class-wise percentile rank rather than by averaging probabilities. We cannot put the actual ProtoSSM/SED fusion on the OOF axis, but we *can* run a controlled comparison on the two first-pass predictors we evaluate out-of-fold — the

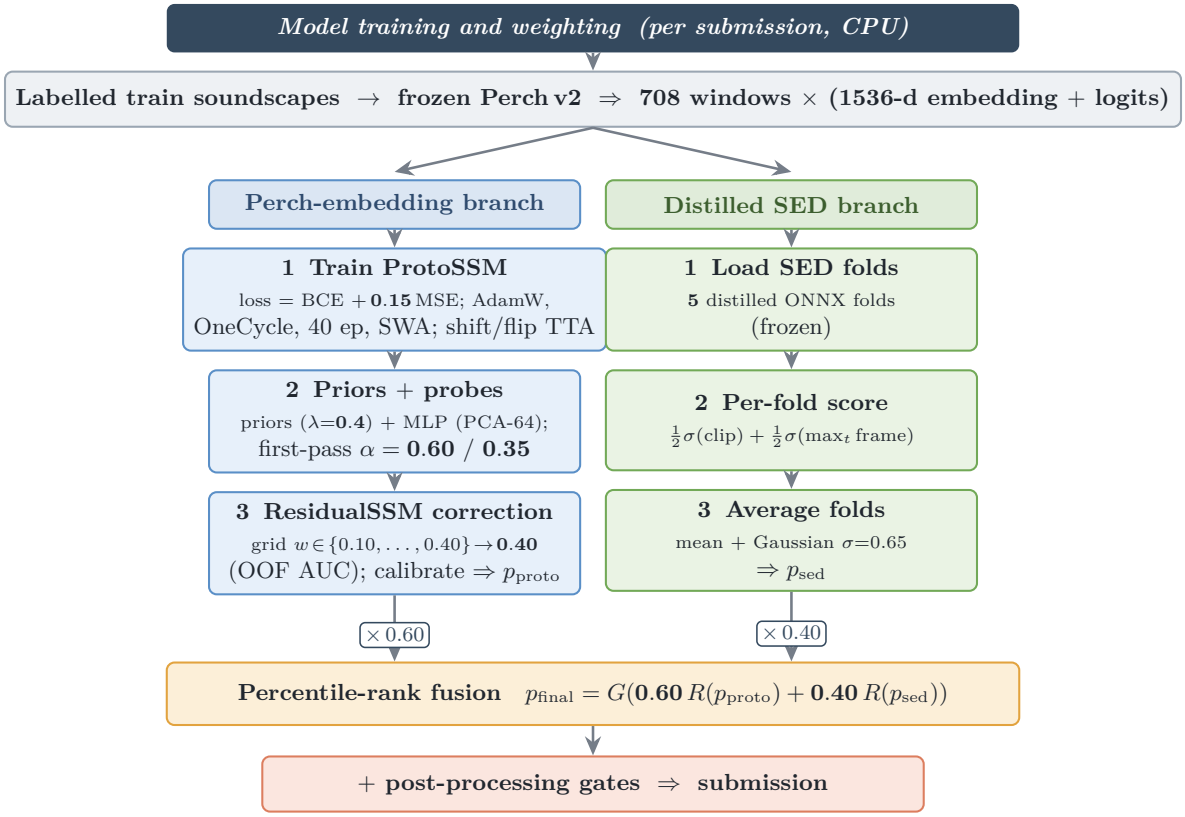

Figure 4: Model training and weighting procedure. The Perch-embedding branch (p_{proto}) trains a ProtoSSM on the cached training-soundscape embeddings, adds activity priors and MLP probes, blends them per class ($\alpha = 0.60$ mapped / 0.35 unmapped), and adds a ResidualSSM correction whose strength (0.40) is grid-selected on OOF macro-AUC. The frozen distilled-SED branch (p_{sed}) averages five ONNX folds. The two branches are combined by class-wise percentile-rank fusion with weights $0.60/0.40$. Bold values are the blend/loss weights.

Table 5
Out-of-fold ablation of the Perch-embedding branch: five seeds, each a freshly shuffled 5-fold GroupKFold (by file) over the 708 fully-labelled windows; macro ROC-AUC over the 71 active species, mean ± std. Each row adds one component to the group above it; the final row is the full branch output p_{proto}. The large std shows the in-domain set is noisy.

