CALM: Class-Conditional Sparse Attention Vectors for Large Audio-Language Models

Videet Mehta
MIT CSAIL
mvideet@mit.edu

Liming Wang
MIT CSAIL
limingw@mit.edu

Hilde Kuehne
Tuebingen AI Center
MIT-IBM Watson AI Lab
kuehne@cs.uni-bonn.de

Rogerio Feris
MIT-IBM Watson AI Lab
rsferis@us.ibm.com

James R. Glass
MIT CSAIL
glass@mit.edu

M. Jehanzeb Mirza
MIT CSAIL
jmirza@mit.edu

Abstract

Large audio-language models (LALMs) show strong zero-shot performance on tasks such as audio question answering, but still lag behind specialized systems on discriminative tasks such as audio classification. Recent work shows that sparse subsets of attention heads can serve as discriminative feature extractors for few-shot classification, but existing methods weight all selected heads equally. We propose **C**lass-Conditional Sparse **A**ttention Vectors for Large Audio-**L**anguage **M**odels (CALM), a few-shot classifier that learns class-conditional importance weights over attention heads. By treating heads as class-specific experts and aggregating them with reliability-weighted soft voting, CALM allows different classes to rely on different subsets of heads. Across diverse few-shot audio, audiovisual, visual, and spoofing benchmarks, CALM consistently outperforms state-of-the-art uniform-voting baselines, with gains of up to 14.52, 1.53, and 8.35 absolute points on audio, audiovisual, and spoofing tasks, respectively. [1]

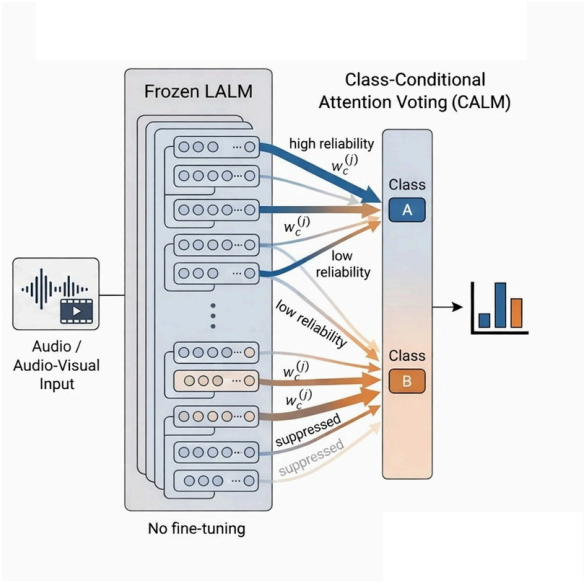


Figure 1: **Fine-tuning-free audio classification with CALM**. Given a frozen audio-language model, we extract per-head last-token hidden-state representations and compute class centroids. Each attention head produces cosine-similarity scores to all class centroids. For every head-class pair, we estimate a reliability score based on the margin to the next most confident class, which is visualized by the arrow width and color in the figure. Highly reliable heads contribute strongly (thick, saturated arrows), while unreliable heads are suppressed by lower weights. At inference time, class predictions are obtained via reliability-weighted soft voting across all heads. This yields accurate classification without task-specific fine-tuning.

1 Introduction

Large audio-language models (LALMs), such as Qwen2-Audio (Chu et al., 2024) and Qwen2.5-Omni (Xu et al., 2025), have shown strong zero-shot and few-shot performance across a wide range of audio understanding tasks (Chu et al., 2024; Xu et al., 2025). Trained on large-scale paired audio-text data, these models generalize well without task-specific fine-tuning (Radford et al., 2022; Huang et al., 2023; Lyu et al., 2023).

Despite these strengths, repurposing LALMs for discriminative audio classification remains challenging. Specialized audio classifiers often outperform prompt-based approaches, but they require substantial labeled data and task-specific training pipelines (Gao et al., 2022; Xie and Virtanen, 2019; Elizalde et al., 2022; Guzhov et al., 2022). In contrast, directly leveraging frozen LALMs for classification offers reduced training cost (Hanif et al., 2024), lower sample complexity in few-shot settings, and a unified model that supports both generative and discriminative use without retraining (Yang et al., 2024).

Recent work has shown that frozen generative models can be turned into strong discriminative

[1] Code can be found at https://github.com/mvideet/CALM

classifiers by extracting intermediate representations rather than relying on prompt-based predictions (Mitra et al., 2025; Huang et al., 2024; Zhang et al., 2024; Hendel et al., 2023; Hojel et al., 2024; Todd et al., 2024). In particular, Sparse Attention Vectors (SAV) (Mitra et al., 2025) show that a small subset of attention heads in a frozen audio-language model can serve as effective feature extractors for few-shot classification. The method selects the most discriminative heads and aggregates their predictions through majority voting, substantially narrowing the gap between generative models and specialized discriminative systems without requiring fine-tuning on labeled data.

