Can Generative AI Improve Bioacoustic Classification?

A Study on BirdCLEF-2026 with BirdNET Embeddings, Fine-Tuning, and AudioLDM 2 Synthetic Audio

Ademar Tutor

COMPX525 Deep Learning — Assignment 2

Due: 19 June 2026

Abstract

This report investigates whether generative artificial intelligence can improve classification in a long-tailed, multi-taxa bioacoustic task using passive acoustic monitoring recordings from the Pantanal wetlands provided through the Kaggle BirdCLEF 2026 competition. The dataset contains 234 target classes across five taxa and is highly imbalanced, with several classes represented by very few recordings. Following a two-by-two experimental design, we compare a frozen pretrained audio representation with an end-to-end fine-tuned neural network, each trained using either the original dataset or a dataset augmented with synthetic audio. The synthetic recordings are generated using AudioLDM 2 and filtered with BirdNET to retain clips that can be recognised as the intended class. Synthetic augmentation produces modest improvements for both approaches, with the strongest performance achieved by the fine-tuned model trained on the augmented dataset. However, only a small proportion of the generated recordings pass the verification step, and the success rate varies considerably across taxa. These findings suggest that generative AI can improve bioacoustic classification, but its usefulness depends on how reliably the vocalisations of each taxon can be generated and verified.

1 Introduction

This report responds to the COMPX525 Deep Learning assignment, which is built on the Kaggle *birdclef-2026* competition. The task is multi-taxa species identification from passive acoustic monitoring (PAM) recordings, and the report's contributions are set out below. The wider motivation is biodiversity monitoring: continuous PAM recordings can survey remote ecosystems at scale, but that same scale makes manual annotation infeasible, so automatic species identification is essential. The Pantanal wetland that birdclef-2026 targets is large, seasonally flooded, and under-monitored (Wikipedia contributors, 2026), and its recordings cover not only birds but also amphibians, mammals, insects, and a reptile (Figure 1).

The task presents two central difficulties. First, most species have far too few recordings to learn from: the dataset is dominated by a handful of common species, while many rare species (the long *tail* of the class distribution) have only a few examples each. Second, the model trains on clean recordings of a single animal (*focal* clips) but must perform on messy real-world recordings in which many sounds overlap (*soundscapes*). This mismatch between training and deployment conditions is known as a *domain shift*. Both problems are extreme for this dataset, and Section 4 gives the exact figures. The rare species matter most of all. A method that does well on common species but ignores the rare ones misses the point of monitoring, because those under-studied species are exactly the ones conservation needs to find.

The assignment frames a specific question, in its own words: *"Can GenAI help getting better classification results?"* This study sharpens that into a concrete question: can generative AI produce synthetic audio that improves classification of rare, under-represented species? To answer

Figure 1. An aerial view of the Pantanal wetlands, the seasonally flooded, multi-taxa habitat in which the birdclef-2026 soundscapes are recorded. The patchwork of open water and vegetation is the kind of acoustically crowded environment that the focal-to-soundscape domain shift (Section 4.5) must bridge. Photograph by Alicia Yo, CC BY-SA 3.0, via Wikimedia Commons.

it, the assignment prescribes a two-by-two experimental matrix. One axis varies the model (frozen BirdNET embeddings with a simple classifier, or a fine-tuned network); the other varies the data (original, or enriched with AudioLDM 2-generated synthetic audio that is filtered with a BirdNET check). This design isolates the effect of the model and the effect of the synthetic data, and shows how the two interact.

This report makes the following contributions:

- A complete, brief-compliant two-by-two evaluation on birdclef-2026, with an 80/20 split in which the four single-example classes are forced entirely into validation.
- A genuine implementation of the synthetic-audio track using AudioLDM 2 with a BirdNET verification filter, rather than a substitute, and a report of its keep-rate by taxon.
- Evidence that filtered generated audio improves both the frozen-embedding and the fine-tuned model, with the largest gain for the fine-tuned model.
- An analysis of *where* the benefit accrues, showing that generative audio helps selectively by taxon and cannot rescue the single-example classes.

2 Study Design

The study comprises four experimental conditions, summarised in the experimental matrix below to clarify their relationships before each condition is described in detail. The design crosses two axes: the *model family* (frozen BirdNET embeddings with a classical classifier, versus an end-to-end fine-tuned network) and the *data condition* (original data, versus original data enriched with BirdNET-filtered AudioLDM 2 synthetic audio). Crossing these axes yields the four conditions summarised in Table 1, and isolates three quantities: the effect of the model, the effect of the synthetic data, and their interaction. The per-condition methods are described in Sections 5.2–5.5, and the corresponding focal-validation results are reported in Table 6.

The fine-tuned conditions train an EfficientNet-b0 from an ImageNet-pretrained initialisation rather than fine-tuning BirdNET, because BirdNET's public release does not provide a supported trainable backbone checkpoint or end-to-end fine-tuning interface; this choice, and the decision to initialise from pretrained rather than random weights, is justified in detail in Section 5.3. Reading the matrix down a column isolates the effect of the synthetic data while holding the model fixed,

Table 1. The two-by-two study design. Each row pairs a data condition with a model family. The two B conditions differ from their A counterparts only in adding BirdNET-filtered AudioLDM 2 audio to the training pool; the validation set, including the four single-example classes, is held fixed across all four conditions, so any change is attributable to the training data or the model rather than to what is measured.

Condition	Training data	Classification approach
A1	Original data only	Frozen BirdNET embeddings (1024-d) + per-class logistic-regression classifier
A2	Original data only	End-to-end fine-tuned EfficientNet-b0 (ImageNet-pretrained) on log-mel spectrograms
B1	Original + AudioLDM 2 synthetic audio (BirdNET-filtered)	Frozen BirdNET embeddings (1024-d) + per-class logistic-regression classifier
B2	Original + AudioLDM 2 synthetic audio (BirdNET-filtered)	End-to-end fine-tuned EfficientNet-b0 (ImageNet-pretrained) on log-mel spectrograms

and reading across a row isolates the effect of the model while holding the data fixed; the Results section exploits exactly this structure to attribute the observed changes.

3 Related Work

Prior work relevant to this study falls into five lines: bioacoustic embeddings and transfer learning, the BirdCLEF task lineage and its domain shift, the behaviour of fine-tuning relative to frozen features under shift, synthetic data and augmentation for rare classes, and the verification and evaluation of synthetic data under extreme imbalance. Three results from this work matter most here: synthetic audio helps as a supplement, not a replacement; general generators produce less faithful bird calls than bird-specific ones; and fine-tuning can degrade good frozen features. We revisit each in Sections 6 and 7.

3.1 Bioacoustic embeddings and transfer learning

A first line of work pairs pretrained, multi-taxa audio embeddings with a simple downstream classifier. BirdNET is a robust avian classifier trained on very large collections of recordings (Kahl et al., 2021) (Appendix A), and Ghani et al. (2023) show that its frozen embeddings, with only a lightweight classifier on top, give strong few-shot transfer across taxa. Perch 2.0 extends this paradigm to multiple taxa and reports that frozen embeddings with a linear probe are highly competitive (van Merriënboer et al., 2025). These approaches are assessed on standardised benchmarks: BEANS (Hagiwara et al., 2023) covers birds, mammals, and insects, whereas BirdSet (Rauch et al., 2025) focuses on avian multi-label classification under the focal-to-soundscape domain shift. This approach is useful because it is data-efficient and robust on small classes, which motivates our A1/B1 cells. However, it leaves open how to help a class that has almost no examples at all, which is the gap the synthetic-data track targets.

3.2 The BirdCLEF lineage and the focal-to-soundscape domain shift

The BirdCLEF competition lineage both defines the standard pipeline and characterises the central difficulty. The BirdCLEF 2024 overview frames the focal-to-soundscape shift explicitly and reports that strong solutions combine pseudo-labelling, test-time augmentation, and diverse ensembles (Kahl et al., 2024). The BirdCLEF+ 2025 overview is the closest published precedent to birdclef-2026: it shares the multi-taxa scope (birds, amphibians, mammals, and insects) and the same focal-to-soundscape shift in a Neotropical soundscape setting (Cañas et al., 2025). We follow

this line of work in two concrete ways: macro-averaged ROC-AUC as the primary metric, and a fine-tuned spectrogram CNN (EfficientNet-b0) as the model for the A2 and B2 cells. The long tail and the domain shift remain the main obstacles, and our exploratory analysis measures both for the Pantanal data.

3.3 Fine-tuning versus frozen features under distribution shift

Pretrained features transfer well and beat from-scratch training on small data (Yosinski et al., 2014), and better pretrained models transfer better (Kornblith et al., 2019), which supports building on a strong ImageNet backbone such as EfficientNet (Tan & Le, 2019). However, Kumar et al. (2022) show that fine-tuning under distribution shift can *distort* good pretrained features and underperform a frozen linear probe. This tension is directly relevant to our model axis: it predicts why our closed-form adapter on the frozen BirdNET embeddings (the A2 ablation) falls below the A1 linear baseline under the focal-to-soundscape shift, and why our headline A2 and B2 cells obtain a benefit only by fine-tuning the EfficientNet backbone end-to-end with strong augmentation (mixup and SpecAugment) rather than through minimal adaptation (Section 7).

3.4 Synthetic data and augmentation for rare classes

When a class has too few examples to learn from, additional training examples can be manufactured, and these methods form a progression from modifying existing data to generating new data. The simplest approach, SMOTE (Synthetic Minority Over-sampling Technique), creates synthetic minority-class examples by interpolating between a real example and its nearest neighbours in feature space, adding variety without any new recordings (Chawla et al., 2002). A second approach perturbs the spectrograms the model already has: mixup blends pairs of examples and their labels (Zhang et al., 2017), and SpecAugment masks and warps regions of the log-mel spectrogram (Park et al., 2019). The most expansive approach generates entirely new audio: AudioLDM 2, a diffusion-based text-to-audio model, synthesises novel clips from a text prompt (Liu et al., 2024) and is the generator the brief specifies. Each step adds more capacity to manufacture training signal where data is scarce, but also more freedom to drift from the real data, which motivates the verification step we examine in Section 6.

The empirical evidence on generative augmentation is consistent but qualified, and it directly shapes the expectations for this study. Across small-scale audio classification, aligning a text-to-audio model to the target data and filtering its output yields consistent gains (Ghosh et al., 2025), and for bioacoustic event detection in particular, synthetic data can improve few-shot performance (Hoffman et al., 2025). The recurring caution, however, is that synthetic audio helps as a *supplement* rather than a replacement: training on synthetic data alone tends to degrade performance on real test data (Ronchini et al., 2024). A further qualification concerns fidelity at the species level. A general text-to-audio model is not specialised for bird calls, and a bird-specific generator can reach substantially higher class fidelity than a general one (Song & Ta, 2025), which suggests that off-the-shelf generation may be most useful when its output is verified before use.

What remains unclear is whether a general-purpose audio generator can produce calls that are accurate at the species level. This is the question the present study tests, by using a BirdNET check to verify each generated clip before it is used.

