# REGION-PROTOTYPE ATTENTION FOR CALIBRATED FEW-SHOT MULTI-LABEL AUDIO TAGGING

**Anonymous authors**
**Paper under double-blind review**

## ABSTRACT

Few-shot multi-label audio tagging, the task of assigning several co-occurring sound labels to a clip from only a handful of examples per novel class, has been studied only in narrow, non-standardised settings, and we find that standard prototypical models expose a subtle but serious failure: while their *ranking* of labels is strong, their *decisions* are badly miscalibrated, with models over-predicting and assigning almost every candidate label to each query. We make three contributions. (i) We establish a few-shot multi-label benchmark on FSD50K and AudioSet-Balanced with a disjoint base/novel class split, a chance floor, and identically trained baselines in single-class and compositional (LC-Protonets-style) prototypical networks, a frozen AudioSet Transformer, and a post-hoc temperature-scaling calibration baseline. (ii) We show that over-prediction and miscalibration are pervasive across convolutional, compositional, and transformer backbones; a large AudioSet-pretrained Transformer attains high recall only by predicting nearly all labels, with an expected calibration error (ECE) far above that of metric-based models. (iii) We introduce **RAFTa**, whose *region-prototype readout* lets class prototypes attend over a query's time–frequency regions before matching. RAFTa is the *best-calibrated* model in every setting, with the lowest ECE among all baselines and the predicted label-cardinality closest to the truth, while remaining competitive on ranking (mAP), and its calibration edge is not matched by post-hoc temperature scaling of a strong baseline. Code and benchmark splits are released via an anonymised repository.

## 1 INTRODUCTION

Understanding real-world audio requires models that reason over complex, overlapping sound events and generalize from limited supervision. *Audio tagging*—assigning one or more class labels to an audio clip—has advanced through Convolutional Neural Networks (CNNs) and Transformer architectures that learn robust mappings from spectrogram features to semantic labels Kong et al. (2020); Gong et al. (2021). Yet most of this progress assumes abundant labeled data per class. In practice, many acoustic categories are rare, annotation is costly, and recordings are inherently *multi-label*: several sound events (e.g., *speech*, *traffic*, and *a dog barking*) co-occur within a single clip. Systems that must recognize new sounds from only a few examples, such as bioacoustic monitoring, industrial fault detection, and on-device personalisation, are therefore outside the regime in which current taggers are trained and evaluated.

These pressures converge in *few-shot multi-label audio tagging*: recognising novel sound classes, several at a time, from only a handful of labelled examples per class. This setting is both difficult and, to date, largely unbenchmarked. Prior few-shot audio research centres on single-label classification or on sound-event detection, and multi-label few-shot tagging on the standard tagging datasets (FSD50K, AudioSet) lacks agreed class splits, chance references, and comparable baselines—so it is hard even to say what a good number is. We argue this gap conceals a concrete and consequential failure mode. When we evaluate strong prototypical models under an honest base/novel protocol, we find that they *rank* candidate labels well but make poorly *calibrated* decisions: they systematically over-predict, assigning close to the maximum number of labels to every query regardless of its true content. A large AudioSet-pretrained Transformer, used as a frozen feature extractor, exhibits this most starkly—it reaches near-perfect recall only by predicting almost every label, at the cost of pre-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

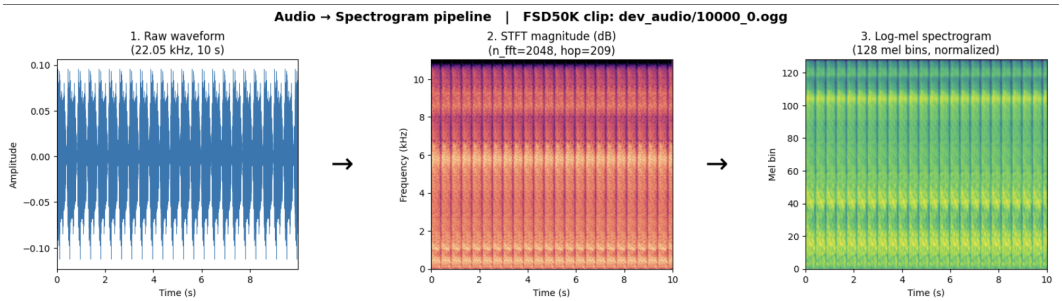


Figure 1: From raw audio to model input, on an FSD50K clip. **(1)** The waveform is resampled to 22.05 kHz, made mono, and fixed to 10 s. **(2)** An STFT ($n_{\mathrm{FFT}}$=2048, hop=209) gives a linear-frequency magnitude spectrogram. **(3)** A 128-band mel filterbank, decibel conversion, and normalization produce the log-mel spectrogram passed to the encoder.

cision and a severely inflated expected calibration error. In a multi-label system that must ultimately *emit* a label set, this decision-quality gap matters as much as ranking.