In this paper, we extend this line of work to audio, audiovisual, and visual classification. Compared with vision or text, audio classification presents unique challenges, including multiple overlapping acoustic events (Dogan et al., 2024; Sridhar et al., 2024), higher intra-class and inter-class variability, and temporal ambiguity in semantic cues. These characteristics make it unlikely that a single uniform aggregation strategy over attention heads is optimal across all classes. While effective, SAV relies on a strong simplifying assumption: all selected attention heads are treated as *equally reliable* across all classes. By contrast, prior work on interpretability and representation learning shows that attention heads exhibit *functional specialization*, often responding differently to different semantic patterns or sensory stimuli (Basile et al., 2025; Jo and Myaeng, 2020; Ma et al., 2025; Wang et al., 2025). This exposes a key limitation of uniform voting: weak or noisy heads can still influence predictions for classes where they are not discriminative (Qian et al., 2025). To address this limitation, we propose **C**lass-Conditional Sparse **A**ttention Vectors for Large Audio-**L**anguage **M**odels (CALM), a principled extension of SAV that models attention heads as *class-conditional experts* rather than uniformly reliable voters. CALM estimates a class-specific reliability score for each selected head from few-shot data and uses these scores in a reliability-weighted voting scheme at inference time. Figure 1 shows the overall pipeline. Because CALM operates directly on a frozen LALM, it introduces only minimal computational overhead. Our contributions are threefold:

1. We introduce CALM, a probabilistically weighted framework that models attention heads as class-conditional experts.
2. We present the first application of attention vectors as discriminative classifiers to audio, audiovisual, and visual classification, despite modality-specific challenges.
3. We demonstrate improvements over uniform-voting SAV baselines across audio, audiovisual, visual, and spoofing benchmarks, including ESC-50, VGGSound, AudioSet, EuroSAT, Oxford-IIIT Pets, and LA-Spoof.

2 Related Work

Generative Models for Audio Classification. Recent work adapts large generative audio or multimodal models to classification through prompting, in-context demonstrations, or label-likelihood scoring (Taylor and Mack, 2025; Gong et al., 2024; Brown et al., 2020; Cheng et al., 2019; Kumar et al., 2023; Hendrycks et al., 2021). While appealing because they reuse a single frozen model, these approaches are often sensitive to prompt design and can lag behind specialized discriminative systems or adapted audio-language models (Olvera et al., 2024; Sahoo et al., 2025; Salinas and Morstatter, 2024; Choudhary et al., 2022; Deshmukh et al., 2024). Other methods improve classification through instruction tuning, preference optimization, or test-time adaptation, but they require additional optimization or supervision and therefore move away from the fine-tuning-free setting we target (Choi et al., 2025; Bucher and Martini, 2024; Ouyang et al., 2022; Sun et al., 2020; Zhang et al., 2025a).

Internal Representation of LALMs. A complementary direction uses frozen language or multimodal models as feature extractors rather than relying on decoded outputs (Belinkov, 2021; Hendel et al., 2023; Huang et al., 2024; Taylor and Mack, 2025). In audio, this includes representation-based classifiers formed from audio-text similarity in a shared embedding space, as well as methods that use intermediate hidden states for downstream prediction. Interpretability studies further show that attention heads exhibit functional specialization, suggesting that different heads capture different semantic factors (Clark et al., 2019; Voita et al., 2019; Basile et al., 2025; Jo and Myaeng, 2020). SAV (Mitra et al., 2025) turns this observation into a classifier by selecting a sparse set of discriminative heads and aggregating them with uniform

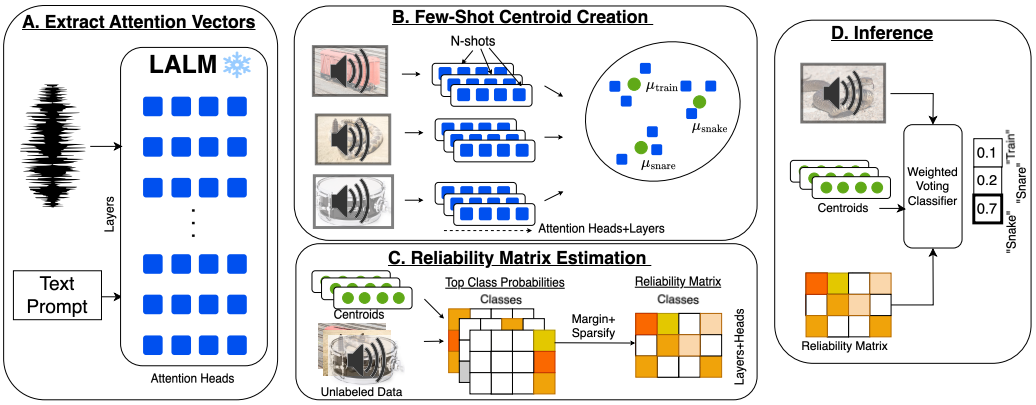


Figure 2: **Overview of CALM**: We extract attention-head vectors from a frozen LALM given an audio input and prompt. Then we build class centroids from few-shot training examples and estimate a class-conditional head reliability matrix using margin-based confidence and sparsifying to k head experts per class. At inference, CALM combines centroid similarities with these reliability weights to perform weighted voting over heads. This allows for fine-tuning-free classification.

voting. Our work builds directly on SAV, but replaces uniform aggregation with class-conditional reliability weighting so that different classes can rely on different expert heads.