Configuration	OOF macro-AUC
Perch logits only (no training)	0.739 ± 0.000
+ activity priors	0.785 ± 0.013
+ per-class MLP probes	0.814 ± 0.015
ProtoSSM sequence model (alone)	0.794 ± 0.029
first-pass blend (ProtoSSM + adjusted Perch)	$\mathbf{0.817 \pm 0.020}$
+ ResidualSSM (full branch, p_{proto})	0.815 ± 0.022

ProtoSSM output and the prior/probe-adjusted Perch evidence — holding the inputs and the $0.60/0.40$ weights fixed and varying only the combination rule (Table 6). Percentile-rank fusion is at least as good as raw-probability averaging on *every* seed (paired mean $+0.0023$, never negative), consistent with its invariance to per-predictor calibration; on these two already similarly-scaled predictors, however, the margin is small and a logit blend did marginally better. Because the real branch fusion combines *more* differently-calibrated predictors (a trained SSM versus a distilled CNN detector), we expect rank fusion's calibration-robustness to matter more there — but that is the part we can only validate at the submission level (Table 7).

Table 6
Fusion-rule comparison on the two first-pass predictors (ProtoSSM and adjusted-Perch evidence, weights $0.60/0.40$): only the combination rule changes. OOF macro-AUC, mean $\pm$ std over five seeds.

Combination rule	OOF macro-AUC
Raw-probability average	0.805 ± 0.024
Logit blend (the pipeline's first pass)	0.813 ± 0.022
Percentile-rank fusion	0.807 ± 0.024

Would a simpler pipeline do as well? Read together, these results raise a question our analysis cannot fully settle. On the OOF set, removing the sequence models entirely — keeping only the priors, probes and the SED branch — costs nothing within noise: the prior/probe path reaches 0.814 against 0.815 for the full Perch-embedding branch (Table 5), and the fusion rule barely moves the score (Table 6). The decisive test is whether this also holds on the hidden test: a *paired* leaderboard submission of an SSM-free pipeline (priors + probes + SED, blended) against the full system would measure the sequence models' true contribution directly. We were unable to run it — the competition closed before this post-hoc analysis was complete — so we flag it as the single most informative next experiment, and, wherever the OOF and leaderboard pictures disagree, we defer to the leaderboard (on which the full pipeline produced our best score).

5.5. Inference Efficiency

BirdCLEF imposes a strict inference budget: submissions run as a CPU-only Kaggle notebook without internet access and within a limited runtime (90 minutes in recent editions), which rewards compact models. Our pipeline meets this comfortably because the expensive backbone is run only once per window and every learned component is executed through ONNX Runtime on CPU (four intra-op threads). On the public dry-run (20 one-minute soundscapes) the complete pipeline — Perch v2 feature extraction, on-the-fly training of the ProtoSSM (13 s) and ResidualSSM (2 s), the MLP probes, and the five-fold distilled-SED branch — runs end-to-end in about three and a half minutes of compute, after roughly two minutes of one-off environment and model loading. Per-file inference is dominated by the Perch backbone ($\approx$2.5 s per one-minute file) and the five SED folds ($\approx$3.3 s per file), so the full hidden-test submission completes well within the CPU time limit. Reusing a cached set of Perch embeddings for the training soundscapes removes the only other sizeable cost.

5.6. Final Submission

Our final submission is the percentile-rank fusion of the ProtoSSM and distilled-SED branches (Table 4) with the taxonomy- and time-aware post-processing of Section 4.6. Table 7 lists the four versions we submitted of this rank-fusion notebook, an earlier rank fusion of our own ensemble-of-solutions blends, and — for reference only — the best configuration of a public OpenVINO starter notebook. Version 2 of the rank fusion was selected as the final submission; it clearly outperforms both the OpenVINO baseline and our earlier blend on both leaderboards.

Table 7
Final leaderboard results (macro ROC-AUC). "Ours — rank fusion" rows are versions of the submitted Perch-embedding / SED rank-fusion notebook; v2 was selected (**bold**). The public OpenVINO starter is listed only as a reference baseline: this is the best-tuned version of that notebook (0.930 public), whereas an early version of the *same* starter scored only 0.791 public / 0.835 private.