### 1.1 Our Approach

We propose **RAFTa** (**R**egion-**A**ttentional **F**ew-shot **T**agging of **a**udio), whose key component is a region-prototype readout for few-shot multi-label audio tagging. RAFTa encodes a log-mel spectrogram as a grid of time–frequency *region* nodes and optionally refines them with a lightweight Graph Attention Network (GAT) so that acoustic context propagates across non-local regions. Unlike a purely temporal encoder that averages frequency away, and complementary to patch-token Transformers such as AST Gong et al. (2021) (which retain spatial structure via positional embeddings but are data- and compute-hungry), the graph reasons explicitly over sparse relations between spectro-temporal regions. Its central component, however, is not the graph but a *region-prototype readout*. Standard prototypical matching pools each clip to a single vector and takes a dot product with a fixed class prototype, discarding *where* in the spectrogram the evidence for a label lives. Instead, we let each class prototype *attend* over the query's region nodes with multi-head cross-attention, pooling the spectro-temporal evidence most relevant to that class before matching. Because this attention sits on the forward path, gradients flow back into the encoder, so the readout *shapes* the learned representation rather than merely re-ranking a frozen one.

Prototype–region attention has been studied for single-label few-shot *image* classification Doersch et al. (2020); Hou et al. (2019); our contribution is to adapt it to few-shot *multi-label audio* and to show that, in this regime, its principal benefit is not higher ranking but markedly better-calibrated, lower-cardinality decisions. Prototypes are formed from a small support set within a standard pretrain-then-meta-train recipe. Importantly, every backbone we study—our CNN encoder and the Transformer baseline alike—is initialized from AudioSet-pretrained weights, so our comparisons isolate architecture and readout rather than differences in pretraining data.

### 1.2 Contributions

Our contributions are threefold:

- A **few-shot multi-label benchmark** on FSD50K and AudioSet-Balanced, with a disjoint base/novel class split, an explicit chance floor, and identically-trained baselines under one protocol: single-class and compositional (LC-Protonets-style Papaioannou et al. (2024)) prototypical networks, a frozen AudioSet Transformer Gong et al. (2021), and a post-hoc temperature-scaling calibration baseline Guo et al. (2017).
- A **diagnosis of over-prediction and miscalibration** in few-shot multi-label audio tagging. We show the failure persists across convolutional, compositional, and transformer backbones, and we quantify it with label-cardinality and calibration (ECE) analyses in addition to ranking metrics.

- The **region-prototype readout** (RAFTa), the **best-calibrated** model in every setting: it attains the lowest ECE among all baselines and the closest-to-true label-cardinality while remaining competitive on mAP, and we show this edge is *not* reproduced by temperature-scaling a strong baseline, so it stems from the architecture rather than post-hoc rescaling. A diagnostic further shows the graph-attention stage is well-conditioned but contributes modestly, attributing the gains to the readout.

## 2 Related Work

**Audio tagging.** CNNs on log-mel spectrograms remain standard, with PANNs Kong et al. (2020) a common pretrained backbone; Transformers such as AST Gong et al. (2021) add global context. AST retains time–frequency structure through patch tokens and positional embeddings, but is compute-hungry and, as we show, poorly calibrated as a frozen few-shot extractor. Recent audio foundation models (BEATs Chen et al. (2023), CLAP Elizalde et al. (2023)) are stronger but heavier; we use AST as a representative frozen transformer and leave foundation-model backbones to future work.

**Multi-label few-shot audio.** Prior multi-label few-shot audio work Cheng et al. (2019); Liang et al. (2024) uses bespoke datasets and matching schemes, and community few-shot sound-event-detection challenges (DCASE Nolasco et al. (2022)) target temporal *detection* rather than multi-label clip *tagging*; none report calibration. We instead provide a standardized episodic benchmark on FSD50K and AudioSet-Balanced with disjoint base/novel splits, a chance floor, and—unlike prior work—explicit calibration and label-cardinality analysis alongside ranking.