3 Methods

We consider a C-way classification problem with a few-shot training set $\mathcal{D}_{\text{train}} = \{(x_i, c_i)\}_{i=1}^{N}$ and a test set $\mathcal{D}_{\text{test}} = \{(x_i, c_i)\}_{i=1}^{N_{\text{test}}}$, where $c_i \in \mathcal{C} = \{1, \ldots, C\}$ and n_c denotes the number of training examples for class c. Given an input x, we prompt a frozen audio-language or multimodal language model and extract the last-token representation from each attention head $j \in \{1, \ldots, K\}$, denoted $h_j(x) \in \mathbb{R}^d$. We then build a fine-tuning-free classifier by computing class centroids in each head space and aggregating head-level evidence. Figure 2 provides an overview.

We first briefly recap SAV (Mitra et al., 2025), then introduce CALM, our class-conditional weighting scheme. We finally describe a simple pseudo-labeling extension for unlabeled data.

3.1 Recap: Sparse Attention Vectors

The SAV classifier (Mitra et al., 2025) forms a nearest-centroid classifier in each attention head and aggregates the top-k heads with uniform voting.

Per-head Centroids. For each class c and head j, SAV computes the class centroid

$$\mu_c^{(j)} = \frac{1}{n_c} \sum_{i:c_i=c} h_j(x_i). \tag{1}$$

It scores a query x against class c in head j using cosine similarity

$$s_j(x, c) = \frac{h_j(x)^\top \mu_c^{(j)}}{\|h_j(x)\|_2 \, \|\mu_c^{(j)}\|_2}. \tag{2}$$

Top-k head selection. Each head is ranked by its nearest-centroid classification accuracy on $\mathcal{D}_{\text{train}}$. Concretely, head j predicts

$$\hat{y}^{(j)}(x) = \arg\max_{c \in \mathcal{C}} \; s_j(x, c), \tag{3}$$

and the top-k heads define $\mathcal{H}_{\text{SAV}}$.

Majority Voting (Uniform Aggregation). At inference time, SAVs aggregates the selected heads by uniform voting:

$$\hat{y}(x_{\text{test}}) = \arg\max_{c \in \mathcal{C}} \sum_{j \in \mathcal{H}_{\text{SAV}}} \mathbf{1}\left[\hat{y}^{(j)}(x_{\text{test}}) = c\right]. \tag{4}$$

This uniform voting scheme assumes that all selected heads are equally reliable across all classes. CALM removes this assumption.

3.2 CALM

CALM keeps the same per-head centroids as SAV, but replaces hard uniform voting with weighted aggregation of per-head posteriors. The key idea is to treat attention heads as *class-specific experts*, so that classes can rely on different subsets of heads with different strengths.

Head Posteriors. Using the per-head class centroids from Eq. 1 and the similarity scores from

Eq. 2, we convert each head's score into a probability distribution over classes.

$$p_c^{(j)}(x) = \frac{\exp\left(s_j(x,c)/\tau_p\right)}{\sum_{c' \in \mathcal{C}} \exp\left(s_j(x,c')/\tau_p\right)}, \tag{5}$$

for some hyperparameter $\tau_p > 0$.

3.2.1 CALM-Global: Head Weighting

We first describe a simpler global-weighting variant, which assigns each head a single reliability shared across all classes. For a labeled example (x_i, c_i), we define the clamped margin

$$m^{(j)}(x,c) = \max\Big(0, p_c^{(j)}(x) - \max_{c' \neq c} \; p_{c'}^{(j)}(x)\Big). \tag{6}$$

The clamp ensures that incorrect or low-confidence cases contribute zero. We then average margins over all training examples to obtain a single global reliability per head:

$$r^{(j)} = \frac{1}{N} \sum_{i=1}^{N} m^{(j)}(x_i, c_i). \tag{7}$$

Retaining the top-k heads with the largest reliabilities gives $\mathcal{H}_G$, and we normalize their weights as

$$w^{(j)} = \begin{cases} \dfrac{\exp(r^{(j)}/\tau_w)}{\sum_{j' \in \mathcal{H}_G} \exp(r^{(j')}/\tau_w)}, & j \in \mathcal{H}_G, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

At test time, we aggregate per-head posteriors using these global weights:

$$p_c(x_{\text{test}}) = \sum_{j=1}^{K} w^{(j)} p_c^{(j)}(x_{\text{test}}), \tag{9}$$

$$\hat{y}(x_{\text{test}}) = \arg\max_c \; p_c(x_{\text{test}}). \tag{10}$$

This variant is useful for comparison, but it does not capture class-specific head specialization.

3.2.2 Class-Conditional Head Weighting

Our full CALM model assigns a separate reliability to each head-class pair. After computing the margins in Eq. (6), we average them class-wise:

$$r_c^{(j)} = \frac{1}{n_c} \sum_{i; c_i = c} m^{(j)}(x_i, c). \tag{11}$$

Let the top-k heads for class c be $\mathcal{H}_c$. We normalize the class-conditional reliabilities with a softmax:

Algorithm 1 CALM (class-conditional weighting)

Require: Frozen LALM, train set $\mathcal{D} = \{(x_i, y_i)\}$, num heads K, and sparsity k