3.5 Rare classes, synthetic filtering, and evaluation under imbalance

The final consideration is the extreme-scarcity regime that this study's single-example classes occupy, and how to evaluate synthetic data fairly within it. When a class has only one real recording and that recording is held out for validation, the model never sees a real example of the class during training; the synthetic-data condition is therefore closer to synthetic-to-real transfer than to conventional supervised learning. Useful results are still achievable in this regime: a few-shot study detected a critically endangered species from only three recordings using cosine similarity over

frozen bird-classifier embeddings (Jana et al., 2025). This shows that extreme-low-data transfer is possible, but a single validation example per class supports only a case-study reading rather than a reliable estimate, which is how this report treats its single-example classes.

The BirdNET filter the brief suggests also warrants caution. BirdNET confidence scores are species-specific detector outputs rather than calibrated probabilities (Wood & Kahl, 2024), so using BirdNET as the sole gate improves fidelity but also introduces bias: it tends to admit clips that match what BirdNET has already learned a species should sound like, and to reject atypical or unrecognised ones, rather than acting as an objective oracle. This matters here because the filter is most informative precisely where BirdNET's verification is strongest, namely for the bird species it covers well.

4 Data and Exploratory Analysis

This section describes the dataset and the properties that shape the experimental design, in particular the severe class imbalance, the taxonomic asymmetry, and the focal-to-soundscape domain shift. Understanding these properties matters because they explain both why a synthetic-data track is worth attempting and why its benefit is concentrated in specific places. The subsections that follow document each property in turn and connect it to the line of related work it bears on.

4.1 Dataset overview

The dataset combines two kinds of audio. The focal recordings are clean, mostly single-species clips contributed by citizen-science platforms, and the soundscapes are continuous passive-acoustic-monitoring (PAM) recordings from the Pantanal that match the deployment setting. The training data comprises 35,549 focal recordings, all provided at a 32 kHz sampling rate. The target label set has 234 classes, but only 206 of them appear among the focal recordings; the remaining 28 classes have *no* focal recording at all and are represented only in the soundscapes. Table 2 summarises the sources and their characteristics. The focal recordings are drawn from two collections that differ in quality metadata: Xeno-Canto provides quality ratings, whereas the iNaturalist recordings are largely unrated (about 36% of all recordings carry no rating). Recording durations vary widely, with a median of 20.8 seconds and a range from a fraction of a second to over 40 minutes, which is why all audio is processed in fixed 5-second windows (Section 5.1).

Table 2. Sources of the birdclef-2026 training data. The focal recordings supply most of the labelled examples but are dominated by Xeno-Canto and span only 206 of the 234 target classes; the soundscapes match the evaluation setting but are sparsely labelled. The split between the collections matters because the two differ in quality metadata.

Source / property	Value
Focal recordings (total)	35,549
from Xeno-Canto	23,043 (64.8%)
from iNaturalist	12,506 (35.2%)
Target classes	234
with at least one focal recording	206
with no focal recording	28
Labelled soundscape segments	1,478 (across 66 files)
Unlabelled soundscape files	10,592
Sampling rate	32 kHz
Median focal duration	20.8 s

4.2 A multi-taxa target

The task is multi-taxa, which is the first property the experimental design must account for. Figure 2 shows the composition of the 234 classes across the five taxa. Birds (Aves) account for 162 classes, with the remaining classes split between amphibians (32), insects (28), mammals (8), and a single reptile. This composition matters because BirdNET, the embedding model used in A1 and B1, is trained predominantly on birds. As Ghani et al. (2023) show, frozen bird embeddings transfer few-shot across taxa (Section 3.1); the non-Aves classes here are precisely the off-distribution case where that transfer is least assured.

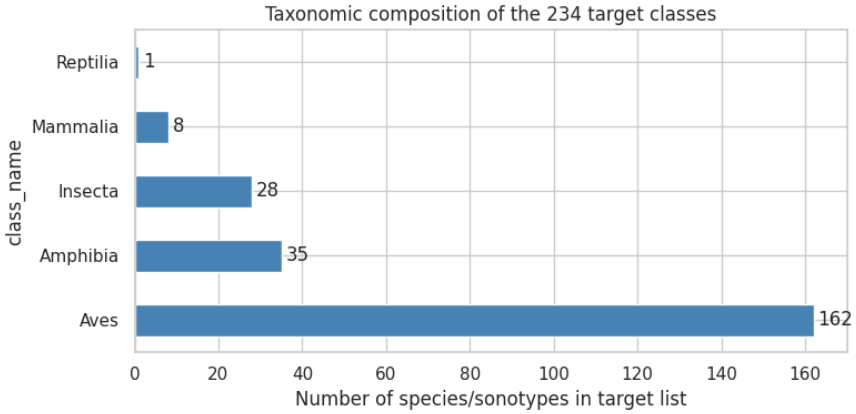

Figure 2. Taxonomic composition of the 234 target classes. Birds dominate the label set, while amphibians, insects, mammals, and a single reptile make up the remainder. Because the embedding model used in A1/B1 is bird-trained, the non-Aves classes are where transfer is least assured.

4.3 The long tail

The class distribution is severely long-tailed, which is the property that motivates the synthetic-data track; this is the class-imbalance regime that minority-oversampling methods such as SMOTE were designed to address (Chawla et al., 2002). The ratio between the most and least populous class is approximately 499×, the median class has 125 recordings, and four classes have only a single recording. Following the brief, these four single-example classes are placed entirely in the validation set, so in the original-data condition the model receives no training signal for them at all. Figure 3 plots the rank-frequency curve and the distribution of class sizes, and Table 3 counts how many classes fall into successive rarity bands. A substantial fraction of classes are data-scarce: 18 classes have at most five recordings and 27 have at most ten, the regime in which synthetic data is most likely to help (Section 3.4). The ≤ 5 and ≤ 10 bands are the same bands used in the rare-tail analysis of the results (Section 6); the counts here are over all classes, whereas the results table counts focal-training support, so band membership differs slightly.

4.4 Scarcity is taxon-dependent

The scarcity is not spread evenly across taxa, which is what makes the long tail and the multi-taxa structure interact. Figure 4 and Table 4 show that birds are comparatively data-rich (median 168 recordings per class), whereas amphibians, mammals, and the reptile are severely under-resourced (median around seven recordings or fewer). The taxa that are scarcest are therefore also the taxa furthest from BirdNET's training distribution. This pairing matters for two parts of the study: it is where a frozen bird-trained embedding is least assured (Section 3.1), and it is where the verification filter is least able to check generated clips, because the filter relies on BirdNET scores whose meaning Wood and Kahl (2024) show is species-specific (Section 3.5).

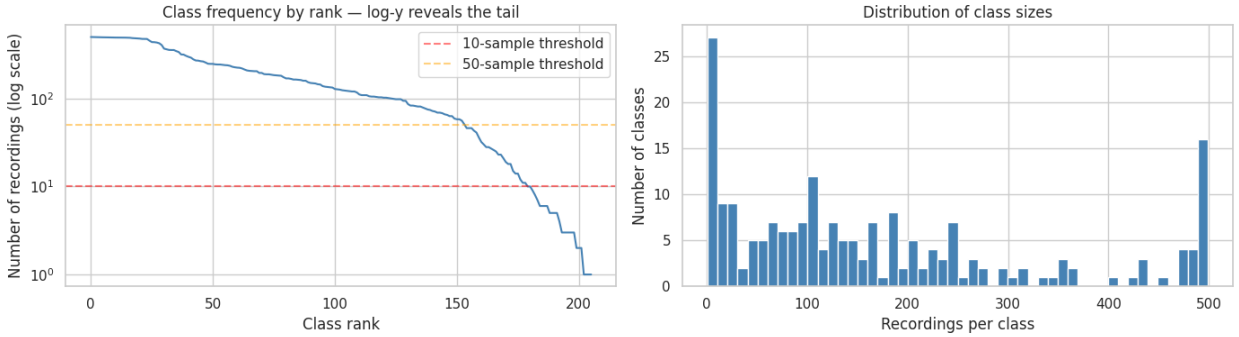

Figure 3. The long tail of birdclef-2026. Left: number of recordings per class against class rank, on a log scale, showing a rapid drop below the 50- and 10-recording thresholds. Right: the distribution of class sizes, with a heavy concentration of small classes. Together they show that many classes are too scarce to learn well from real data alone, motivating targeted synthetic augmentation of the tail.

Table 3. Number of training classes within successive rarity bands. The ≤ 5 and ≤ 10 bands correspond to the rare-tail bands analysed in Section 6; the four single-example classes are forced entirely into validation. Counts are over all classes with focal recordings.

Rarity band (recordings)	Classes	% of classes with focal data
$= 1$	4	1.9%
≤ 5	18	8.7%
≤ 10	27	13.1%
≤ 25	40	19.4%
≤ 50	52	25.2%
≤ 100	83	40.3%

4.5 The focal-to-soundscape domain shift

The training audio and the evaluation audio are not the same kind of audio, and they differ in two ways that compound each other. This mismatch is the central difficulty that the BirdCLEF overviews describe (Cañas et al., 2025; Kahl et al., 2024) (Section 3.2). The first difference is *where* the recordings come from. The focal recordings are spread across the world, but the test region is the Pantanal alone. Every focal recording is geotagged, yet only 727 of them (2.0%) fall inside the Pantanal, as Figure 5 shows; the models therefore learn mostly from birds recorded elsewhere and are judged on birds recorded there. The second difference is *what the audio sounds like*. The test recordings are passive soundscapes left running in the field, so they are crowded: many calls overlap and the background is noisy. The focal clips are the opposite, clean and usually one species at a time. Figure 6 places a focal recording next to a soundscape segment so this contrast is visible. A third problem sharpens the first two: some test species are barely present in the training clips at all. Of the 75 species labelled in the soundscapes, only 47 also have focal recordings; the remaining 28 have none, so in the original-data condition there are simply no examples to learn them from.

4.6 Acoustic diversity across taxa

Finally, the taxa differ markedly in the structure of their vocalisations, which is why a bird-trained model and a bird-trained verification filter behave differently across them. Representative log-mel spectrograms for each taxon, provided in Appendix C (Figure 13), show that their time-frequency structure differs markedly from one group to the next. This diversity is the data-side reason behind the taxon-dependent behaviour reported later: a generator and a filter built around bird vocalisations are best matched to the bird classes and least matched to the others, which anticipates the taxon-selective keep-rate examined in Section 6.