**Graph methods.** GNNs model audio's non-Euclidean structure for event detection; AT-GNN Singh et al. (2024) attends between spectrogram patch and label nodes. We adapt this patch–label idea to the few-shot regime, using support prototypes as the label nodes. A separate line places the graph over *episode examples* García Satorras & Bruna (2018); Wang et al. (2020); RAFTa's graph is instead *within* a clip, over time–frequency regions—orthogonal to, and composable with, sample-level graphs.

**Metric-based few-shot.** Prototypical Snell et al. (2017), matching Vinyals et al. (2016), and relation Sung et al. (2018) networks are the standard metric-based few-shot learners; we adopt the prototypical formulation. LC-Protonets Papaioannou et al. (2024) extend prototypes to multi-label tagging via compositional label-subset prototypes; we include a compositional-prototype baseline in this spirit on our shared encoder. Prototype–region cross-attention itself exists for single-label few-shot *images* Doersch et al. (2020); Hou et al. (2019); we bring it to multi-label *audio*, where its benefit is calibration and cardinality control rather than ranking. For calibration, temperature scaling Guo et al. (2017) is the standard post-hoc remedy; we compare against it to establish that our gains are architectural rather than a rescaling effect.

## 3 Methodology

We describe the RAFTa pipeline in four parts: spectrogram preparation (Fig. 1); a graph-attentional encoder that produces time–frequency region nodes; the region-prototype readout that matches those regions to class prototypes; and the base/novel protocol under which prototypes are formed and evaluated.

### 3.1 Data Preparation and Augmentation

**Spectrogram Conversion.** Raw audio $x(t)$ is resampled to $f_s = 22{,}050$ Hz, converted to mono, and either trimmed or tiled to a fixed 10-second length. We compute the STFT with window size $n_{\mathrm{FFT}} = 2048$ and hop length $h = 209$, yielding $X \in \mathbb{C}^{F \times T}$. A fixed mel filterbank $\Phi \in \mathbb{R}^{128 \times F}$ projects the power spectrogram:

$$M = \Phi\,|X|^2, \quad M \in \mathbb{R}^{128 \times T}, \tag{1}$$

which is converted to decibels and min–max normalized to $[0, 1]$.

**Augmentation.** At each training iteration, we apply:

- **Mixup** Zhang et al. (2018): $\widetilde{M} = \lambda M + (1-\lambda)M'$, $\widetilde{y} = \min(y + y', 1)$, with $\lambda \sim \mathrm{Beta}(0.5, 0.5)$.
- **SpecAugment** Park et al. (2019): mask up to 15% of mel bins and 20% of time frames.

**Channel Replication and Standardization.** Each $128 \times T$ spectrogram is stacked to 3 channels to match a CNN14 backbone. Finally,

$$S = \frac{M_{\mathrm{rep}} - \mu}{\sigma}, \tag{2}$$

using dataset-specific $\mu, \sigma$. Figure 1 illustrates this pipeline, from the raw waveform to the log-mel spectrogram that the encoder consumes.

### 3.2 Encoder and Region Nodes

A CNN14 backbone from PANNs Kong et al. (2020), initialized from AudioSet-pretrained weights, encodes the spectrogram. Following standard few-shot practice, the earliest convolutional blocks (which capture generic low-level acoustics) are frozen, and the later blocks together with the graph and readout are adapted. Crucially, rather than averaging frequency into a single temporal chain, we read out the backbone's final feature map as a two-dimensional grid of *region nodes*: we pool to $f$ frequency bands and $T'$ time steps, giving $N = f \cdot T'$ nodes, each a $d$-dimensional embedding,

$$R = \mathrm{Enc}(S) \in \mathbb{R}^{N \times d}, \qquad f{=}4,\ d{=}256, \tag{3}$$

with $N{=}132$ in our configuration. Retaining the time–frequency grid—instead of collapsing frequency, as a purely temporal encoder would—is what allows both the graph and the readout to reason over localized spectro-temporal regions.