Extract attention-head features $h_j(x_i)$ for all $j \in \{1, \ldots, K\}$
Compute class centroids $\mu_c^{(j)} = \frac{1}{n_c} \sum_{i:y_i=c} h_j(x_i)$
for each class c and head j **do**
  Compute posteriors $p_c^{(j)}(x_i)$ via Eq. 5
  Compute margins via Eq. 6
  $r_c^{(j)} \leftarrow \frac{1}{n_c} \sum_{i:y_i=c} m_i^{(j)}$
end for
Select top-k reliable heads per class and normalize weights $w_c^{(j)}$
Predict $\hat{y} = \arg\max_c \sum_j w_c^{(j)} p_c^{(j)}(x_{\text{test}})$

$$w_c^{(j)} = \begin{cases} \dfrac{\exp(r_c^{(j)}/\tau_w)}{\sum_{j' \in \mathcal{H}_c} \exp(r_c^{(j')}/\tau_w)}, & j \in \mathcal{H}_c, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

for some hyperparameter $\tau_w > 0$. This generalizes the shared top-k set and uniform voting used by SAV to a class-specific set of head experts.

$$p_c(x_{\text{test}}) = \sum_{j=1}^{K} w_c^{(j)} p_c^{(j)}(x_{\text{test}}), \tag{13}$$

and predict with the same argmax rule as in Eq. (10). Algorithm 1 summarizes the class-conditional CALM procedure.

3.3 Self-supervised CALM

To use CALM without ground-truth labels, we generate M stochastic predictions for each unlabeled example using decoding temperatures $\tau \sim \mathcal{U}(1, 2.5)$. We assign a pseudo-label by majority vote, retain only examples where at least 50% of rollouts agree, and then run SAV or CALM on the retained pseudo-labeled set $\tilde{\mathcal{D}}_{\text{train}} = \{(x_i, \tilde{c}_i)\}$.

4 Experiments and Results

We evaluate our method on a range of audio and audiovisual classification benchmarks. We first describe the implementation details, models, datasets, and evaluation protocols, and then analyze the impact of head sparsity and class-specific weighting through ablations.

Dataset	ZS	SAV	SAV+PL	CALM	CALM+PL
Qwen2-Audio					
ESC-50	73.75	99.00	96.50	**99.00**	98.75
VGGSound	72.94	85.64	85.65	**88.15**	87.67
AudioSet	46.59	44.88	49.06	**59.40**	51.80
LA-Spoof	16.19	**72.00**	70.03	61.10	22.73
Qwen2.5-Omni					
ESC-50	86.00	99.00	91.00	**99.25**	98.50
VGGSound	61.57	87.11	81.78	**87.85**	86.28
AudioSet	61.52	69.52	63.54	**70.50**	66.70
VGGVideo	70.71	89.49	–	**91.02**	–
LA-Spoof	15.82	73.64	67.01	**81.99**	70.13
EuroSAT	59.67	68.72	58.10	**72.80**	62.70
Pets	96.05	98.27	97.92	**99.23**	97.16

Table 1: **Main classification results.** Accuracy (%) for zero-shot (ZS), SAV, and CALM with and without pseudo-labeling (PL), except LA-Spoof, for which we report Macro-F1 (%). Best results within each model block are shown in **bold**. "–" denotes settings that were not evaluated.

4.1 Implementation Details

All experiments are implemented in PyTorch and run on NVIDIA A6000 GPUs. We use frozen pretrained models from the Qwen family (Chu et al., 2024; Xu et al., 2025). *Qwen2-Audio* (Chu et al., 2024) is an audio-language model in which an audio encoder feeds into the language model. The language model consists of 32 Transformer layers with 32 attention heads per layer (1024 heads total), each with a hidden dimension of 128. We extract attention head activations from the language model layers rather than the audio encoder to capture cross-modal audio-text interactions. *Qwen2.5-Omni* (Xu et al., 2025) is a multimodal model supporting audio, image, video, and speech inputs. As with Qwen2-Audio, we extract attention activations exclusively from the language tower. To select τ_p and τ_w, we perform a hyperparameter sweep either on a small labeled validation split or purely from pseudo-labeled data (see Hyperparameter Selection without Labels). We sweep $\tau_p, \tau_w \in \{0.001, 0.03, 0.1, 0.5, 1, 2\}$. We use $\tau_p = 0.03$ and $\tau_w = 0.5$ for almost all datasets; for AudioSet, we use $\tau_p = 0.001$ and $\tau_w = 0.5$. For ESC-50, VGGSound, and AudioSet, an example prompt is: `What caption does the given audio belong to? A. washing_machine B. toilet_flush C. car_horn D. wind`. For spoofing detection, an example prompt is: `Is this audio bona fide or spoofed? A. bona_fide B. spoofed`. We use the same input-formatting logic for image inputs.

4.2 Datasets

We primarily evaluate our approach on audio classification benchmarks, and further extend it to audiovisual recognition and audio spoofing detection. Across all datasets, we follow the input formatting protocol of SAV (Mitra et al., 2025), casting each task into a multiple-choice question-answering formulation suitable for LALMs.

ESC-50 (Piczak, 2015) is a standard environmental sound classification benchmark consisting of 2,000 five-second audio clips across 50 classes. We sample class-balanced subsets according to the desired shot setting and frame the task as multiple-choice audio classification.