Submission	Public LB	Private LB
Public OpenVINO starter (best public baseline)	0.930	0.942
Our ensemble-of-solutions blend (earlier)	0.949	0.940
Ours — rank fusion v1	0.954	0.951
Ours — rank fusion v2 (final)	**0.955**	**0.950**
Ours — rank fusion v3	0.954	0.951
Ours — rank fusion v4	0.952	0.946

6. Results and Discussion

6.1. Main Results

Our final submission — the percentile-rank fusion of the Perch-embedding sequence branch and the distilled-SED branch with taxonomy- and time-aware post-processing — reached a public macro ROC-AUC of **0.955** and a private score of **0.950**, placing **39th of 4,094 teams** and earning a **silver medal** (Table 7). The qualitative picture is consistent with our design: the frozen Perch v2 embeddings carry most of the discriminative signal, the activity priors and per-class probes refine it and carry most of the measurable in-domain gain (Section 5.4), the ProtoSSM sequence model adds temporal context, and the distilled-SED branch adds the event-level evidence needed for species outside Perch's label space. We quantify the contribution of the Perch-embedding components in a controlled out-of-fold ablation (Section 5.4, Table 5); the two-branch fusion weights and post-processing gates, by contrast, were tuned on the public leaderboard, as per-component hidden-test scores were not available.

6.2. Public–Private Gap and Validation Reliability

The public LB is computed on only a subset of the test soundscapes and is a noisy proxy for the private LB. In our case the gap was small and benign: the selected submission scored 0.955 public / 0.950 private (a 0.005 difference), and across all 45 scored submissions the public and private scores moved together along the diagonal (Figure 7; Pearson $r = 0.989$), so the ranking was stable and we did not observe a damaging shake-up. We selected v2 by its public-LB score — the only signal available at submission time — and that v1 and v3 ended marginally higher on the private LB (Table 7) simply reflects that differences within the rank-fusion family are themselves within the noise. This is consistent with our validation strategy (Section 4.2): because the OOF macro-AUC on the small labelled-soundscape set (around 0.82; Table 5) is high-variance and correlates only weakly with the leaderboard — especially at high scores, as also reported for the 2025 edition [14] — we preferred robust choices over components that merely chased public-LB jumps, which limited overfitting to the public subset.

6.3. Error Analysis

A per-window error breakdown on the hidden test is not available, but the structure of the Perch label mapping (Section 4.3) gives a concrete, data-grounded view of where the system is strong and where it must be weak. Of the 234 target species, **203 map directly to a Perch logit** and are the easiest; **31 are unmapped**, of which only **3** recover a weak signal through a genus proxy and **28 have no Perch signal at all**. These 28 are almost exactly the 28 focal-free species of Table 2 — overwhelmingly insects — so they have neither focal training audio nor a Perch logit and can be reached only through the SED branch and the activity priors. Table 8 groups the target classes accordingly. Two further factors compound the difficulty: the rare taxa (Amphibia, Mammalia, Reptilia; 44 species) are handled

by an explicit suppression gate because their sparse evidence is easily swamped by false positives, and densely polyphonic windows (mean 4.2 species, Section 3.3) remain the hardest acoustically. We therefore expect performance to fall off along the Aves → mapped-non-Aves → proxy-only → no-signal axis, with the 28 no-signal species contributing most of the residual error.

Table 8
Difficulty groups of the 234 target species by the evidence available to the system. "Perch logit" is a directly mapped Perch class; "genus proxy" is the maximum logit over same-genus Perch classes; "no signal" species are recoverable only via the SED branch and the activity priors.

Group	Species	Available evidence
Perch-mapped	203	Perch logit + embedding + SED
Genus-proxy only	3	proxy logit + embedding + SED
No Perch signal	28	SED branch + activity priors only
Total	234	

6.4. Negative Experiments

We also record what did *not* help, drawing on the full submission history; Table 9 summarises the main findings with the evidence behind each. Together they point future participants towards the foundation-model features and the fusion rule, and away from re-tuning blend weights or growing the ensemble.