### 3.3 Graph-Attention Stage

We treat the $N$ region nodes as a graph whose structure is inferred from the features themselves. For each node pair $(i, j)$ we compute cosine similarity on the encoder features $R$ (not the projected features) and retain the top-$k$ most similar neighbors per node plus self-loops, yielding a sparse, *directed* adjacency $A \in \{0,1\}^{N \times N}$ (we do not symmetrize). A single multi-head Graph Attention layer Veličković et al. (2018) then updates each node from its neighbors,

$$r'_i = \Big\|_{m=1}^{H} \sum_{j:\, A_{ij}=1} \alpha_{ij}^{(m)}\, W^{(m)} r_j, \tag{4}$$

where $\alpha_{ij}^{(m)}$ follow the standard GAT parameterization—a LeakyReLU-activated single-layer scorer on the concatenation of the transformed endpoints, normalized over each node's neighbors—and $H$ heads are concatenated ($\|$). A residual connection and layer normalization, $r'_i \leftarrow \mathrm{LN}(r'_i + r_i)$, preserve node identity and guard against over-smoothing. This stage lets non-local regions exchange context before matching. As our ablations show (and as we verify with a graph diagnostic), the learned graph is well-conditioned—edges connect regions that are distant in time rather than merely adjacent, and node variance is essentially preserved through the layer—yet the stage contributes only modestly to accuracy, motivating the readout below as the primary source of RAFTa's gains.

### 3.4 Region-Prototype Readout

The readout is the central component of RAFTa. Let $P = \{p_c\}_{c=1}^{C}$ be the class prototypes for the current episode (formed as described below), stacked as $P \in \mathbb{R}^{C \times d}$, and let $R' \in \mathbb{R}^{N \times d}$ be the refined region nodes of a query clip. Rather than pooling $R'$ to one vector, we let each prototype attend over the query's regions with multi-head cross-attention. With prototypes as queries and region nodes as keys and values,

$$Q = PW_Q, \quad K = R'W_K, \quad V = R'W_V, \tag{5}$$

$$\mathrm{Att} = \mathrm{softmax}\Big(\tfrac{QK^\top}{\sqrt{d/H}}\Big) \in \mathbb{R}^{C \times N}, \tag{6}$$

$$\tilde{p}_c = \big(\mathrm{Att}\, V\big)_c\, W_O \in \mathbb{R}^d, \tag{7}$$

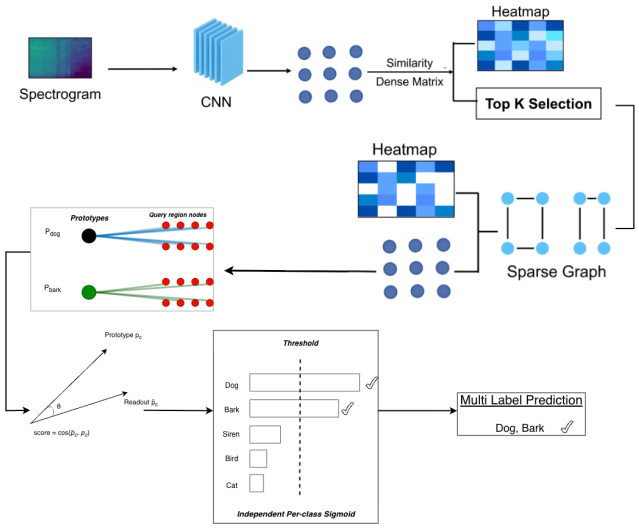


Figure 2: Overall architecture. A spectrogram is encoded into a grid of time–frequency region nodes, refined by a graph-attention layer; class prototypes formed from the support set then attend over the query's region nodes (the region-prototype readout) to produce per-class logits.

The projections $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ are split into $H$ heads of dimension $d/H$; attention is computed per head with the $1/\sqrt{d/H}$ scaling above, and $W_O \in \mathbb{R}^{d\times d}$ concatenates and projects the $H$ heads back to $d$ (standard multi-head attention, $H$=4, $d$=256). Thus $\tilde{p}_c$ aggregates the spectro-temporal evidence in the query most relevant to class $c$. The class logit is a temperature-scaled cosine match between this class-conditioned readout and the *original* prototype $p_c$ (the mean-support embedding, before projection; gradients are not stopped through it),

$$\ell_c(q) = \frac{\cos\big(\tilde{p}_c,\, p_c\big)}{\tau}, \qquad \hat{y}_c = \sigma\big(\ell_c(q)\big). \tag{8}$$

Three properties matter. **(i)** Attention and projections lie on the forward path, so gradients reach the encoder—the readout *shapes* the representation, not merely the final comparison. **(ii)** Each class is scored by an independent sigmoid (no softmax coupling classes), so the readout is natively multi-label. **(iii)** *Why this calibrates.* A pooled baseline forces every class to read the *same* clip vector, so a query that broadly resembles the episode scores high on *all* classes—the over-prediction we observe. Class-conditioned attention instead lets each prototype gather only its own evidence, so a class scores high only when *its* regions are present. This widens the margin between present and absent classes, which shrinks the predicted label-cardinality and, empirically, the ECE.