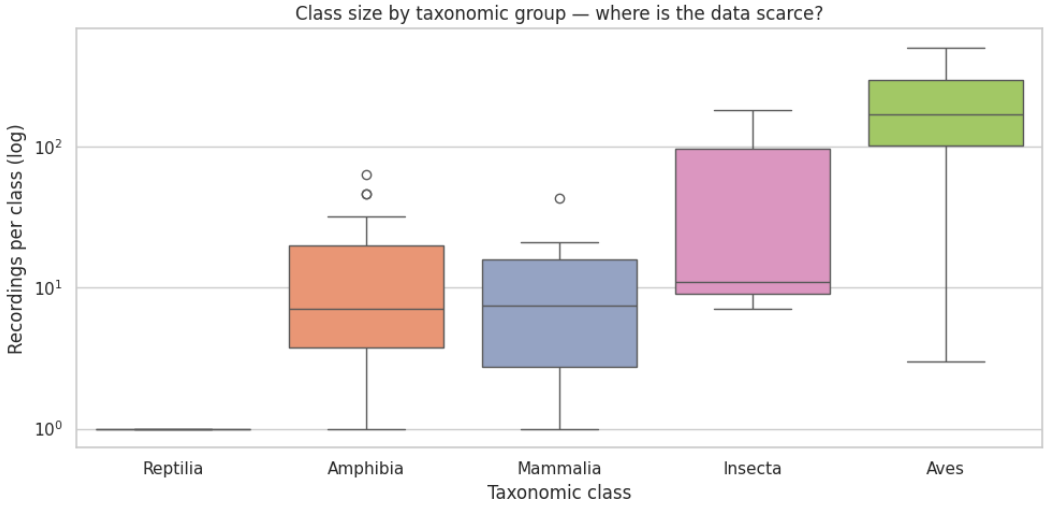

Figure 4. Recordings per class by taxon, on a log scale. Birds have by far the most recordings per class, while reptiles, amphibians, and mammals sit near the bottom. The scarcest taxa are also the ones least represented in BirdNET's training data, which compounds their difficulty.

Table 4. Per-taxon class counts and recordings-per-class summary. Birds dominate both in the number of classes and in recordings per class; the non-Aves taxa are far scarcer, with median support at or below about seven recordings per class.

Taxon	Classes	Median	Mean	Max
Aves	162	168.0	214.9	499
Amphibia	32	7.0	14.9	63
Insecta	28	31.0	31.8	71
Mammalia	8	7.5	12.4	43
Reptilia	1	1.0	1.0	1

5 Methods

This section describes the preprocessing, the train/validation split, and the four experimental cells. The methods follow the brief's two-by-two structure so that the effect of the model and the effect of synthetic data can be read off separately. We first fix the shared protocol, then describe each cell and the synthetic-audio generation pipeline, and reference the architecture diagrams that accompany them.

5.1 Preprocessing and the brief-compliant split

All four cells share a common raw-audio front-end, after which the two model families diverge in how a window is represented; the train/validation split is then shared again. Each recording is cut into 5-second windows, matching the segment length used at evaluation time. Focal recordings are read from the start of the file, whereas a labelled soundscape segment is read at its annotated offset. Multi-channel audio is averaged to a single channel, the window is resampled to a rate that depends on the branch, and windows shorter than five seconds are zero-padded while longer reads are truncated to the window length.

The embedding cells (A1 and B1) turn each window into a frozen BirdNET feature vector rather than an image. Because BirdNET expects audio at its native 48 kHz with a 3-second analysis window, each 5-second clip is resampled to 48 kHz and covered by two overlapping 3-second windows, aligned to the start ($[0, 3]$ s) and to the end ($[2, 5]$ s) of the clip. Each is embedded separately and the two 1024-dimensional vectors are mean-pooled into one, which is then L_2-normalised.

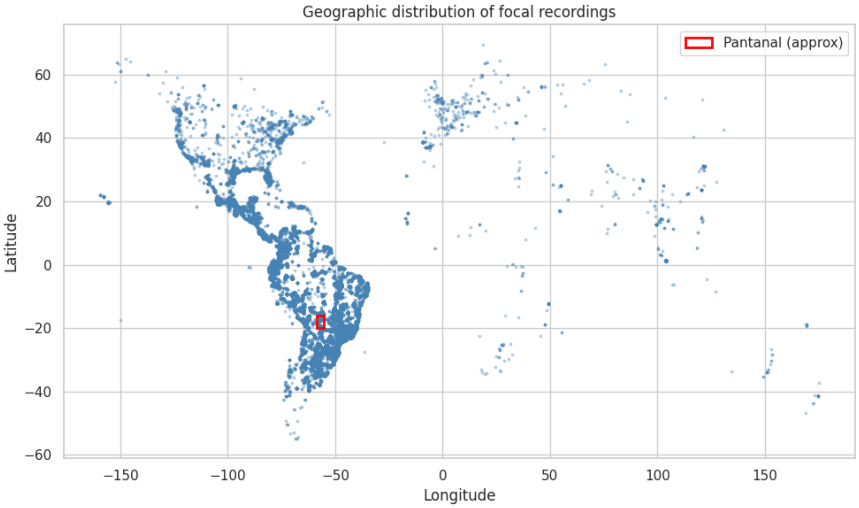

Figure 5. Geographic distribution of the focal recordings, with the approximate Pantanal bounding box overlaid. Although the recordings are geocoded worldwide, only about 2% fall inside the Pantanal, illustrating the geographic component of the focal-to-soundscape shift.

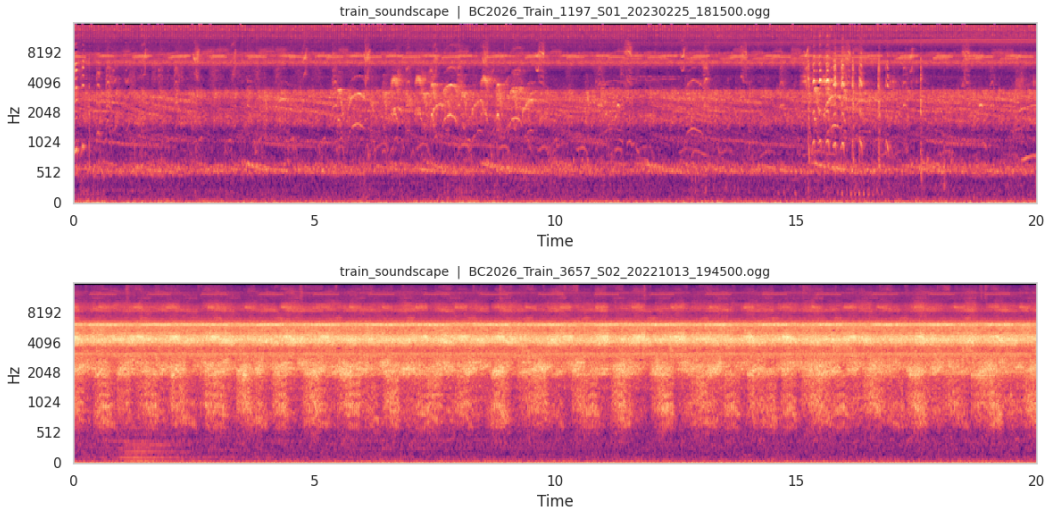

Figure 6. A Pantanal soundscape (PAM) segment as a log-mel spectrogram. Unlike a clean focal recording, it contains dense, overlapping signals from multiple sources and substantial background noise, which is the acoustic component of the domain shift that the models must bridge at evaluation time.

A per-class logistic regression (regularisation C=1.0, class-balanced weighting) is trained on the cached embeddings. In B1, the BirdNET-filtered AudioLDM 2 clips are embedded in the same way and added to the training pool.

The spectrogram cells (A2 and B2) turn each window into an image for the convolutional backbone. The window is resampled to 32 kHz and converted to a log-mel spectrogram (1024-point FFT, hop length 512, 128 mel bands spanning 50 to 16,000 Hz) expressed in decibels relative to its maximum, standardised to zero mean and unit variance, resized to 224×224, and repeated across three channels so that an ImageNet-pretrained backbone can consume it. During training the spectrograms are augmented with SpecAugment (two frequency and two time masks) and mixup (α=0.15); in B2, the generated rare-class renders additionally receive pitch- and time-shifting on the waveform before the mel transform. Figure 7 shows each of these augmentations applied to the same focal recording, which illustrates how they perturb the time-frequency structure and so widen the effective training distribution, particularly for the rare classes that have few real examples. The multi-label binary-cross-entropy loss uses a per-class positive weight capped at 50 to counter imbalance, and labelled soundscape segments are weighted 10× relative to focal windows. In B2,

the AudioLDM 2 clips are added to the training pool, and the cell falls back to augmentation alone if no clips were generated.

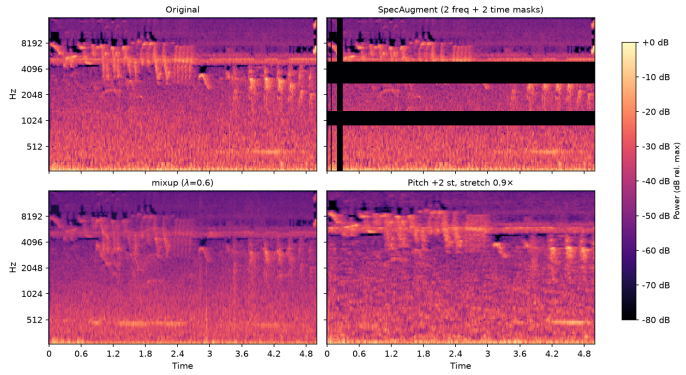

Figure 7. The three spectrogram-domain augmentations used by the convolutional cells, each applied to the same focal recording: SpecAugment (frequency and time masks), mixup (a blend with a second clip), and the B2 waveform pitch- and time-shift. A full-size version with the exact parameters is provided in Appendix B (Figure 12).

The 80/20 train/validation split is applied to the focal recordings (`train.csv`). These are split into training and validation, stratified by class, with one exception required by the brief: the four single-example classes are placed *entirely* in validation. As a consequence, these four classes have no training signal in the original-data condition, which is the brief's deliberate test of whether synthetic data can help where no real example is available for training. The labelled soundscape segments are not part of this split; they are added in full to the training pool as in-domain passive-monitoring signal, and the unlabelled soundscape files (Table 2) are not used for supervised training or validation. Because the labelled soundscapes are therefore inside the training set, performance is reported only on the held-out focal-validation split (Section 5.6).

5.2 A1: frozen BirdNET embeddings with a linear classifier

The first cell, A1, is the frozen-embedding baseline. Each 5-second window is embedded by BirdNET, and a per-class logistic-regression classifier is trained on the cached 1024-dimensional embeddings. The backbone is not updated. Figure 8 illustrates this pipeline. This design is motivated by the evidence in Section 3.1 that frozen bioacoustic embeddings with a simple classifier give strong, data-efficient transfer, and it serves as the reference point against which the other cells are measured.

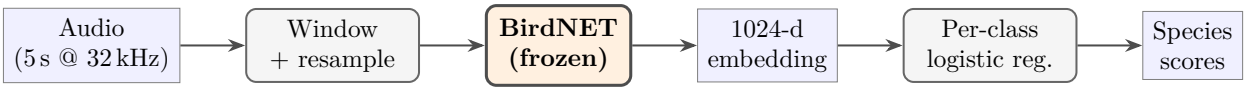

Figure 8. The A1 pipeline. Audio windows are embedded by a frozen BirdNET model, and a per-class logistic-regression head is trained on the cached embeddings. The backbone is never updated, so the only learned component is the linear head. B1 reuses this pipeline, adding generated audio to the training pool before embedding.