VGGSound (Chen et al., 2020) is a large-scale audiovisual dataset containing over 200k videos spanning 309 sound event classes. Compared to ESC-50, it exhibits substantially greater semantic and acoustic diversity. We adopt the same multiple-choice formulation and use both audio and video modalities for audiovisual experiments.

AudioSet (Gemmeke et al., 2017) is a large-scale ontology-based audiovisual dataset with 527 non-mutually exclusive sound event classes. Due to its multi-label nature, we decompose each audio clip into independent classification queries, one per ground-truth label, and report strict *exact-match* accuracy at the clip level.

ASVspoof (Wang et al., 2020) is used for spoofing detection. We study the LA (Logical Access) subset of ASVspoof 2019, which requires distinguishing bona fide speech from synthetic audio generated via speech synthesis and voice conversion methods. Given the class imbalance, we report

Macro-F1 as the primary evaluation metric (Zhang et al., 2025b).

EuroSAT (Helber et al., 2019) is a vision classification dataset for land-cover recognition. It contains 10 classes with approximately 2,000–3,000 images per class. We randomly sample a 20-shot support set and use the remaining images for evaluation.

Oxford-IIIT Pets (Parkhi et al., 2012) is another vision classification dataset, containing 37 cat and dog breeds. The main challenge is the subtle visual differences between breeds. We follow the same multiple-choice protocol as above, pairing the ground-truth breed with three randomly sampled distractor breeds.

UrbanSound8K (Salamon et al., 2014) contains 8,732 short urban-sound clips spanning 10 classes. We use it in our full-label-space evaluation in Section 4.4.2.

4.3 Evaluation Protocol

For the main results, each multiple-choice query contains the ground-truth label together with three distractor labels sampled from the remaining label space. We use identical candidate sets and test examples for ZS, SAV, and CALM within each run. CALM uses these examples both to construct class centroids and to estimate the class-conditional reliability matrix. Each example contributes $1/n_c$ to its corresponding class centroid; this reuse can introduce bias. Thus, we measure robustness across 10 independently sampled 20-shot support sets (Table 7) and also evaluate without distractors over the full label space in Section 4.4.2.

4.4 Results

4.4.1 Main Results

Table 1 summarizes the main comparison against zero-shot inference, SAV, and the pseudo-labeling variant from Section 3.3. Across both LALMs and most datasets, CALM improves over uniform SAV aggregation, with the largest gains appearing on the most challenging tasks. On ESC-50, both methods are near saturation, so the margin is small. On VGGSound, CALM improves over SAV by +2.51 points on Qwen2-Audio and +0.74 points on Qwen2.5-Omni. The strongest gain appears on AudioSet, where CALM improves over SAV by +14.52 points on Qwen2-Audio, suggesting that class-conditional weighting is especially useful when the label space is large and heterogeneous. Notably, zero-shot inference exceeds SAV on Qwen2-Audio AudioSet (46.59 vs. 44.88). Because AudioSet contains non-mutually-exclusive events, we hypothesize that uniform voting is vulnerable to heads that are poorly calibrated for particular classes. Meanwhile, CALM down-weights such heads and reaches 59.40%, improving by 14.52 points over SAV and 12.81 points over zero-shot inference.

The same trend extends beyond audio-only classification. On Qwen2.5-Omni, CALM improves over SAV on VGGVideo, EuroSAT, and Oxford-IIIT Pets, showing that class-conditional head weighting transfers not only to audiovisual recognition but also to pure image classification. LA-Spoof is the main exception: CALM improves over SAV for Qwen2.5-Omni but underperforms for Qwen2-Audio. This is consistent with the binary nature of spoofing detection, where estimating separate class-specific weights from few examples can overfit more easily than in large multi-class problems.

Pseudo-labeling is best viewed as a complementary self-training setting rather than the main source of CALM's gains. As shown in Table 1, CALM+PL is usually below fully supervised CALM, although it remains competitive on most settings where additional unlabeled data helps stabilize the centroids. Qwen2-Audio LA-Spoof is a particularly severe failure case; CALM+PL falls to 22.73 while the zero-shot baseline is only 16.19. This is consistent with pseudo-labeling propagating a poor initial prediction prior rather than correcting it.

Comparison to Fine-tuning. Table 2 compares CALM against standard fine-tuning techniques such as linear probing and LoRA, and additional few-shot baselines such as nearest neighbors using CLAP embeddings (Elizalde et al., 2022). CALM matches the best result on ESC-50, improves over both LoRA and SAV on VGGSound, and remains substantially stronger than CLAP. The main failure case is again LA-Spoof, where discriminative fine-tuning is more effective.

4.4.2 Beyond Multiple-Choice Evaluation

To test whether CALM's gains depend on sampled multiple-choice distractors, we remove the multiple-choice constraint and evaluate over the full label space of each dataset. For zero-shot inference, we generate free-form model predictions, embed the generated output and all ground-truth class labels using `sentence-transformers/all-MiniLM-L6-v2`,

Method	ESC-50	VGGSound	LA-Spoof F1
CLAP	95.50	35.12	59.79
Linear Probe	97.50	81.48	**85.13**
LoRA	98.00	87.23	69.68
SAV	99.00	85.64	72.00
CALM	**99.00**	**88.15**	61.10

Table 2: **Comparison to stronger baselines.** CALM remains competitive with linear probing and LoRA while outperforming SAV on VGGSound.