7. Limitations and Conclusions

7.1. Limitations

Several limitations should be acknowledged. First, *validation is only weakly reliable*: the sequence models are fitted and validated on only 59 labelled soundscape files from the single Pantanal region, the OOF macro-AUC (around 0.82) is high-variance and correlates poorly with the leaderboard, so a model tuned on it need not transfer elsewhere. Second, *the work is one of reuse rather than novelty*: the system is built on a frozen Perch v2 backbone and adapts several public 2026 notebooks, so our contribution is the integration, tuning and analysis rather than a new model or training recipe. Third,

Table 9
Negative and low-yield experiments, with the evidence behind each. "Public" denotes the public leaderboard.

What we tried	Outcome
From-scratch / OpenVINO spectrogram route	Plateaued at 0.930 public (an early version scored only 0.791; Table 7); never approached the 0.955 of the Perch route.
Averaging raw probabilities (vs. rank fusion)	In a controlled OOF proxy, rank fusion beat raw averaging on every seed (by $+0.002$ on average; Table 6) and is calibration-invariant, though the margin was small.
Re-weighting the two-branch blend	Diminishing returns — many weightings landed at ≈ 0.949 public, so beyond a sensible 0.60/0.40 split the weights mattered little.
Adding a third heavy sequence-model member	*Timed out* under the CPU runtime limit, so the final design keeps two branches.
ResidualSSM correction strength	In the cross-seed OOF ablation it does *not* improve the branch ($0.817 \rightarrow 0.815$, within the ± 0.02 seed noise; Table 5).

the ablation reaches only the Perch-embedding branch: per-component scores on the hidden test were not available, so the SED branch and the two-branch fusion are attributed at the submission level (and, for the fusion rule, on a controlled OOF proxy) rather than isolated on the private leaderboard. Fourth, *several hyper-parameters were inherited from the public notebooks and only lightly retuned*: the fusion weights, the activity-prior shrinkage and the post-processing gates were taken from those notebooks rather than tuned in a principled, leakage-free way. Finally, *the system is structurally weak on the species Perch cannot represent*: the 28 target classes with no Perch logit (mostly the focal-free insects) depend entirely on the SED branch and the priors, and their performance is likely poor and was not measured directly.

7.2. Conclusions

This paper presents the system our team submitted to BirdCLEF+ 2026, an ensemble of solutions built on the pretrained Perch v2 foundation model. Rather than training spectrogram classifiers from scratch, the system encodes each one-minute soundscape into a sequence of Perch embeddings and logits, models that sequence with a bidirectional selective state-space network (ProtoSSM) refined by activity priors, MLP probes and a residual SSM correction, and fuses it by class-wise percentile rank with a distilled five-fold sound-event detector, followed by taxonomy- and time-aware post-processing. Everything runs CPU-only through ONNX Runtime within the competition budget. The final submission reached a public score of 0.955 and a private score of 0.950, placing 39th of 4,094 teams for a silver medal. The decisive ingredient was the Perch v2 embeddings, on top of which the sequence model, the complementary SED branch and the rank fusion were each retained because they improved out-of-fold stability or public-leaderboard performance (Section 5.4); carefully combining publicly released components under the CPU constraint proved an effective route to a competitive result. The most informative next experiment is a paired leaderboard test of an SSM-free pipeline against the full system (Section 5.4), to confirm on the hidden test what our out-of-fold analysis suggests — that the foundation-model features and the cheap priors, rather than the sequence models, carry most of the value. Beyond that, future work may focus on better coverage of the species Perch does not represent (the 28 no-signal classes), on training the sequence model on a larger pool of labelled soundscapes, and on a principled, leakage-free tuning of the fusion and post-processing in place of the inherited settings used here.

Acknowledgements

The author thanks the members of the team for discussions and for granting permission to document the team's BirdCLEF+ 2026 submission in this working note. The system stands on publicly released work, which we gratefully acknowledge: Google's Perch v2 bioacoustic foundation model [9]; the public BirdCLEF+ 2026 rank-fusion notebook [18] and the Perch-embedding sequence-model reproduction it builds on [19]; the distilled sound-event detector shared as open data [20]; and the public OpenVINO starter [21] used only as a reference baseline. We also thank the wider BirdCLEF community for the discussions and notebooks consulted during development.