The architecture deliberately separates a fixed feature extractor from a small trainable head. The BirdNET backbone is *frozen* and reused only to map audio to a 1024-dimensional embedding; the only trainable components are the per-class linear heads. This choice follows directly from the problem and the data. A frozen bioacoustic embedding with a light linear probe gives data-efficient transfer on the long tail (Section 3.1), which matters because most classes here have few examples. It also avoids the risk, identified by Kumar et al. (2022), that fine-tuning a pretrained backbone under distribution shift distorts otherwise-useful features, a risk that is acute under the focal-to-soundscape shift. Finally, keeping the learned capacity small (about 0.23 million parameters in total, against a backbone that is never updated) limits overfitting on the rare classes that dominate

the label set. Table 12 in Appendix F summarises the resulting configuration stage by stage, including the frozen/trainable split and parameter counts.

A linear head is trained for each of the 224 of 234 classes that have at least three positive examples in the training pool; the remaining classes are too sparse to fit a head and default to a neutral score. B1 reuses this architecture unchanged, embedding the BirdNET-filtered generated clips in the same way and adding them to the training pool before the heads are fit.

5.3 A2: end-to-end fine-tuned EfficientNet

The second cell, A2, instantiates the brief's first model option, namely an *architecture of our choice* trained on this task, rather than the alternative of fine-tuning BirdNET. We choose this option because BirdNET's public release provides inference-oriented model artifacts and supports training custom classifiers on frozen embeddings, but it does not provide a supported trainable backbone checkpoint or an end-to-end fine-tuning interface. Consequently, end-to-end optimisation of the BirdNET backbone was not practically available for this study. The literature-standard substitute is therefore to train a convolutional network on bird spectrograms, which is the standard BirdCLEF pipeline (Section 3.2). We initialise that architecture from ImageNet-pretrained weights rather than from scratch: with roughly 35k focal recordings under a 499× class imbalance, random initialisation is likely to underperform, and the evidence in Section 3.3 indicates that a strong pretrained backbone transfers better than from-scratch training on data of this scale. We fine-tune an EfficientNet-b0 backbone end-to-end on the 224×224 log-mel spectrograms, with a multi-label classification head, mixup, and SpecAugment. Figure 9 illustrates this model, which is shared by A2 and B2. We also implemented a closed-form adapter on the frozen embeddings as an *ablation*; consistent with Kumar et al. (2022), this light adaptation underperformed the A1 linear baseline, and it is reported only as an ablation rather than as the headline A2.

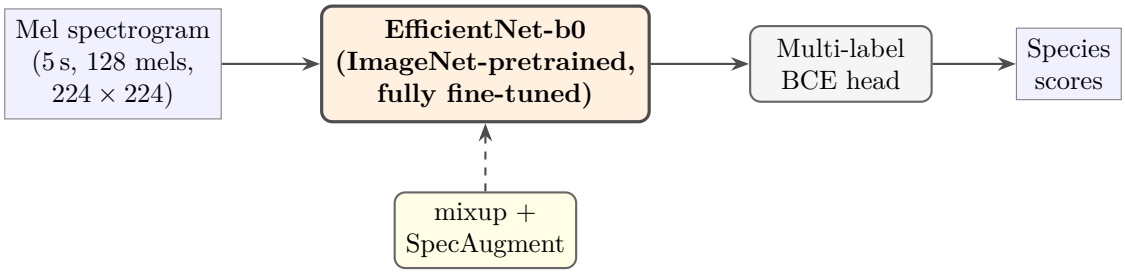

Figure 9. The A2/B2 model. A log-mel spectrogram is classified by an EfficientNet-b0 backbone that is fine-tuned end-to-end, with a multi-label binary-cross-entropy head and mixup and SpecAugment applied during training. B2 differs from A2 only in that the training pool also contains the generated audio.

Unlike A1, this architecture has no frozen feature extractor: the entire EfficientNet-b0 backbone is trainable and is fine-tuned end-to-end together with the classification head. The backbone is ImageNet-pretrained rather than randomly initialised, because with roughly 35k focal recordings under a severe class imbalance a strong pretrained initialisation transfers better than from-scratch training at this scale (Section 3.3). Full fine-tuning is a deliberate choice that runs against the caution of Kumar et al. (2022), and the contrast between A1 and A2 is precisely the test of when it pays off: here the backbone being adapted is a generic image network rather than the already-bioacoustic BirdNET probe, so it has more to gain from task adaptation, and the feature-distortion risk is countered by strong regularisation, namely mixup and SpecAugment together with a per-class positive weighting in the loss that keeps the rare classes from being ignored. The classification head is a single linear layer producing one logit per class, trained under a multi-label binary-cross-entropy objective so that overlapping vocalisations in the soundscapes can be represented. The full layer configuration and the training settings are given in Appendix G (Tables 13 and 14).

5.4 The synthetic-audio pipeline: AudioLDM 2 with a BirdNET filter

The synthetic-audio track implements the brief's track-B requirement directly. The brief frames this track as generating synthetic audio to balance all classes; in practice the pipeline targets the under-represented classes, and because no generated bird clip passed the BirdNET verification step (Section 6), the augmentation rebalances the non-bird tail rather than achieving balance across every class. For each under-represented class, a text prompt is constructed from the taxonomy and passed to AudioLDM 2 (Liu et al., 2024) to generate candidate clips. The prompt is built from two taxonomy fields, the species name and the taxon: the name is the common name where available and the scientific name otherwise, and the taxon selects a call-type phrase. For example, the bird class `sptnig1` yields the prompt "a natural field recording of a Spot-tailed Nightjar bird call". A fixed negative prompt, "low quality, music, speech, human voice, silence", is supplied alongside every prompt to steer generation away from non-target audio.

The taxonomy lookup tables are read directly from `taxonomy.csv`, which is the concrete sense in which the prompt is "constructed from the taxonomy". The per-class loop builds each prompt with `prompt_for`, passes it to the AudioLDM 2 diffusion pipeline (here `cvssp/audioldm2`, 50 diffusion steps, guidance scale 3.5, generating 10 s from which the loudest 5 s is cropped), and retains a clip only if it passes the BirdNET check. A representative code snippet of this procedure is given in Appendix E (Listing 1). Because AudioLDM 2 exceeded the memory available on the local machine, generation was run on a cloud GPU.

Each generated candidate is verified with a BirdNET check, as the brief suggests: a clip is kept only if BirdNET recognises it as the target species (for species BirdNET covers) or if its BirdNET embedding is sufficiently close to the class's real-example centroid (for taxa BirdNET does not cover). A representative code snippet of this check is given in Appendix E (Listing 2). Figure 10 illustrates this pipeline. The verification step is essential: it is the mechanism by which low-fidelity clips are discarded before they can mislead the classifier, and its keep-rate is itself an informative result (Section 6).

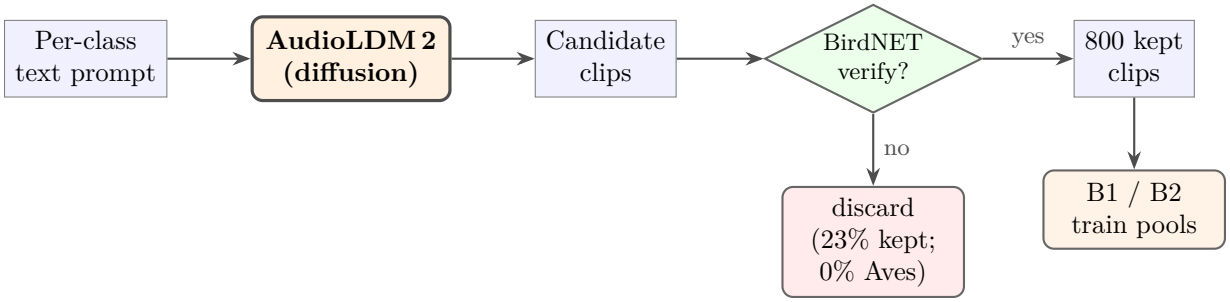

Figure 10. The track-B generation pipeline. AudioLDM 2 generates candidate clips from per-class text prompts; BirdNET verifies each clip and only verified clips enter the B1 and B2 training pools. Of 3,484 candidates, 800 (23%) passed the filter; none of the clips for the BirdNET-recognised bird species passed, an outcome discussed in Section 7.

5.5 B1 and B2: adding generated audio

The two B cells reuse the A models unchanged and differ only in the training pool. B1 embeds the kept generated clips with BirdNET and adds them to the A1 logistic-regression training pool. B2 adds the kept clips, as real audio, to the EfficientNet training pool of A2. Figure 11 shows one such kept clip beside a real recording of the same class. The validation set, including the four single-example classes, is never altered, so any change in validation macro-AUC is attributable to the generated training audio rather than to a change in what is measured.

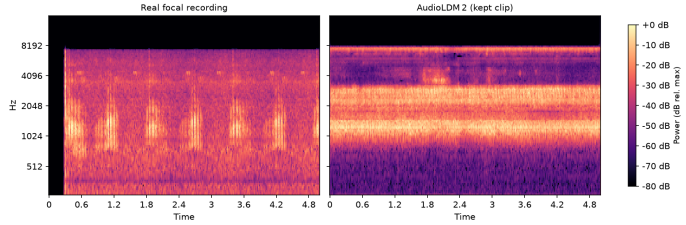

Figure 11. A real focal recording (left) and a verified AudioLDM 2 clip (right) for the same under-represented amphibian class (`23724`, Waxy Monkey Tree Frog), as log-mel spectrograms. The generated clip passed the BirdNET filter and entered the training pool, yet its time-frequency structure is visibly less defined than the real call. A full-size version is provided in Appendix D (Figure 14).

5.6 Evaluation metrics

The *primary* measure is *focal-validation macro-AUC*: for each class the area under its receiver-operating-characteristic curve is computed, and these per-class values are averaged with equal weight across classes. Intuitively, the AUC for a species is the probability that the model scores a held-out focal clip that does contain that species above one that does not, so 0.5 is no better than chance and 1 is perfect; macro-averaging then treats every species equally, so that ranking a rare amphibian well counts as much as ranking a common bird well. The macro average is the appropriate choice for this dataset because the class distribution is severely long-tailed, with the most populous class holding roughly 499 times as many recordings as the rarest; a sample-weighted score would be dominated by a few common classes, whereas equal weighting lets a change on the rare tail register. It is also the metric the BirdCLEF line of work and the Kaggle leaderboard adopt, which the study follows for comparability. Because AUC is threshold-free, it scores how well each species' clips are ranked without committing to a cutoff for calling a species present or absent; this suits the multi-label setting here, where a single clip may contain several of the 234 species and a separate decision threshold would otherwise have to be tuned for each one. All values are computed on the held-out focal-validation split. They are not reported on the soundscapes, because the labelled soundscape segments were part of training and their scores are therefore biased upward; the focal-validation metric measures performance on clean, single-species clips and does not by itself capture the focal-to-soundscape domain shift discussed in Section 4.5.