### 3.5 Prototypes and Protocol

In each 5-way 5-shot episode, the prototype $p_c$ is the mean support representation of class $c$. We follow the standard prototypical, pretrain-then-meta-train recipe Snell et al. (2017): the encoder is first pretrained by supervised multi-label classification, then meta-trained on episodes with the region-prototype readout in place. To make this a genuine few-shot test, we split the *class indices* into disjoint base and novel sets: pretraining and meta-training use only base classes, while every evaluation episode draws its classes exclusively from the novel set. Because clips are multi-label, a training clip may still *contain* a novel sound, but the model is never *supervised* on a novel label—so evaluation measures generalization to genuinely unseen classes. Our readout is motivated by compositional multi-label prototype methods such as LC-Protonets Papaioannou et al. (2024), but instead of enumerating label-subset prototypes we let single-class prototypes attend over the query's regions; we therefore reference that line of work as inspiration rather than reproducing its music-tagging system.

# 4 ABLATIONS

## 4.1 COMPUTATIONAL COST

RAFTa adds the graph-attention and region-prototype modules over a shared CNN14 encoder at negligible cost (Table 1): only $\sim$0.13M extra parameters and under 0.1 GFLOPs relative to the CNN+ProtoNet baseline. Parameters and FLOPs are therefore essentially identical across the two, so RAFTa's calibration and decision-quality gains come at no meaningful compute overhead.

| | **RAFTa** | **CNN+ProtoNet** |
|---|---|---|
| Params (M) | 76.19 | 76.06 |
| FLOPs (G) | 42.81 | 42.79 |
| Time (ms/clip) | 25.14 | 23.44 |
| Node dim | 256 | 256 |

Table 1: Computational cost on a single GPU. RAFTa's added modules are lightweight relative to the shared CNN14 backbone.

## 4.2 EFFECT OF SHOT COUNT

A known failure of pooled few-shot taggers is that over-prediction *worsens* with more support examples: additional shots enlarge and blur the prototypes, so the model predicts ever more labels and accuracy degrades. Table 2 sweeps the support size for RAFTa (FSD50K, random protocol, 5-way, 100 episodes). Two things stand out. First, mAP improves from 1- to 5-shot and then plateaus (the slight 10-shot dip is within the confidence interval). Second, and more importantly, the predicted label-cardinality stays controlled across all shot counts ($3.4$–$3.8$), rather than drifting toward the 5-way maximum as shots increase. The region-prototype readout thus does not inherit the over-prediction-with-shots pathology: more support data sharpens matching without inflating the predicted label set, and ECE stays stable ($0.36$–$0.40$).

| **Shot** | **mAP** | **F1** | **ECE$\downarrow$** | **Card.** |
|---|---|---|---|---|
| 1 | 0.642 | 0.458 | **0.360** | **3.41** |
| 3 | 0.669 | 0.452 | 0.386 | 3.76 |
| 5 | **0.717** | **0.467** | 0.384 | 3.68 |
| 10 | 0.707 | 0.453 | 0.395 | 3.74 |

Table 2: Effect of support-set size (RAFTa, FSD50K, random protocol, 5-way, 100 episodes). Predicted cardinality stays near the true value ($\approx$ 1–2) across shot counts—no over-prediction blow-up—while mAP peaks at 5-shot.

# 5 EXPERIMENTS

## 5.1 DATASETS

We evaluate **RAFTa** on two widely used audio tagging benchmarks:

- **FSD50K** Fonseca et al. (2021): A 200-class dataset containing 37,134 training, 4,170 validation, and 10,231 test clips. Clips vary in duration from 0.3 to 30 seconds and may contain multiple overlapping sound events.
- **AudioSet–Balanced** Gemmeke et al. (2017): A curated 527-label subset of AudioSet with 20,550 training and 1,887 evaluation clips. Each 10-second clip may include multiple co-occurring acoustic events.