Declaration on Generative AI

During the preparation of this work, the author used OpenAI's ChatGPT and Anthropic's Claude for grammar and spelling checking, language and wording refinement, LaTeX formatting, and assistance with scaffolding analysis and plotting code. These tools were not used to design the modelling system, run the experiments, or generate the reported results. After using these tools, the author reviewed and edited all content and takes full responsibility for the publication's content.

References

[1] 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.

[2] 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. de Almeida Júnior, R. C. Kridler, M. Lasseck, A. V. de Melo, H. Müller, M. de O. Neves, M. G. dos Reis, L. K. Sampaio Teles, K.-L. Schuchmann, J. L. M. M. Sugai, K. T. P. Azevedo, P. do 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*, CEUR-WS.org, 2026.

[3] S. Kahl, T. Denton, H. Klinck, H. Reers, F. Cherutich, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, "Overview of BirdCLEF 2023: Automated bird species identification in Eastern Africa," in: *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 1934–1942.

[4] S. Kahl, T. Denton, H. Klinck, V. Ramesh, V. Joshi, M. Srivathsa, A. Anand, C. Arvind, H. Cp, S. Sawant, V. V. Robin, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, "Overview of BirdCLEF 2024: Acoustic identification of under-studied bird species in the Western Ghats," in: *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1948–1957.

[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. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, "Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the Middle Magdalena, Colombia," in: *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 2909–2919.

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

[7] M. Tan, Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in: *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, 2019, pp. 6105–6114.

[8] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, "A ConvNet for the 2020s," in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11966–11976.

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

[10] M. Lasseck, "Bird species identification in soundscapes," in: *Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum*, volume 2380 of *CEUR Workshop Proceedings*, 2019.

[11] M. Lasseck, "Improving bird recognition using pseudo-labeled recordings from the target location," in: *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2110–2118.

[12] G. Hinton, O. Vinyals, J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531* (2015).

[13] L. Hong, "Domain adaptation for birdcall recognition: Progressive knowledge distillation with semi-supervised and self-supervised soundscape labeling," in: *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2103–2109.

[14] 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*, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 3208–3228.

[15] L. Hong, "Acoustic bird species recognition at BirdCLEF 2023: Training strategies for convolutional neural network and inference acceleration using OpenVINO," in: *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 2043–2050.
[16] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in: *Proceedings of UAI*, 2018.
[17] A. Gu, T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752* (2023).
[18] Pilkwang Kim, "BirdCLEF+ 26: Acoustic time-window rank fusion," Kaggle notebook, 2026. https://www.kaggle.com/code/pilkwang/birdclef-26-acoustic-time-window-rank-fusion. Accessed 2026-06-04.
[19] yukiZ, "Bird26: Reproduce Perch + ProtoSSM + ResidualSSM (inference / train)," Kaggle notebook, 2026. https://www.kaggle.com/code/hideyukizushi/bird26-reproduce-perch-protossm-resssm-inf-train. Accessed 2026-06-04.
[20] T. Arrants, "BC2026 distilled-SED," Kaggle dataset and notebook, 2026. https://www.kaggle.com/code/tuckerarrants/bc2026-distilled-sed. Accessed 2026-06-04.
[21] R. Ramakrishnan, "BirdClef2026 | OpenVINO | Starter," Kaggle notebook, 2026. https://www.kaggle.com/code/ravi20076/birdclef2026-openvino-starter-v1. Accessed 2026-06-04.
[22] Jae John, "Perch-meta: cached Perch v2 embeddings for the BirdCLEF+ 2026 train soundscapes," Kaggle dataset, 2026. https://www.kaggle.com/datasets/jaejohn/perch-meta. Accessed 2026-06-04.

A. Implementation Details

For reproducibility we record the remaining implementation details. The Perch v2 backbone and the five distilled-SED folds are executed with ONNX Runtime on CPU (four intra-op threads); the sequence models (ProtoSSM and ResidualSSM) are small PyTorch modules trained on the fly from the cached training-soundscape embeddings, and the per-class probes are scikit-learn multilayer perceptrons over a 64-dimensional PCA of the Perch embeddings. The random seed is fixed (seed 4).

Front ends. The two branches use different front ends. The Perch branch consumes raw 5-second 32 kHz waveforms and emits a 1536-dimensional embedding and Perch logits per window. The distilled-SED branch consumes a log-mel spectrogram of each window; Table 10 lists its parameters.