The *generation keep-rate* is the fraction of AudioLDM 2 candidate clips that pass the BirdNET verification check. It is a diagnostic of the synthetic-audio pipeline rather than of a classifier, and it is reported because the keep-rate, and especially its variation across taxa, is itself an informative result and one of the study's stated contributions. Its interpretation rests on the property that BirdNET outputs are species-specific detector scores rather than calibrated probabilities (Wood & Kahl, 2024): the keep-rate is most meaningful where BirdNET's verification is strongest, namely the bird species it can recognise, and is expected to behave differently for the taxa it does not cover.

Top-1 accuracy is the fraction of validation windows whose highest-scoring class is the true primary label. It is reported as a secondary, single-label measure alongside macro-AUC because the two capture different things: macro-AUC rewards ranking the correct class highly across all classes, whereas top-1 rewards placing the single correct class first. Reporting both makes visible a trade-off that the macro-AUC ranking alone would hide, in which a model can rank better on average while classifying the single label less often.

6 Results

The central question is whether the synthetic-audio track helps, and if so where: the long tail established in Section 4.3 and its taxon dependence (Section 4.4) predict that any benefit should appear among the rare, under-resourced classes rather than uniformly. The primary metric is focal-validation macro-AUC, defined with the other two measures in Section 5.6; recall that it weights rare and common classes equally and is computed on the held-out focal-validation split

Table 5. The three measures used in the Results section. Macro-AUC scores the classifiers, the keep-rate scores the synthetic-audio pipeline, and top-1 accuracy is a single-label complement that exposes a ranking-versus-classification trade-off. The rationale ties each measure to the dataset (Section 4) and the related work (Section 3).

Measure	What it quantifies	Role	Why it is used
Focal-validation macro-AUC	Mean over classes of per-class ROC-AUC on the focal-validation split	Primary	Equal per-class weight suits the long-tailed distribution; BirdCLEF/Kaggle standard; threshold-free
Generation keep-rate	Fraction of AudioLDM 2 candidates passing the BirdNET check	Pipeline diagnostic	Measures usable synthetic audio per taxon; BirdNET scores are species-specific, so the rate varies by taxon
Top-1 accuracy	Fraction of windows whose top-scoring class is the true label	Secondary	Single-label complement to macro-AUC; reveals the ranker-versus-classifier trade-off

rather than on the soundscapes. We report the two-by-two matrix, the generation keep-rate, the rare-tail breakdown, the per-taxon outcome, and the ranking-versus-classification trade-off, interpreting each in turn.

Table 6 presents the two-by-two matrix. Both synthetic-data cells exceed their original-data counterparts, which provides evidence that the filtered generated audio helps both model families. The fine-tuned model benefits more than the frozen baseline (a gain of +0.0084 versus +0.0034), and the fine-tuned-plus-generated cell, B2, attains the best result in the study at 0.9549.

Table 6. Focal-validation macro-AUC for the two-by-two matrix. Rows vary the data condition and columns vary the model. Both synthetic-data cells improve on their originals, and the fine-tuned model with generated audio (B2) is best overall.

	A. Frozen BirdNET + logreg	B. Fine-tuned EfficientNet
1. Original data	A1 = 0.9429	A2 = 0.9465
2. + AudioLDM 2 audio	B1 = 0.9463	**B2 = 0.9549**

Table 7 reports the generation keep-rate, which is the fraction of AudioLDM 2 candidates that passed the BirdNET filter. Two patterns stand out. First, the overall keep-rate is low (23%), and it varies strongly by taxon: insect clips are kept most often, whereas amphibian and mammal clips are kept rarely. Second, the keep-rate is 0% for the bird species that BirdNET can actually recognise, and the kept clips fall overwhelmingly in taxa that BirdNET does not directly recognise.

The keep-rate determines which classes were actually augmented, and the effect is uneven across taxa. The pattern is starkest for insects, whose targeted classes held only 18 real recordings between them yet received 625 synthetic clips, and for birds, whose targeted classes held 144 real recordings but received *none*, because every generated bird clip was rejected by the filter. In other words, the synthetic track topped up the non-bird tail substantially while leaving the bird tail unchanged. The full per-taxon real-versus-synthetic counts are given in Appendix I (Table 16).

The fidelity gap behind these counts is visible directly in the audio. Figure 11 (in Section 5.5) places a real focal recording next to one of the verified AudioLDM 2 clips for the same scarce amphibian class; the generated clip reproduces the broad spectral envelope but lacks the sharp, repeated call structure of the real recording. A full-size version is provided in Appendix D

Table 7. AudioLDM 2 generation keep-rate, that is, the fraction of generated candidates passing the BirdNET filter, by taxon and by whether BirdNET covers the species. No clip for a BirdNET-covered bird species passed the filter.

Group	Generated	Kept	Keep-rate
All	3,484	800	23.0%
Insecta	920	625	67.9%
Reptilia	32	25	78.1%
Amphibia	1,720	125	7.3%
Mammalia	392	25	6.4%
Aves	420	0	0.0%
BirdNET-covered species	240	0	0.0%
Not BirdNET-covered	3,244	800	24.7%

(Figure 14).

Table 8 breaks down per-class AUC by training-support band, where most of the headline difference lives. The generated audio lifts the rare and mid-tail bands in both families, and the fine-tuned model gains the most: at the band with at most five training examples, B2 reaches 0.890 against the A1 baseline's 0.666, the largest improvements in the table.

Table 8. Mean per-class AUC by training-support band. The bands count training examples per class; the B1 column counts *real* focal-training support while the A1 and B2 columns count focal-training support, so band membership differs by a few classes, but the direction is unambiguous. Generated audio lifts the rare and mid-tail bands, most strongly for the fine-tuned B2. The "all classes" row averages over the 203 classes that can be scored, that is, the classes with at least one positive in the focal-validation split; per-class AUC is undefined for the remaining classes (of the 234 targets, 206 have focal recordings and 203 of those appear in the validation split).

Training-support band	A1	B1 (+AudioLDM 2)	B2 (+AudioLDM 2)
≤ 5 examples	0.666	0.754	**0.890**
≤ 10 examples	0.764	0.825	**0.918**
all classes (203)	0.9429	0.9463	**0.9549**

The support-band view shows that the gains concentrate in the rare classes; the per-taxon view in Table 9 shows that they also concentrate in the taxa that actually received synthetic audio, which is the outcome predicted in Sections 4.4 and 5.4. Amphibians, the main taxon for which verified clips were generated, improve steadily from A1 to B1 to B2 (from 0.860 to 0.951), whereas birds, for which no generated clip passed the filter, are essentially unchanged (around 0.96 throughout). The synthetic benefit is therefore taxon-selective rather than uniform, and it appears precisely where the pipeline was able to add audio. The insect and reptile rows rest on very few scorable classes and are indicative only.

The same taxon pattern governs the extreme case of the four single-example classes that the brief forces entirely into validation. In the original-data condition two of them score at chance (0.5), having no training signal at all. Adding synthetic audio rescues exactly the two that received verified clips, both amphibians, which rise from 0.5 to near-perfect under B2 (0.9997 and 0.897); the reptile and the mammal do not benefit, and one degrades. Because each of these classes contributes a single validation point, the values are reported per class rather than aggregated, and Section 7 interprets what they do and do not show.

Finally, Table 10 reports a trade-off that the macro-AUC ranking alone hides. The fine-tuned cells win on macro-AUC but have *lower* top-1 accuracy than the frozen baseline. In other words,

Table 9. Mean per-class focal-validation AUC by taxon, for A1, B1, and B2. The count n is the number of scorable classes (those with at least one validation positive) in each taxon. Amphibians, which received most of the verified synthetic audio, improve markedly with the generated data, while birds, which received none, do not. The insect (n=3) and reptile (n=1) rows rest on very few classes and should be read as indicative only.

Taxon	n	A1	B1 (+AudioLDM 2)	B2 (+AudioLDM 2)
Aves	162	0.960	0.960	0.959
Amphibia	30	0.860	0.897	**0.951**
Insecta	3	0.994	0.994	0.968
Mammalia	7	0.890	0.856	0.903
Reptilia	1	0.850	0.669	0.801

the fine-tuned model is the better *ranker* (it tends to place the correct species highly across the board), while the frozen baseline is the better single-label *classifier*, in that it more often places the single correct species first. We report both so that the choice between them can be made against the intended use.

Table 10. Macro-AUC against top-1 accuracy across the four cells. The fine-tuned cells (A2, B2) win on macro-AUC but have lower top-1 accuracy than the frozen cells (A1, B1), reflecting a ranker-versus-classifier trade-off rather than a uniform improvement.

Cell	Macro-AUC	Top-1 accuracy
A1	0.9429	0.789
A2	0.9465	0.646
B1	0.9463	0.796
B2	**0.9549**	0.669

7 Discussion

We interpret the results in light of the assignment's question and the related work, and states the limitations. We compare the result against the line of work that motivated it, then assembles those readings into an answer to the research question, and finally bounds how far that answer should be taken.

7.1 Interpreting generative augmentation against prior bioacoustic classification work

Each design decision followed a strand of the related work, so the results are read here against what each strand predicts (Section 3).

Frozen bioacoustic embeddings with a light classifier (Section 3.1) predict data-efficient transfer, and A1 agrees, reaching a strong macro-AUC overall with only per-class linear heads (Table 6; Ghani et al. (2023)); the agreement is with data efficiency, not with robustness on the rarest band, where A1 is weakest (Table 8, Figure 15) and which the synthetic track targets.

The BirdCLEF lineage (Section 3.2) motivates the pipeline, which produced the headline result, but focal-validation macro-AUC does not test the focal-to-soundscape shift that this work names as central.

The caution that fine-tuning can distort good features under shift (Section 3.3; Kumar et al. (2022)) is both confirmed and bypassed: adapting the already-bioacoustic *BirdNET* embeddings falls below the A1 baseline (0.9231 against 0.9429), whereas the generic *ImageNet*-pretrained

backbone of A2 and B2 has little task structure to distort, so full fine-tuning helps, with strong augmentation limiting drift and the cost surfacing only as lower top-1 accuracy (Table 10).

Synthetic data for rare classes (Section 3.4) agrees: filtered audio improves both families, most in the rare bands (Table 6, Table 8), and adding rather than replacing real data follows the recommended "supplement, not replacement" design (Ghosh et al., 2025; Hoffman et al., 2025; Ronchini et al., 2024); but the filter kept few candidates and none for verifiable bird species (Table 7), consistent with general text-to-audio models being weaker on bird calls than bird-specific generators (Song & Ta, 2025).

Finally, on evaluation under imbalance (Section 3.5), a zero keep-rate reflects disagreement with BirdNET's class boundary rather than proven unfaithfulness (Wood & Kahl, 2024), single-example classes behave as case studies rather than stable estimates (Jana et al., 2025), and the best synthetic method depends on the downstream model: embedding-space synthesis helped the frozen probe more, real generated audio helped the fine-tuned model more.