Dataset	ZS	SAV	CALM
UrbanSound8K	40.71	77.78	**78.88**
EuroSAT	22.80	64.23	**69.17**
VGGSound	12.94	48.40	**50.68**

Table 3: **Full-label-space evaluation** without sampled multiple-choice distractors. Accuracy (%).

and assign the class with the highest cosine similarity. SAV and CALM instead classify directly over all class centroids with no sampled distractors.

Table 3 shows that CALM continues to outperform SAV and ZS across all three datasets. These results suggest that the benefit of class-conditional reliability persists when classification is not limited to three sampled distractors and is performed over the complete label space.

4.4.3 Class-Specific Head Specialization

To better understand head specialization, Figure 3 visualizes per-class weight survival functions for representative VGGSound classes, while Figure 4 shows the most influential head for each class across the 309 VGGSound classes. Under SAV, all selected heads receive the same weight $\frac{1}{k}$, shown by the dashed baselines in Figure 3. In contrast, CALM learns class-conditional, non-uniform weightings: the survival curves differ substantially across classes, and several have heavy tails, indicating that a small subset of heads carries most of the evidence. Figure 4 further shows that the strongest heads concentrate in later layers, suggesting that LALMs contain both class-specific experts and a smaller set of broadly useful heads and that CALM exploits structured specialization rather than diffuse contributions from all heads.

4.4.4 Ablations

We next validate the two main design choices in CALM: class-conditional weighting and sparse reliability-based aggregation.

Algorithmic Ablations		
Model	**VGGSound**	**LA-Spoof**
Global		
Qwen2-Audio	87.02	**69.07**
Qwen2.5-Omni	86.84	**83.60**
Class-cond.		
Qwen2-Audio	**88.15**	61.10
Qwen2.5-Omni	**87.85**	81.99

Table 4: **Algorithmic ablations.** Comparison between global and class-conditional weighting on VGGSound and LA-Spoof.

Class-Conditional vs Global Weighting. Table 4 and the left panel of Figure 5 compare the full class-conditional method with the simpler global variant from Section 3.2.1. Local weighting is consistently better on VGGSound, where class-specific head specialization matters most, while global weighting is better on LA-Spoof. This reinforces the interpretation from Table 1: in a binary setting with limited support data, sharing one reliability score per head can regularize better than fitting separate class-wise weights. The pattern is also consistent with the head specialization analysis in Section 4.4.3: if different classes repeatedly route through different late-layer heads, then class-conditional weighting should outperform any scheme that forces all classes to share a single reliability profile.

Component Ablations. Table 5 isolates the effect of the two main ingredients in CALM. Removing L2 normalization consistently hurts performance, which suggests that raw head activations have scale variation that degrades centroid quality. Removing margin-based reliability also lowers accuracy, especially for Qwen2.5-Omni, showing that selecting heads by discriminability is important in addition to sparse aggregation. Together, these results show that both normalization and reliability weighting are necessary parts of the final method.

Shot and Sparsity Sensitivity. The right panel of Figure 5 shows CALM improves as support examples increase, with the largest gains appearing between 5 and 10 shots. The plot also shows that the best performance usually comes from intermediate sparsity rather than either extreme, supporting the central design intuition: too few heads discard useful evidence, while too many heads dilute the contribution of the most discriminative experts.

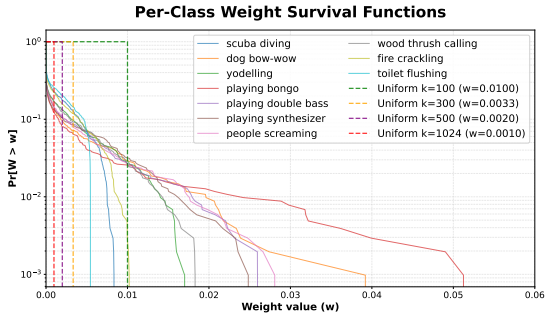


Figure 3: **Per-class weight survival functions for VGGSound.** Different classes concentrate weight on different subsets of attention heads. Dashed lines show the corresponding uniform-voting baselines for $k = 100, 300, 500$, and 1024 heads.

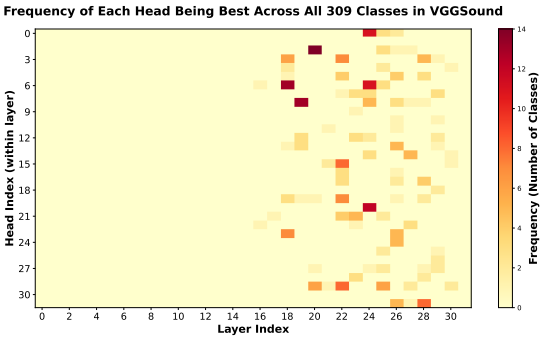
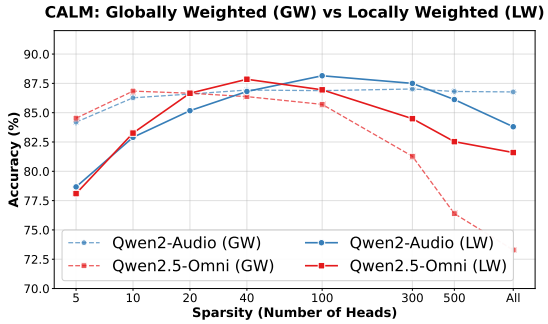


Figure 4: **Most influential attention heads across classes.** Heatmap of the most influential head (highest $w^{(j)}$) for each VGGSound class.