Table 10
Log-mel spectrogram parameters of the distilled-SED branch. The Perch branch uses no hand-designed spectrogram (raw-waveform input).

Parameter	Distilled-SED branch
Sample rate (Hz)	32,000
FFT size	2048
Hop length	512
Mel bins	256
Frequency range (Hz)	20–16,000
Amplitude	dB (top 80), standardised per instance

Inference procedure. At test time each 60-second soundscape is read, divided into twelve consecutive 5-second windows, and embedded once by Perch v2. The Perch-embedding branch produces p_{proto} by running the ProtoSSM (with shift/flip test-time augmentation), adding the activity priors and MLP probes, applying the per-class first-pass blend (Eq. 2) and the ResidualSSM correction, and finishing

Table 11
Public components the system builds on (Kaggle datasets / notebooks and the Perch v2 model release).

Component	Source	Role
Perch v2 backbone	google/bird-vocalization-classifier [9]	frozen feature extractor
ProtoSSM / ResidualSSM	reproduction notebook [19]	sequence model
Distilled SED (5 folds)	`bc2026-distilled-sed` [20]	SED branch
Perch embedding cache	`perch-meta` [22]	precomputed features
Rank-fusion notebook	Pilkwang Kim [18]	fusion + post-processing
OpenVINO starter	ravi20076 [21]	reference baseline only

with the calibration and confidence/rank scalings of Section 4.4. The SED branch produces p_{sed} as the smoothed mean of the five folds. The two branches are fused by class-wise percentile rank (Eq. 1) and passed through the post-processing gates of Section 4.6; the result is clipped to $[0, 1]$ and written in the per-window `row_id` format expected by the competition. On the public CPU dry-run the whole procedure takes a few minutes (Section 5.5).

Public components used. For reproducibility, Table 11 lists the public artefacts the pipeline builds on; the final submission is a fork of the rank-fusion notebook with these inputs attached.

B. Additional Data Analysis

This appendix collects further quantitative views: two of the focal-to-soundscape domain shift discussed in Section 3.3 — Figure 5 (geographic origin of the focal recordings against the Pantanal test region) and Figure 6 (how many species are annotated simultaneously per chunk) — and one of the public–private leaderboard behaviour, Figure 7 (Section 6.2).

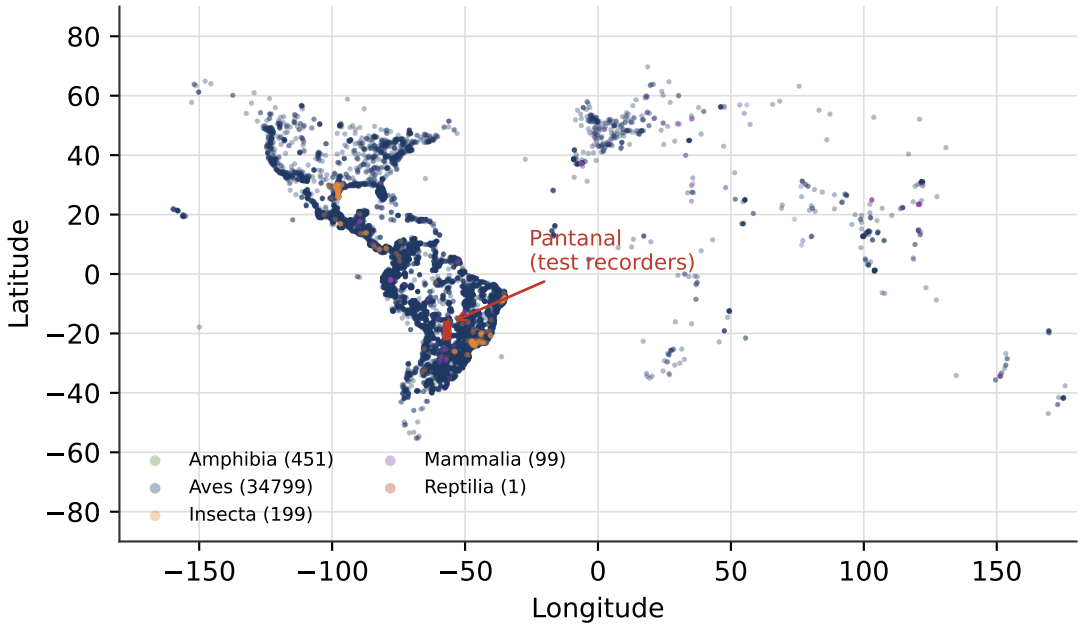

Figure 5: Geographic domain shift. Origin of the 35,549 focal training recordings (coloured by taxonomic class, with counts in the legend); the red rectangle marks the Pantanal test-recorder deployment box (16.5°–21.6°S, 55.9°–57.6°W). The training data span the globe, yet only 2.4% of recordings originate inside the target region.