7.2 Can generative AI improve classification?

Taken together, these readings answer the question the study set out to address, namely whether generative AI can improve classification on this long-tailed, multi-taxa task. The answer is that *it can, but conditionally*, and two conditions govern when:

1. the class must be data-scarce enough to have something to gain from extra examples; and
2. the generator must be able to produce audio for that class that passes the BirdNET verification check.

In this study both conditions were met, and the results bear them out. Adding BirdNET-filtered AudioLDM 2 audio improves both the frozen-embedding and the fine-tuned model on focal-validation macro-AUC, with the largest gain for the fine-tuned cell, so the benefit is not confined to one model family. Consistent with the first condition, the gains fall almost entirely in the data-scarce bands and not in the well-populated ones (Table 8, Figure 15); consistent with the second, they reach the non-bird taxa for which the generator produced verifiable audio, chiefly the amphibians, but not the bird classes, where no clip passed the filter (Table 9).

The qualified answer is off-the-shelf generated audio helps a class only when that class needs more data and the generator can make audio realistic enough to pass the filter for it. What limits the benefit on this task is the generator's ability to meet that bar, not the idea of using synthetic data itself.

7.3 Limitations of the synthetic-data approach

Several limitations bound how far that answer should be generalised. The kept clips are skewed toward a few taxa, so the benefit is uneven across the matrix; model capacity and the synthetic pool were both bounded by the GPU access available for this study, so the convolutional backbone was kept at EfficientNet-b0 rather than a larger variant, the synthetic pool was capped at a few dozen candidates per class, and only a subset of possible model variants could be explored in full (Appendix H lists those that were run); and the BirdNET filter, while appropriate, can only verify the taxa BirdNET covers, so its selectivity is itself taxon-dependent. The headline metric is computed on the focal-validation split and therefore does not measure the focal-to-soundscape shift, so the answer speaks to clean-clip performance rather than to deployment on soundscapes. Finally, several of the supporting sources in the synthetic-augmentation literature are preprints rather than peer-reviewed studies, so the comparisons drawn from them are indicative rather than settled. None of these undermine the central finding, but each bounds how far it should be taken.

8 Conclusion

This report asked whether generative AI can improve classification on a long-tailed, multi-taxa bioacoustic task, and answered it through a brief-compliant two-by-two study on birdclef-2026. The evidence indicates that adding AudioLDM 2-generated audio, filtered by a BirdNET check, improves both the frozen-embedding and the fine-tuned model on focal-validation macro-AUC, with the fine-tuned model reaching the best result of 0.9549. The overall gain is modest because the metric is already high on the common classes; the improvement is concentrated where real data is scarcest, where the fine-tuned model with generated audio raises mean per-class AUC on the rarest classes (those with at most five training examples) from 0.666 to 0.890, a relative gain of about 34% over the frozen baseline (Table 8). The benefit is selective rather than uniform: the generation keep-rate is low overall and is zero for the bird species BirdNET can verify, which indicates that the gains come mainly from taxa whose vocalisations are easier to synthesise, and that the single-example classes cannot be reliably rescued.

Several directions may strengthen the synthetic-audio track in future work. Species-conditioned or bird-specialised generative models could raise the keep-rate for birds, which is where off-the-shelf generation failed verification here; recent bird-specific diffusion work that reports markedly higher class fidelity than general models (Song & Ta, 2025) indicates that this direction is promising. Scaling generation beyond the compute bound used in this study would provide a larger and less taxon-skewed synthetic pool. Finally, the BirdNET filter could be loosened carefully so that more clips are kept, accepting some lower-fidelity audio in exchange for wider coverage, as long as the effect on validation accuracy is checked rather than assumed. Together, these directions would test whether the benefit seen here, which currently reaches only some taxa, can be spread more evenly across all of them.

References

Cañas, J. S., et al. (2025). Overview of BirdCLEF+ 2025: Multi-Taxonomic Sound Identification in the Middle Magdalena, Colombia. *CLEF 2025 Working Notes (CEUR-WS)*. https://www.researchgate.net/publication/396180283_Overview_of_BirdCLEF_2025_Multi-Taxonomic_Sound_Identification_in_the_Middle_Magdalena_Colombia

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Ghani, B., Denton, T., Kahl, S., & Klinck, H. (2023). Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, *13*, 22876. https://doi.org/10.1038/s41598-023-49989-z

Ghosh, S., Kumar, S., Kong, Z., Valle, R., Catanzaro, B., & Manocha, D. (2025). Synthio: Augmenting Small-Scale Audio Classification Datasets with Synthetic Data (Y. Yue, A. Garg, N. Peng, F. Sha, & R. Yu, Eds.). *2025*, 66689–66711. https://proceedings.iclr.cc/paper_files/paper/2025/file/a6e1f6963f65bcc4854691a15460dbd8-Paper-Conference.pdf

Hagiwara, M., Hoffman, B., Liu, J.-Y., Cusimano, M., Effenberger, F., & Zacarian, K. (2023). BEANS: The Benchmark of Animal Sounds, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096686

Hoffman, B., et al. (2025). Synthetic Data Enables Context-Aware Bioacoustic Sound Event Detection. *arXiv preprint arXiv:2503.00296*. https://arxiv.org/abs/2503.00296

Jana, A., Uili, M., Atherton, J., O'Brien, M., Wood, J., & Brickson, L. (2025). An Automated Pipeline for Few-Shot Bird Call Classification: A Case Study with the Tooth-Billed Pigeon. https://arxiv.org/abs/2504.16276

Kahl, S., Denton, T., Klinck, H., Ramesh, V., Joshi, V., Srivathsa, M., Anand, A., Arvind, C., Cp, H., Sawant, S., Glotin, H., Goëau, H., Vellinga, W.-P., Planqué, R., & Joly, A. (2024). Overview of BirdCLEF 2024: Acoustic identification of under-studied bird species in the Western Ghats. *Conference and Labs of the Evaluation Forum (CLEF 2024)*, 1948–1957. https://ceur-ws.org/Vol-3740/paper-184.pdf

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do Better ImageNet Models Transfer Better? *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html

Kumar, A., Raghunathan, A., Jones, R., Ma, T., & Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2202.10054

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., & Plumbley, M. D. (2024). AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 2871–2883. https://doi.org/10.1109/TASLP.2024.3399607

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. https://arxiv.org/abs/1904.08779

Rauch, L., Schwinger, R., Wirth, M., Heinrich, R., Huseljic, D., Herde, M., Lange, J., Kahl, S., Sick, B., Tomforde, S., & Scholz, C. (2025). BirdSet: A Dataset and Benchmark for Classification in Avian Bioacoustics. *International Conference on Learning Representations*, *2025*, 29482–29520. https://proceedings.iclr.cc/paper_files/paper/2025/file/484d254ff80e99d543159440a06db0de-Paper-Conference.pdf

Ronchini, F., Serizel, R., Cornell, S., Wisdom, S., Hershey, J. R., Plakal, M., & Ellis, D. P. W. (2024). Synthetic Training Set Generation Using Text-To-Audio Models for Environmental Sound Classification. *arXiv preprint arXiv:2403.17864*. https://arxiv.org/abs/2403.17864

Song, T., & Ta, T. V. (2025). Towards high-fidelity and controllable bioacoustic generation via enhanced diffusion learning. *arXiv preprint arXiv:2509.00318*. https://arxiv.org/abs/2509.00318

Tan, M., & Le, Q. V. (2019, September). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97). PMLR. https://proceedings.mlr.press/v97/tan19a.html

van Merriënboer, B., Dumoulin, V., Hamer, J., Harrell, L., Burns, A., & Denton, T. (2025). Perch 2.0: The Bittern Lesson for Bioacoustics. *arXiv preprint arXiv:2508.04665*. https://arxiv.org/abs/2508.04665

Wikipedia contributors. (2026). *Pantanal*. Wikipedia, The Free Encyclopedia. Retrieved June 15, 2026, from https://en.wikipedia.org/wiki/Pantanal

Wood, C. M., & Kahl, S. (2024). Guidelines for appropriate use of BirdNET scores and other detector outputs. *Journal of Ornithology*. https://doi.org/10.1007/s10336-024-02144-5

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, *27*. https://proceedings.neurips.cc/paper_files/paper/2014/file/532a2f85b6977104bc93f8580abbb330-Paper.pd

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond Empirical Risk Minimization. https://arxiv.org/abs/1710.09412

Appendices

A BirdNET reference performance (Kahl et al., 2021)

This appendix records the headline performance and training-data figures reported for BirdNET by Kahl et al. (2021), the source model used in this study. They are reproduced here to substantiate the claim in Section 3.1 that BirdNET is a robust avian classifier trained on large collections of recordings, and to motivate its use as the embedding extractor for the A1 and B1 cells and as the verification model for the synthetic-audio filter. The values are drawn from the abstract and results of Kahl et al. (2021) and are presented in Table 11.

Table 11. Reference performance and scale of BirdNET, as reported by Kahl et al. (2021). The evaluation figures are for the 2021 model release; this study uses a later BirdNET release for embedding extraction, so these values characterise the model family rather than the exact checkpoint.

Metric / property	Reported value
Focal (single-species) mean average precision	0.791
Annotated soundscape F0.5 score	0.414
Hotspot occurrence correlation (r)	0.251 (121 species, 4 years)
Training data (Xeno-canto)	>500,000 recordings, >10,000 species, >7000 h
Training data (Macaulay Library)	>750,000 recordings
Architecture	Wide ResNet, 157 layers (36 weighted)
Input representation	64×384 mel spectrogram, 3 s @ 48 kHz

The training-data scale supports the description of BirdNET as learned from large recording collections, while the gap between the focal score (0.791) and the annotated-soundscape score (0.414) is an early, in-source illustration of the focal-to-soundscape domain shift discussed in Section 3.2.

B Full-size augmentation panel

This appendix provides a full-size version of the augmentation figure summarised inline in Section 5.1 (Figure 7). It is reproduced at full size in Figure 12 so that the effect of each augmentation on the time-frequency structure is legible, including the masked bands of SpecAugment and the frequency and timing changes introduced by the waveform augmentation.