### 5.2 Implementation Details

**Feature Extraction and Augmentation.** All audio clips are converted to mono, resampled to 22.05 kHz, and padded or tiled to a fixed 10-second duration. We compute 128-bin log-Mel spectrograms using an STFT window of 2048 and hop size of 209. Spectrograms are replicated across 3 channels and normalized using dataset-specific mean and standard deviation. We apply Mixup (Beta distribution, $\alpha = 0.5$) and SpecAugment with 15% frequency masking and 20% time masking during training.

**Model.** The encoder is CNN14 Kong et al. (2020), initialized from AudioSet-pretrained weights; the first two convolutional blocks are frozen and the remainder adapted. Its final feature map is read out as a $4 \times 33$ grid of $N{=}132$ region nodes of dimension $d{=}256$. The graph-attention stage is a single layer with $H{=}8$ heads whose outputs are concatenated back to 256, cosine-similarity top-$k$ neighbours ($k{=}10$), a residual connection, and layer normalization ($\alpha{=}0.3$ LeakyReLU in the attention). The region-prototype readout uses 4 cross-attention heads. All feature dimensions are 256 throughout; the temperature $\tau$ is fixed on base-class validation episodes.

**Training.** We use a pretrain-then-meta-train recipe. The encoder is first pretrained by supervised multi-label classification on base classes (Adam Kingma & Ba (2015), learning rate $3\mathrm{e}{-4}$, weight decay $1\mathrm{e}{-4}$, with a class-balanced loss to counter the heavy negative skew), then meta-trained on 5-way 5-shot 5-query episodes with the region-prototype readout in the loop. Every baseline shares this same AudioSet-pretrained CNN14 backbone, so comparisons isolate the decoder rather than pretraining data. Training runs on a single GPU.

### 5.3 Protocol and Metrics

We split class *indices* into disjoint base and novel sets (80% base, fixed seed). Pretraining and meta-training use base classes only; every test episode draws its classes exclusively from the novel set, so we measure generalization to genuinely unseen classes—a true inductive few-shot protocol. We evaluate two episodic settings, each over the *same* 200 fixed 5-way 5-shot 5-query episodes for every model: a **random** protocol (episodes sampled uniformly; queries are predominantly single-label) and a **multi-label** protocol (episodes seeded on co-occurring classes so that queries carry $\geq 2$ active labels). Reporting both prevents a method from looking good only where the task collapses to single-label.

We report four quantities, each as a mean over the 200 episodes with a 95% confidence interval. **mAP** is computed per episode over that episode's classes, directly comparable to a **chance floor** obtained by scoring the same episodes with random values. **Macro-F1** uses a fixed decision threshold of 0.5, shared across all methods (never tuned per model). **Expected Calibration Error (ECE)** uses 10 uniform-width confidence bins and is computed jointly over all (class, query) prediction pairs within each episode's class set. **Predicted label-cardinality** is the mean number of labels above threshold per query, against the true cardinality. The temperature $\tau{=}0.4$ is shared across all models (fixed once on base-class validation episodes), so calibration differences reflect the decoder rather than per-model $\tau$ tuning.

### 5.4 Baselines

The metric-based baselines share the AudioSet-pretrained CNN14 encoder and differ only in the decoder, so each comparison isolates the readout rather than the features. **CNN+ProtoNet**: single-class prototypes with pooled cosine matching and independent per-class sigmoids Snell et al. (2017). **LC-Protonets**: compositional label-subset prototypes (up to size 3) on the same encoder, following Papaioannou et al. (2024) (a re-implementation of the decoder, not their music-tagging system). **CNN+ProtoNet+TS**: the CNN+ProtoNet baseline with post-hoc temperature scaling Guo et al. (2017), its scaling temperature fit on base-class validation—this isolates whether the calibration gain requires the architecture or can be obtained by simple rescaling. **AST+ProtoNet**: a frozen AudioSet AST Gong et al. (2021) feature extractor with pooled prototype matching—the requested transformer baseline. **RAFTa**: the region-prototype readout, evaluated with and without the graph-attention stage (*–GAT*) to isolate its contribution.

