AUDIO-ONLY SEMANTIC ACOUSTIC IMAGING WITH RECOGNITION-PRIOR SCORE FUSION FOR DCASE2026 TASK 3

Technical Report

Runbang Wang[1]*, Zining Liang*[2]*, Yin Cao*[3]*, Qiuqiang Kong*[2*]

[1]Nanjing University, China
[2]The Chinese University of Hong Kong, Hong Kong SAR, China
[3]Institute of Acoustics, Chinese Academy of Sciences, China
[1]221840194@smail.nju.edu.cn, [2]violetliang@link.cuhk.edu.hk
[3]yin.k.cao@gmail.com, [2]qqkong@ee.cuhk.edu.hk

ABSTRACT

Semantic acoustic imaging predicts class-labeled acoustic regions from microphone-array recordings, producing a spherical map of where sound events appear in a scene. The audio-only setting creates a difficult input-output mismatch: region boundaries are not observed in the waveform, and the event class, acoustic extent, and confidence of each prediction must be estimated from multichannel acoustic cues. We present a system for DCASE2026 Task 3 that first forms spherical acoustic evidence from raw MIC-format audio and then decodes this evidence into mask candidates with event classes and detector scores. A separate AudioMAE-based recognition prior estimates class activity for two-second windows and aligns the probabilities to detector frames. The final fusion stage uses this prior to re-rank the detector candidates while preserving their masks. On the full-recording development test set, the system obtains 0.1017 mAP, 0.2378 AP50, and 0.7904 Pearson r. For submission, prediction compression reduces the maximum JSON size from 148.98 MB to 19.03 MB, with mAP changing from 0.1017 to 0.1009.

Index Terms— Semantic acoustic imaging, sound event localization and detection, recognition-prior fusion

1. INTRODUCTION

Semantic acoustic imaging takes audio recordings as input and produces a spatial map showing where different sound events appear in the scene. A scene can be observed visually, but microphone-array audio also provides an acoustic view by recording sound together with spatial cues. In this acoustic view, different sound events are represented by their categories and spatial regions in the acoustic scene. This capability is useful for acoustic cameras, augmented and virtual reality, and spatial audio systems that need scene awareness from sound. DCASE2026 Task 3 evaluates this problem, and this report focuses on the audio-only track using microphone-array recordings [1].

Prior work provides two useful foundations for this task. For multichannel microphone recordings, sound event localization and detection (SELD) has established models for recognizing active sound events and estimating their directions or trajectories. SELDnet jointly models sound event activity and source direction, while ACCDOA and Multi-ACCDOA encode event activity and localization with compact directional representations [2, 3, 4]. Real-scene benchmarks such as STARSS23 further support this line of work with spatial recordings and spatiotemporal event annotations [5]. A recent DCASE2024 NERC-SLIP system further explored multi-task modeling for SELD with source-distance estimation [6]. In visual scene understanding, set-prediction and mask-classification methods represent a scene as a set of region candidates with class labels, masks, and scores [7, 8, 9]. Recent concept-aware segmentation models further show the value of connecting semantic recognition with mask prediction [10]. DCASE2026 Task 3 connects these two lines by using multichannel microphone recordings as input and class-aware acoustic regions as output.

Semantic acoustic imaging changes the modeling problem by asking for acoustic regions from multichannel microphone recordings. The prediction target is a set of class-aware regions rather than a list of event positions: a sound event may occupy an extended area, and multiple events can be active in the same time frame. Compared with conventional SELD, the desired output is not only an event class with a source direction or trajectory, but an acoustic region for the event. Compared with visual segmentation, the image-like output is not directly observed in the input; it must be inferred from inter-channel level, phase, and temporal cues. The audio-only system therefore needs a representation that can carry spatial acoustic evidence, and a prediction stage that can assign region shape, class, and confidence to each candidate.

We therefore organize the system as an audio-only representation-to-region pipeline. The system first builds a spherical acoustic representation from MIC-format recordings, and then predicts class-aware acoustic regions from this representation. This main pipeline is implemented by the Acoustic Mask Detector, which combines an Audio-to-Spherical Feature Backbone with a Mask Decoder to produce candidate masks, classes, and detector scores. To improve semantic ranking, an AudioMAE Prior Module estimates class priors from two-second windows of the same recording and aligns them to detector frames. A Fusion Module uses these priors to re-rank the detector candidates without changing their masks [11].