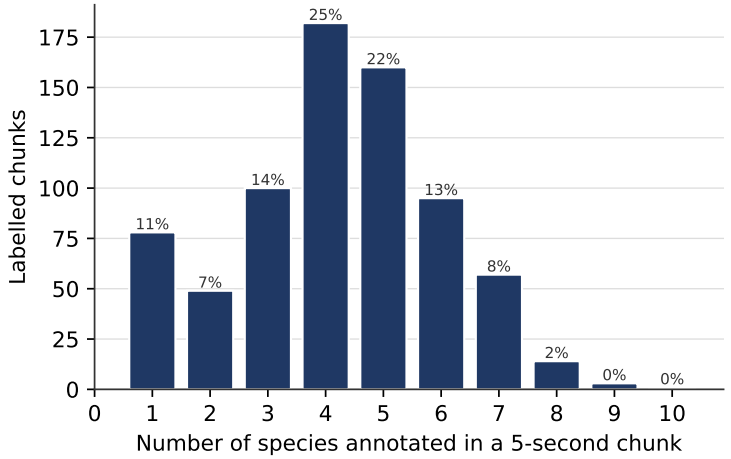

Figure 6: Multi-label density of the labelled training soundscapes: number of species annotated within each 5-second chunk (739 labelled chunks; percentages above bars). Most chunks contain several simultaneous species (mean 4.2, max 10), in stark contrast to the single-species focal clips.

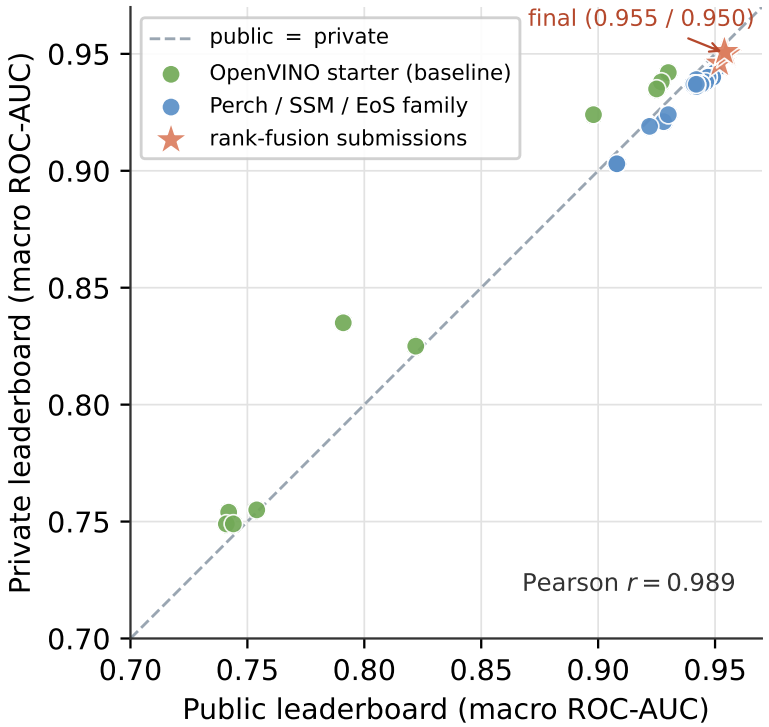

Figure 7: Public versus private leaderboard for all 45 scored team submissions (macro ROC-AUC). Points lie close to the diagonal (Pearson $r = 0.989$), so the public leaderboard was a reliable proxy and we observed no shake-up; the selected rank-fusion submission (star) sits at the top of the cluster. The OpenVINO-starter baselines span the lower range and the Perch / SSM / EoS family the upper range (Section 6.2).