| Data | Model | mAP | F1 | ECE↓ | Card. |
|---|---|---|---|---|---|
| FSD50K | Chance | 0.598 ±.025 | – | – | – |
| | CNN+ProtoNet | 0.724 ±.020 | 0.628 | 0.306 | 4.6 |
| | LC-Protonets | **0.784** ±.020 | 0.685 | 0.317 | 4.5 |
| | AST+ProtoNet | 0.718 ±.024 | 0.676 | 0.519 | 5.0 |
| | RAFTa (–GAT) | 0.772 ±.022 | **0.690** | **0.256** | 4.3 |
| | RAFTa | 0.772 ±.022 | 0.690 | 0.271 | **4.2** |
| AudioSet | Chance | 0.598 ±.025 | – | – | – |
| | CNN+ProtoNet | 0.757 ±.020 | 0.636 | 0.309 | 4.3 |
| | LC-Protonets | 0.747 ±.020 | 0.636 | 0.328 | 4.4 |
| | AST+ProtoNet | 0.673 ±.022 | 0.622 | 0.519 | 5.0 |
| | RAFTa (–GAT) | 0.777 ±.020 | **0.644** | **0.278** | **4.1** |
| | RAFTa | **0.778** ±.020 | 0.639 | 0.280 | **4.1** |

Table 3: Multi-label protocol (5-way 5-shot, 200 episodes; mAP as mean ± 95% CI). Lower ECE and Card. nearer the true cardinality ($\approx 2$) are better; best per column in bold.

## 6 RESULTS

Tables 3 and 4 report every model on both datasets and both protocols, on identical episodes; mAP is a mean over episodes with a 95% confidence interval. *Card.* is the mean predicted labels per query (true cardinality $\approx 2$ multi-label, $\approx 1$ single-label); over-prediction shows up as inflated *Card.* Three findings recur across all four settings.

**Ranking (mAP) does not separate the methods; decisions do.** Under the multi-label protocol the chance floor is high (mAP $\approx 0.60$, because co-occurring episodes have high per-class prevalence), and every metric-based model—and even the frozen AST—clusters in a narrow band above it. Under the single-label protocol the best ranker is dataset-dependent (LC-Protonets on FSD50K, plain CNN+ProtoNet on AudioSet). We therefore make no ranking claim. The methods separate instead on *decision quality*: RAFTa is the best-calibrated model in *every* setting, with the lowest ECE among all baselines and the predicted cardinality closest to the truth, and equal-or-better macro-F1. Its no-GAT variant reaches marginally lower ECE in three settings, consistent with the graph contributing little. To test whether this is merely a rescaling effect, we compare against post-hoc temperature scaling of CNN+ProtoNet (**supplementary material**); the architectural readout, not logit rescaling, is what closes the calibration gap.

**The transformer is the worst-calibrated.** Frozen AST predicts *all five* candidate labels for every query (Card. $= 5.0$ regardless of the true count), reaching high recall at the expense of precision and an ECE of $0.52$–$0.71$—roughly double any metric-based model. Over-prediction is thus not a small-CNN artifact but a general failure of few-shot multi-label audio, most acute for the highest-capacity backbone. Reliability diagrams for every model are provided in the supplementary material.

**The graph is well-conditioned but contributes little.** GAT vs. no-GAT gives a small, consistent single-label mAP gain ($+0.016$ FSD, $+0.010$ AudioSet) and is neutral on multi-label; the diagnostic (see *Graph Diagnostic*) shows non-local edges and no over-smoothing. The readout, not the graph, drives RAFTa's behavior. Per-class AUPRC rankings (easiest and hardest novel classes) and full per-cardinality breakdowns are reported in the supplementary material.

### 6.1 GRAPH DIAGNOSTIC

To check whether the graph-attention stage performs genuine non-local reasoning rather than local smoothing, we inspect the learned graph on test episodes. Edges connect regions that are, on average, $7.5$ time-frames apart (a value of $1.0$ would indicate purely adjacent connections), and node variance is essentially unchanged through the layer ($0.43 \rightarrow 0.43$), indicating no over-smoothing. The stage is thus well-conditioned; its modest effect on accuracy reflects the task, not a degenerate graph. We visualize the learned graph edges and the region-prototype attention maps on example spectrograms in the supplementary material.