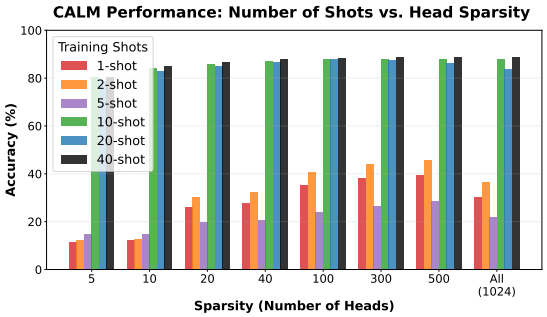


Figure 5: **CALM ablations. Left:** class-conditional local weighting consistently outperforms global weighting, especially when more heads are retained, supporting the hypothesis that attention heads specialize by class. **Right:** performance improves with more training shots and typically peaks at intermediate sparsity levels, showing that CALM benefits from both better centroid estimates and selective head pruning.

L2 Norm and Margin Weighting		
Method	**Qwen2-Audio**	**Qwen2.5-Omni**
SAV Baseline	85.64	87.11
CALM	**88.15**	**87.85**
– L2 Norm	87.03	87.00
– Margin	87.33	85.52

Table 5: **CALM component ablations.** Effect of removing L2 normalization or margin-based reliability weighting.

Random-head Counterfactual. Table 6 tests whether CALM's gains come from many heads or from structured selection and weighting. Random head selection with random weights degrades performance sharply in sparse regimes (27.35% vs. 78.10% at top-$k = 5$ on VGGSound), and CALM remains significantly better across all k.

top-k	**Random**	**CALM**
5	27.35	**78.10**
10	38.04	**83.26**
20	53.08	**86.66**
40	68.52	**87.85**
100	79.79	**86.95**

Table 6: **Random-head counterfactual.** Results with Qwen2.5-Omni on VGGSound.

Seed Stability. Table 7 measures robustness across 10 random 20-shot VGGSound splits on Qwen2-Audio. CALM consistently outperforms SAV across all tested top-k values with low variance, showing stable gains across support-set sampling and sparsity choices.

Hyperparameter Selection without Labels. Hyperparameters can be selected using either a small labeled validation split or zero-shot pseudo-labeled test data. On VGGSound, pseudo-label-selected hyperparameters match labeled-validation selection

top-k	SAV	CALM
20	85.22 ± 0.18	**86.38 ± 0.16**
40	85.20 ± 0.24	**87.46 ± 0.13**
100	85.37 ± 0.19	**88.22 ± 0.10**
300	85.65 ± 0.12	**88.45 ± 0.11**

Table 7: **Seed stability.** Results with Qwen2-Audio on VGGSound across 10 random 20-shot splits.

Backbone	SAV	CALM
Qwen2-VL-7B	71.03	**73.93**
Phi-4-Multimodal	47.99	**49.72**
Audio Flamingo 3	67.93	**69.14**

Table 8: **Cross-architecture transfer.** VGGSound results across additional multimodal and audio-language model backbones.

and yield 88.15% test accuracy, improving over the 72.94% zero-shot baseline by +15.21 points without labeled tuning.

Cross-architecture Transfer. To test whether CALM is specific to the Qwen family, we evaluate additional backbones. As shown in Table 8, CALM improves over SAV on Qwen2-VL-7B (71.03→73.93) (Wang et al., 2024), Phi-4-Multimodal (47.99→49.72) (Abouelenin et al., 2025), and Audio Flamingo 3 (67.93→69.14) (Goel et al., 2025). The consistent performance across model families suggests head specialization is a universal property across all backbones.

5 Conclusion

We presented **CALM**, a fine-tuning-free classifier that treats attention heads in frozen LALMs as class-conditional experts. Across audio, audiovisual, and visual benchmarks, CALM improves over uniform SAV voting. Ablations show that class-specific margin-based reliability and normalization drive these gains, indicating that reliability-aware sparse aggregation exposes discriminative structure already present in frozen multimodal models.

Limitations

CALM remains sensitive to label scarcity and label noise because class reliabilities are estimated from small support sets. We use the same support examples for centroid construction and reliability estimation. However, Table 7 shows low variance, meaning this problem might be mitigated.

Binary spoofing is a notable failure mode discussed. On Qwen2-Audio, class-conditional CALM underperforms SAV, while the global-weighting variant performs better, suggesting that estimating a separate reliability for every head-class pair can have high variance when there are few classes and few support examples. One extension could be $\tilde{r}_c^{(j)} = (1-\alpha)r_c^{(j)} + \alpha r_{\text{global}}^{(j)}$ where we interpolate between global reliability and class reliability. This regularization could be used in lower-data regimes. We leave this extension to future work.