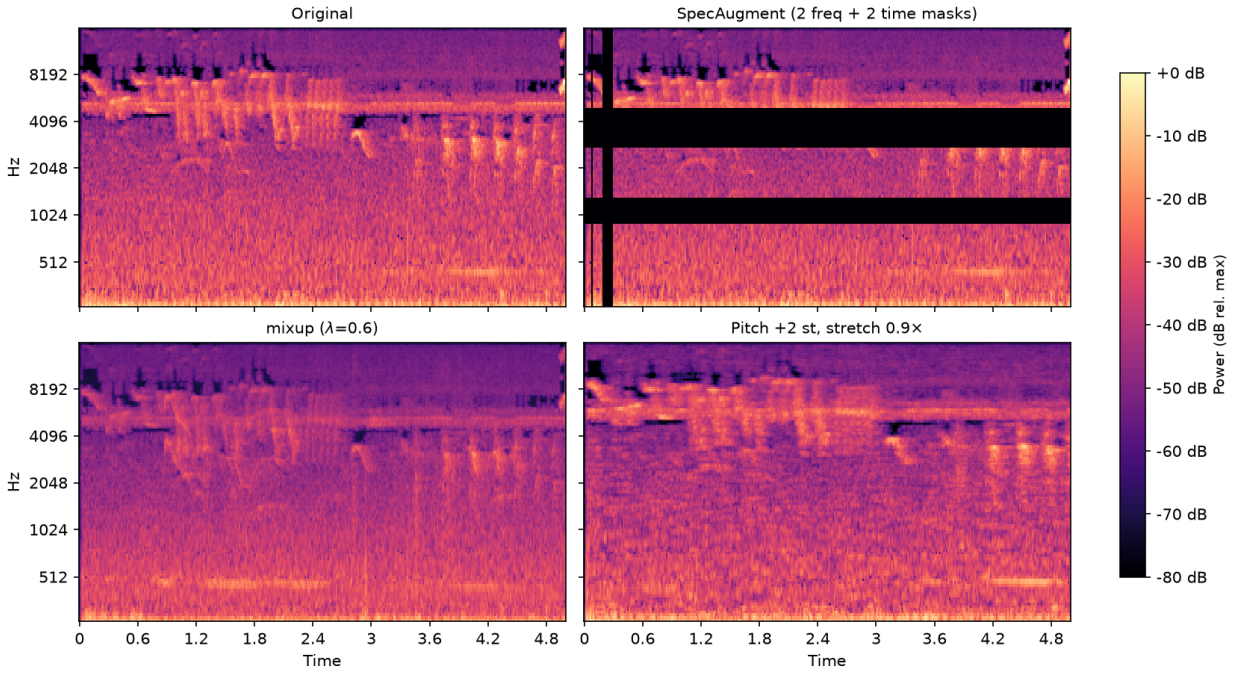

Figure 12. The three spectrogram-domain augmentations used by the convolutional cells, each applied to the same focal recording (a real Xeno-Canto clip, `crebec1/XC70545`). The original log-mel spectrogram (top left) is shown alongside SpecAugment, which blanks two frequency and two time bands; mixup, which convexly blends the clip with a second recording; and the B2 waveform augmentation, which pitch- and time-shifts the audio before the mel transform. The augmentation parameters that are random at training time are fixed here to representative values for clarity (mixup λ=0.6, a +2-semitone pitch shift, and a 0.9× time stretch), and the panels are drawn at the native mel resolution rather than the 224×224 tensor the network consumes.

C Representative spectrograms across taxa

This appendix provides the per-taxon spectrograms referenced in Section 4.6. Figure 13 shows a representative log-mel spectrogram for each of the five taxa, making visible the differences in time-frequency structure that motivate the taxon-dependent behaviour of a bird-trained embedding and a bird-trained verification filter.

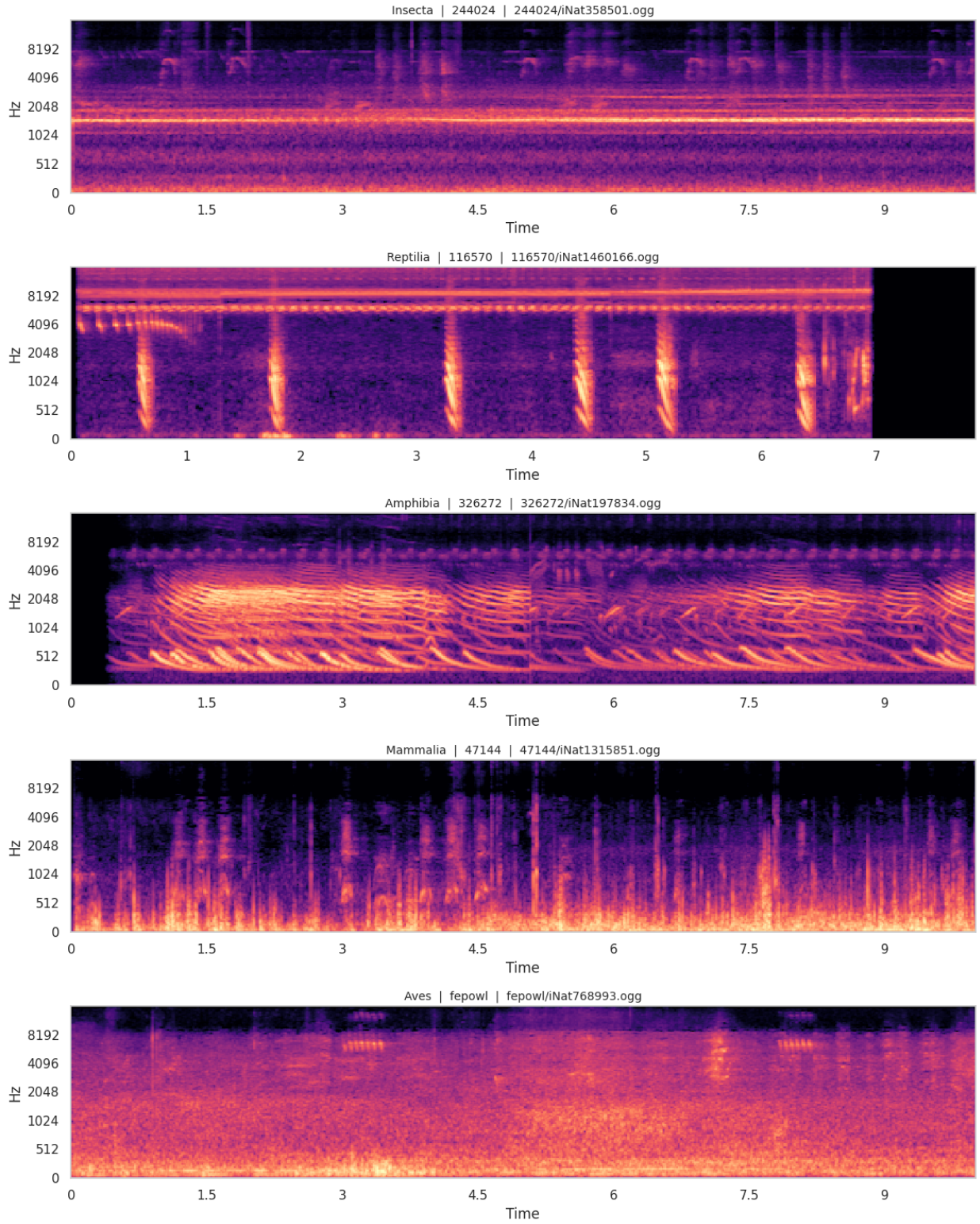

Figure 13. Representative log-mel spectrograms across the five taxa. The time-frequency structure differs substantially between groups, which is the data-side reason a bird-trained embedding and a bird-trained verification filter are best matched to the bird classes and least matched to the others.

D Real versus synthetic audio: full-size comparison

This appendix provides a full-size version of the real-versus-synthetic comparison. Figure 14 places a real focal recording next to a verified AudioLDM 2 clip for the same under-represented amphibian class, so that the difference in time-frequency structure between real and generated audio is legible.

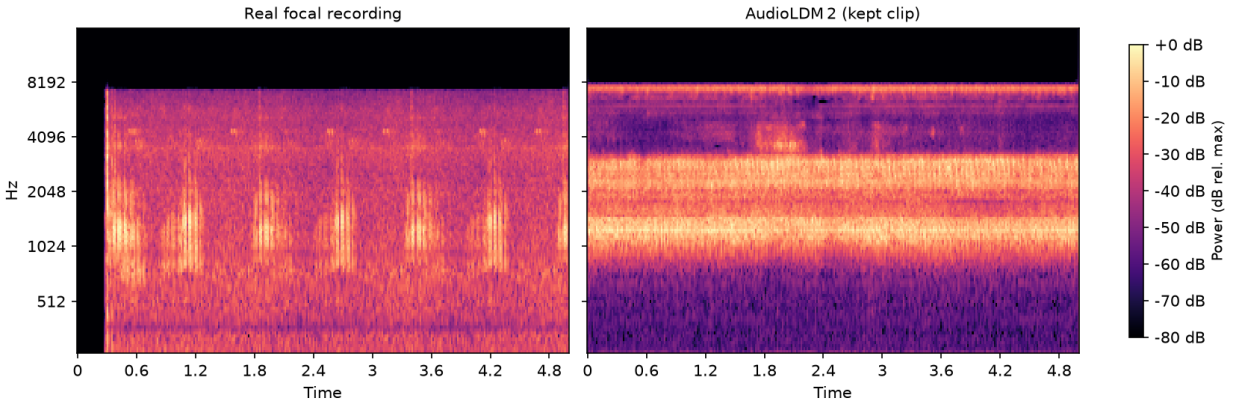

Figure 14. A real focal recording (left) and a verified AudioLDM 2 clip (right) for the same under-represented amphibian class (23724, Waxy Monkey Tree Frog), shown as log-mel spectrograms. The synthetic clip is one of the clips that passed the BirdNET verification filter and was added to the B1 and B2 training pools. It reproduces the broad spectral envelope of the call but lacks the sharp, repeated temporal structure of the real recording, which illustrates the fidelity gap that the keep-rate (Table 7) measures in aggregate.

E Synthetic-audio generation: code snippet

This appendix provides the code snippet for the synthetic-audio generation pipeline summarised in Section 5.4. Listing 1 is a condensed, representative version of what was implemented in the generation notebook (`b_audioldm2_generate_colab.ipynb`): it shows how the taxonomy is turned into a text prompt, how that prompt is passed to AudioLDM 2, and how each generated clip is verified before it is kept. It is intended to make the procedure reproducible in outline rather than to be run verbatim; docstrings, error handling, batching, file input and output, and device setup are omitted for readability, but the parameters shown are the values actually used.

```
from diffusers import AudioLDM2Pipeline

# 1. Build the taxonomy lookup tables from the competition's taxonomy.csv.
taxonomy = pd.read_csv(DATA_DIR / 'taxonomy.csv')
TAX = taxonomy.set_index('primary_label')
COMMON_NAME = TAX['common_name'].to_dict()
SCI_NAME    = TAX['scientific_name'].to_dict()
CLASS_NAME  = TAX['class_name'].to_dict()       # taxon (Aves, Amphibia, ...)

# 2. Construct the text prompt for a class from its taxonomy entry.
def prompt_for(label):
    name = COMMON_NAME.get(label) or SCI_NAME.get(label)
    kind = {'aves': 'bird call',     'amphibia': 'frog call',
            'insecta': 'insect call', 'mammalia': 'mammal call',
            'reptilia': 'reptile call'}.get(CLASS_NAME.get(label, '').lower(),
                                            'animal call')
    return f'a natural field recording of a {name} {kind}'

# 3. Load the AudioLDM 2 text-to-audio diffusion model.
NEG_PROMPT = 'low quality, music, speech, human voice, silence'
pipe = AudioLDM2Pipeline.from_pretrained('cvssp/audioldm2',
                                         torch_dtype=torch.float16).to(DEVICE)

def generate_clips(prompt, n):
    clips = []
    for _ in range(n):
        out = pipe(prompt, negative_prompt=NEG_PROMPT,
                   num_inference_steps=50, guidance_scale=3.5,
                   audio_length_in_s=10.0).audios          # AudioLDM 2 -> 16 kHz
        clips.append(np.asarray(out[0], dtype='float32'))
    return clips

# 4. For each scarce class: generate, crop the loudest 5 s, BirdNET-verify, save.
for c in rare_classes:
    prompt = prompt_for(c)
    while kept < rare_target and n_attempt < n_gen_per_class:
        for w in generate_clips(prompt, batch):
            crop = loudest_window(w, gen_sr)               # loudest 5 s window
            if passes_birdnet(crop, gen_sr, c):            # keep only verified clips
                save(resample(crop, gen_sr, dataset_sr))
```

Listing 1. A representative code snippet of the synthetic-audio generation pipeline: constructing the text prompt from the taxonomy and generating BirdNET-verified synthetic clips with AudioLDM 2. This is a condensed version of the implementation in the generation notebook; docstrings and error handling are omitted for brevity, and the parameters shown are the values actually used.