| Data | Model | mAP | F1 | ECE↓ | Card. |
|---|---|---|---|---|---|
| FSD50K | Chance | 0.287 ±.007 | – | – | – |
| | CNN+ProtoNet | 0.644 ±.013 | 0.384 | 0.450 | 4.2 |
| | LC-Protonets | **0.756** ±.014 | 0.433 | 0.428 | 3.9 |
| | AST+ProtoNet | 0.524 ±.012 | 0.351 | 0.691 | 5.0 |
| | RAFTa (–GAT) | 0.705 ±.015 | 0.457 | **0.380** | 3.7 |
| | RAFTa | 0.721 ±.015 | **0.468** | 0.382 | **3.6** |
| AudioSet | Chance | 0.287 ±.007 | – | – | – |
| | CNN+ProtoNet | **0.716** ±.013 | 0.411 | 0.451 | 3.9 |
| | LC-Protonets | 0.713 ±.013 | 0.341 | 0.544 | 4.9 |
| | AST+ProtoNet | 0.492 ±.012 | 0.337 | 0.709 | 5.0 |
| | RAFTa (–GAT) | 0.675 ±.016 | 0.444 | 0.400 | 3.6 |
| | RAFTa | 0.685 ±.016 | **0.450** | **0.393** | **3.5** |

Table 4: Random protocol (5-way 5-shot, 200 episodes; mAP as mean ± 95% CI). True cardinality $\approx 1$; best per column in bold.

## 7 Limitations

Several limitations remain. First, the region-prototype readout *matches* rather than exceeds the pooled baselines on ranking (mAP); its advantage is calibration and decision quality, and we do not claim a ranking improvement. Second, we study a single readout design and do not fine-tune the full backbone, both for compute reasons; a jointly fine-tuned encoder could change the ranking picture. Third, the AST baseline is used frozen with spectrograms resampled to its expected configuration, making it a conservative estimate of transformer performance; a fully adapted transformer might rank higher, though our claim concerns calibration, where it fails clearly. Finally, we evaluate on two curated datasets; robustness to out-of-domain audio, longer-tailed distributions, and real-world noise remains open.

## 8 Conclusion

We revisited few-shot multi-label audio tagging and found that strong prototypical models, despite good ranking, make poorly calibrated decisions, over-predicting to the point of assigning nearly every candidate label, a failure that persists across convolutional, compositional, and transformer backbones. We introduced RAFTa, whose region-prototype readout lets class prototypes attend over a query's time–frequency regions, and showed that it is competitive with baseline ranking while consistently improving calibration, macro-F1, and label-cardinality on FSD50K and AudioSet-Balanced. Alongside a standardized benchmark with disjoint base/novel splits and explicit calibration analysis, we hope this reframes few-shot multi-label audio tagging around decision quality, not ranking alone.

## 9 Broader Impact

Few-shot audio tagging is aimed at low-resource settings—bioacoustic and biodiversity monitoring, industrial fault detection, and on-device personalization—where labels are scarce and mislabeling has real cost. Our central point is that a well-*ranked* model can still make badly *calibrated* decisions, over-predicting labels; deploying such a model (e.g., flagging every species as present) is misleading. By foregrounding calibration and label-cardinality, we hope to encourage evaluation that reflects deployment risk rather than ranking alone. Risks are those common to audio recognition: acoustic surveillance and biased performance across under-represented sound classes or recording conditions; our long-tail per-class analysis is a step toward surfacing the latter, but real-world use warrants domain-specific auditing.

## REFERENCES


Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Kuan-Hung Cheng, Shih-Yuan Chou, and Yi-Hsuan Yang. Multi-label few-shot learning for sound event recognition. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2019.

Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: Spatially-aware few-shot transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 760–764, 2021.

Víctor García Satorras and Joan Bruna. Few-shot learning with graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Jort F Gemmeke, Daniel P W Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

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

Jinhua Liang, Huy Phan, and Emmanouil Benetos. Learning from taxonomy: Multi-label few-shot classification for everyday sound recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 771–775, 2024.

Inês Nolasco, Shubhr Singh, Veronica Morfi, Vincent Lostanlen, Ariana Strandburg-Peshkin, Hanna Pamuła, Helen Whitehead, Ivan Kiskin, Michael Emmerson, and Dan Stowell. Few-shot bioacoustic event detection at the DCASE 2022 challenge. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2022.

Charilaos Papaioannou, Emmanouil Benetos, and Alexandros Potamianos. LC-Protonets: Multi-label few-shot learning for world music audio tagging. *arXiv preprint arXiv:2409.11264*, 2024.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, 2019.

Shubhr Singh, Christian J Steinmetz, Emmanouil Benetos, Huy Phan, and Dan Stowell. ATGNN: Audio tagging graph neural network. *IEEE Signal Processing Letters*, 31:825–829, 2024.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4077–4087, 2017.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Yushi Wang, Shuai Li, and Florian Metze. Few-shot learning with graph neural networks for audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 331–335, 2020.

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