Pseudo-labeling also depends on the quality of the underlying zero-shot prior, as seen in Qwen2-Audio for LA-Spoof. Stronger filtering may improve robustness to inaccuracies. CALM additionally introduces dataset-dependent temperature hyperparameters (τ_p, τ_w) and validation-dependent sparsity selection, although our experiments show that these can also be selected using pseudo-labeled data. Finally, while we evaluate across several backbones, broader evaluation across multiple tasks is needed to fully show generality.

5.1 Acknowledgements

This work was supported by IBM.

References

Abdelrahman Abouelenin and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. 2025. Head pursuit: Probing attention specialization in multimodal transformers. In *NeurIPS*.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Preprint*, arXiv:2102.12452.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned "small" llms (still) significantly outperform zero-shot generative ai models in text classification. *Preprint*, arXiv:2406.08660.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Kai-Hsiang Cheng, Szu-Yu Chou, and yi-hsuan Yang. 2019. Multi-label few-shot learning for sound event recognition. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5.

Youngwon Choi, Jaeyoon Jung, Hyeonyu Kim, Huu-Kim Nguyen, and Hwayeon Kim. 2025. Exploring fine-tuning of large audio language models for spoken language understanding under limited speech data. *Preprint*, arXiv:2509.15389.

Shwetank Choudhary, C R Karthik, Punuru Sri Lakshmi, and Sumit Kumar. 2022. Lean: Light and efficient audio classification network. In *2022 IEEE 19th India Council International Conference (INDICON)*, page 1–6. IEEE.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at? an analysis of bert's attention. *Preprint*, arXiv:1906.04341.

Soham Deshmukh, Rita Singh, and Bhiksha Raj. 2024. Domain adaptation for contrastive audio-language models. *Preprint*, arXiv:2402.09585.

Duygu Dogan, Huang Xie, Toni Heittola, and Tuomas Virtanen. 2024. Multi-label zero-shot audio classification with temporal attention. *Preprint*, arXiv:2409.00408.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. Clap: Learning audio concepts from natural language supervision. *Preprint*, arXiv:2206.04769.

Heting Gao, Junrui Ni, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson. 2022. Wavprompt: Towards few-shot spoken language understanding with frozen language models. In *Interspeech*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2024. Listen, think, and understand. In *ICLR*.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP*.

Asif Hanif, Maha Tufail Agro, Mohammad Areeb Qazi, and Hanan Aldarmaki. 2024. Palm: Few-shot prompt learning for audio language models. In *EMNLP*.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *Preprint*, arXiv:1709.00029.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of EMNLP*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.

Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. 2024. Finding visual task vectors. In *ECCV*.

Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. 2024. Multimodal task vectors enable many-shot multimodal in-context learning. In *NeurIPS*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. *Preprint*, arXiv:2304.12995.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417, Online. Association for Computational Linguistics.

Sachin Kumar, Chan Young Park, and Yulia Tsvetkov. 2023. Gen-z: Generative zero-shot text classification with contextualized label descriptions. *Preprint*, arXiv:2311.07115.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *Preprint*, arXiv:2306.09093.

Xueqi Ma, Jun Wang, Yanbei Jiang, Sarah Monazam Erfani, Tongliang Liu, and James Bailey. 2025. Cognitive mirrors: Exploring the diverse functional roles of attention heads in llm reasoning. In *NeurIPS*.

Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. 2025. Enhancing few-shot vision-language classification with large multimodal model features. In *ICCV*.

Michel Olvera, Paraskevas Stamatiadis, and Slim Essid. 2024. A sound description: Exploring prompt templates and class descriptions to enhance zero-shot audio classification. *Preprint*, arXiv:2409.13676.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Yukun Qian, Xuyi Zhuang, and Mingjiang Wang. 2025. Head information bottleneck (hib): leveraging information bottleneck for efficient transformer head attribution and pruning. *EURASIP Journal on Audio, Speech, and Music Processing*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *ICML*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

J. Salamon, C. Jacoby, and J. P. Bello. 2014. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of ACL*.

Arvind Krishna Sridhar, Yinyi Guo, and Erik Visser. 2024. Enhancing temporal understanding in audio question answering for large audio language models. In *NAACL*.

Yu Sun, Xiaolong Wang, Liu Zhuang, John Miller, Moritz Hardt, and Alexei A. Efros. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*.

James Taylor and Wolfgang Mack. 2025. Improving audio classification by transitioning from zero- to few-shot. In *Interspeech*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *ICLR*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL*.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. 2025. Differentiation and specialization of attention heads via the refined local learning coefficient.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, and 21 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech and Language*.

Huang Xie and Tuomas Virtanen. 2019. Zero-shot audio classification based on class label embeddings. *Preprint*, arXiv:1905.01926.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. 2024. Uniaudio 1.5: Large language model-driven audio codec is a few-shot audio task learner. In *NeurIPS*.

Haoyu Zhang, Jiaxian Guo, Yusuke Iwasawa, and Yutaka Matsuo. 2025a. Aqa-ttrl: Self-adaptation in audio question answering with test-time reinforcement learning. *Preprint*, arXiv:2510.05478.

Xueping Zhang, Zhenshan Zhang, Yechen Wang, Linxi Li, Liwei Jin, and Ming Li. 2025b. Multiapi spoof: A multi-api dataset and local-attention network for speech anti-spoofing detection. *Preprint*, arXiv:2512.07352.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why are visually-grounded language models bad at image classification? In *NeurIPS*.