Listing 2 expands the `passes_birdnet` step referenced in Section 5.4. For each scarce class a centroid is first built from a few of its real focal recordings, embedded with BirdNET and averaged;

the acceptance threshold is the lower quartile of how close those real seeds sit to their own centroid. A generated clip is then kept by one of two checks: for species BirdNET recognises, the target species must appear in BirdNET's top-k predictions above a confidence floor; for taxa BirdNET does not cover, the clip's BirdNET embedding must be at least as close to the class centroid as that threshold. The few classes with neither BirdNET coverage nor a usable real seed are accepted unverified and recorded separately, so they can be distinguished from filter-verified clips.

```
# Build a real-example centroid and an acceptance threshold for each scarce class.
for c in rare_classes:
    embs = [l2n(birdnet_embed(real_clip)) for real_clip in real_seeds(c)[:8]]
    if embs:
        centroid[c] = l2n(mean(embs))            # class's real-example centroid
        sims = [e @ centroid[c] for e in embs]
        tau[c] = percentile(sims, 25)            # keep threshold = lower-quartile
        ↪ self-similarity

# Decide whether one generated clip is kept.
def passes_birdnet(clip, label, sci):
    topk, emb = birdnet_predict(clip)            # top-k species confidences + 1024-d
    ↪ embedding

    # (a) species-recognition check, for classes BirdNET covers
    if birdnet_covers(sci):
        for species, conf in topk[:5]:           # birdnet_topk = 5
            if species == sci and conf >= 0.10:  # birdnet_min_conf = 0.10
                return True                      # recognised as the target species ->
                ↪ keep
        return False                             # covered but not recognised ->
        ↪ reject

    # (b) embedding-centroid check, for taxa BirdNET does not cover
    if label in centroid:
        return (l2n(emb) @ centroid[label]) >= tau[label]   # close enough to real
        ↪ seeds -> keep

    # (c) no coverage and no real seed -> accept conservatively, flagged as unverified
    return True
```

Listing 2. A representative code snippet of the BirdNET verification check (`passes_birdnet`). Covered species are kept when BirdNET ranks the target species highly enough; uncovered taxa are kept when the generated clip's embedding is close enough to a centroid of the class's real recordings. As in Listing 1, this is a condensed version of the implementation; the thresholds shown are the values actually used.

F A1 architecture configuration

This appendix gives the stage-by-stage configuration of the A1 cell referenced in Section 5.2. Table 12 lists each component, its output, and whether it is frozen or trainable, making explicit that the only learned parameters are those of the per-class linear heads while the BirdNET backbone is used as a fixed feature extractor.

Table 12. The A1 architecture and its frozen/trainable split. The only learned component is the bank of per-class linear heads; the BirdNET backbone is used as a fixed feature extractor and is never updated. Parameter counts are for the trainable heads only.

Stage	Operation	Output	Trainable
Input	5 s window, resampled to 48 kHz	waveform	no
BirdNET backbone	Wide ResNet feature extractor, two 3 s sub-windows mean-pooled	1024-d embedding	frozen
Normalisation	L_2-normalisation	1024-d	no
Classifier	per-class logistic regression, one-vs-rest (C=1.0, class-balanced)	224 class scores ($\approx$ 229.6k params)	yes

G A2 architecture and training configuration

This appendix gives the configuration of the A2 cell (shared with B2) referenced in Section 5.3. Table 13 lists the architecture stage by stage, and Table 14 lists the training settings. In contrast to A1, every component is trainable: the EfficientNet-b0 backbone is fine-tuned end to end rather than used as a fixed feature extractor.

Table 13. The A2 (and B2) architecture. The model is fine-tuned end to end, so all components are trainable. The backbone parameter count is the published figure for EfficientNet-b0; the head count is for the final linear layer.

Stage	Operation	Output	Trainable
Input	5 s window @ 32 kHz	waveform	no
Front-end	log-mel (128 mels, 1024-pt FFT, hop 512, 50–16,000 Hz), standardised, resized	$3 \times 224 \times 224$ image	no
Backbone	EfficientNet-b0 (`tf_efficientnet_b0_ns`, ImageNet-pretrained, $\approx$ 5.3M params)	1280-d features (pooled)	yes
Head	linear, $1280 \to 234$	234 logits ($\approx$ 0.30M params)	yes
Output	sigmoid (multi-label)	234 class scores	no

Table 14. Training configuration for A2 and B2. The two cells differ only in the training pool (B2 adds the BirdNET-filtered generated clips); all settings below are shared.

Setting	Value
Optimiser	AdamW
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Schedule	1-epoch warmup, then cosine decay
Epochs	12
Batch size	64
Precision	mixed precision (AMP)
Loss	multi-label BCE with per-class positive weight (capped at 50)
Soundscape weighting	10$\times$ relative to focal windows
Augmentation	mixup (α=0.15) and SpecAugment (2 frequency, 2 time masks)
Input	224×224, three-channel log-mel spectrogram

H Model variants explored

This appendix records the model and data variants that were run during the study, beyond the four headline cells of the two-by-two matrix. They were explored within the GPU budget available, which bounded both the model capacity and how many variants could be carried through to a full run. The four headline cells were chosen for compliance with the brief's two-by-two design and for a clean, directly comparable comparison, rather than because they maximised the score. Table 15 lists all of the runs, and the variants were set aside for reasons that differ by type.

The two embedding-space methods, SMOTE on the frozen embeddings and the generative oversampler, in fact reach a slightly higher focal-validation macro-AUC than the reported B1. They were not adopted as the headline because they synthesise new examples in BirdNET's embedding space, by interpolating and jittering rare-class embeddings, rather than as audio. They therefore sit outside the brief's track-B method, which is to generate audio and then verify it, and they answer an easier question than the one the study set out to address; they are reported here as a comparison. The closed-form adapter variants, namely the A2 ablation and the two B2 adapter rows, adapt the frozen embeddings instead of fine-tuning the backbone end to end; consistent with the feature-distortion point in Section 7.1, they trail either the A1 linear baseline or the fine-tuned cells. Finally, the augmentation-only B2 row removes the generated audio and is included as an isolation ablation, to show what the generated audio adds on top of augmentation alone.

Table 15. Model and data variants run during the study, with focal-validation macro-AUC. The first four rows are the headline two-by-two cells; the rest are ablations and alternatives explored within the GPU budget. All values are focal-validation macro-AUC, comparable to those in the main text.

Variant	Description	Focal macro-AUC
A1	Frozen BirdNET embeddings + per-class logistic regression	0.9429
A2	Fine-tuned EfficientNet-b0	0.9465
B1	A1 with BirdNET-filtered AudioLDM 2 audio	0.9463
B2	A2 with BirdNET-filtered AudioLDM 2 audio	**0.9549**
A2 (ablation)	Closed-form adapter on the frozen embeddings	0.9231
B1 (variant)	A1 with SMOTE on the embeddings	0.9539
B1 (variant)	A1 with a generative oversampler	0.9541
B2 (variant)	Augmentation only, no generated audio	0.9489
B2 (variant)	Adapter with generated audio	0.9344
B2 (variant)	Adapter with SMOTE	0.9345

I Additional results

This appendix collects two supporting results referenced from Section 6: the per-taxon real-versus-synthetic supply (Table 16) and the support-band AUC comparison between A1 and B2 (Figure 15). Both expand on points made in the main text; the headline results remain in Section 6.

Table 16. Real versus synthetic supply for the under-represented classes targeted by the generator, by taxon. *Targeted* counts the scarce classes for which generation was attempted; *Real* sums their existing focal recordings; *Filled* counts those that received at least one verified clip; and *Synthetic added* counts the kept AudioLDM 2 clips. Birds were targeted but received no synthetic audio, since all generated bird clips failed verification.

Taxon	Targeted classes	Real recordings	Filled classes	Synthetic added
Insecta	27	18	25	625
Amphibia	31	264	5	125
Mammalia	7	56	1	25
Reptilia	1	1	1	25
Aves	7	144	0	0
All	73	483	32	800

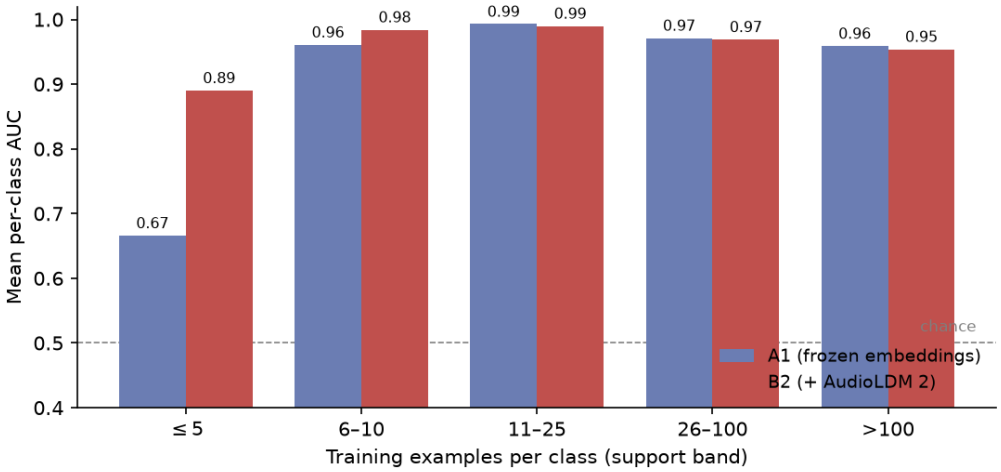

Figure 15. Mean per-class focal-validation AUC by training-support band, for the frozen-embedding baseline (A1) and the fine-tuned synthetic-data cell (B2). The two cells are close except in the smallest band, where A1 falls to 0.67 and B2 reaches 0.89. Band membership is by real focal-training support and differs by a few classes between A1 and B2 because B2's pool also contains generated clips, as noted for Table 8.