This paper is organized as follows. Section 2 introduces the Acoustic Mask Detector, AudioMAE Prior Module, and Fusion Module. Section 3 describes Data preparation, Training, Inference and Export, and Evaluation and Results, including Prediction Compression for Export.

*Corresponding author.

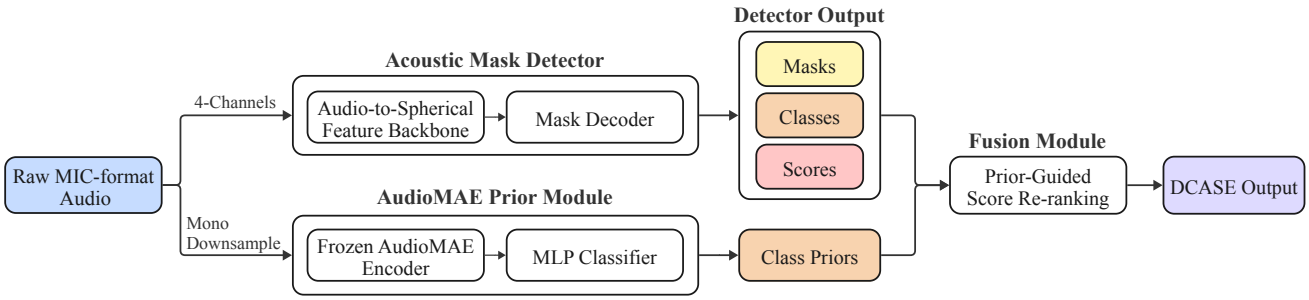

Figure 1: Audio-only system overview. Raw MIC-format audio is processed by the Acoustic Mask Detector and the AudioMAE Prior Module, and the Fusion Module produces ranked candidates for DCASE output.

2. SYSTEM DESCRIPTION

2.1. System overview

Figure 1 summarizes the audio-only inference system, which maps raw four-channel MIC-format audio to ranked class-aware acoustic masks. In this report, each 10 Hz DCASE output frame contains a set of detections, and each detection consists of an event class, a confidence score, and a spherical acoustic mask on a 180×360 elevation–azimuth grid.

The Acoustic Mask Detector is the main prediction path: it maps the MIC-format audio to candidate masks, event classes, and detector scores. In parallel, the AudioMAE Prior Module converts the same recording to a mono downsampled signal and estimates class priors for two-second windows. These priors are aligned to detector frames before fusion. The Fusion Module combines the detector scores with these priors and passes the ranked candidates to the export step.

2.2. Acoustic Mask Detector

2.2.1. Detector interface

The Acoustic Mask Detector maps a raw MIC-format waveform $\mathbf{x} \in \mathbb{R}^{L\times 4}$ to a set of per-frame mask candidates. For each frame, the detector predicts $\{(\mathbf{M}_i, c_i, s_{d,i})\}_{i=1}^{Q}$, where $\mathbf{M}_i$ is an acoustic mask, c_i is an event class, and $s_{d,i}$ is the detector score before prior fusion. The detector has two main components. The Audio-to-Spherical Feature Backbone converts the waveform into a spherical acoustic feature map $\mathbf{F}_{\text{sph}} \in \mathbb{R}^{T_d\times D\times H_s\times W_s}$, where $T_d = 21$, $D = 16$, and $(H_s, W_s) = (45, 90)$ for a 2-s input segment. The Mask Decoder uses this map with learned mask queries to form the candidate masks, classes, and scores.

2.2.2. Audio-to-Spherical Feature Backbone

The backbone converts a 2-s MIC-format waveform into the spherical feature map used by the Mask Decoder. The STFT front end produces a padded magnitude–phase representation $\mathbf{X} \in \mathbb{R}^{202\times 1024\times 12}$, where the 12 channels are log magnitude, sine phase, and cosine phase for the four microphones.

A ConvNeXt-style audio encoder then extracts multi-level time-frequency features [12]. Figure 2 shows the full encoder hierarchy; the spherical projection uses the stage-2, stage-3, and stage-4 features. For the 2-s segment shown in the figure, these features have shapes $101\times128\times192$, $101\times64\times384$, and $101\times32\times768$, respectively. They provide the time-frequency evidence for the spherical projection.

Spherical cross-attention maps the time-frequency features to an elevation–azimuth grid. The learned spherical grid queries have shape $45 \times 90 \times 16$ and serve as the queries. The time-frequency tokens from each ConvNeXt stage serve as keys and values. Each stage produces a $101 \times 16 \times 45 \times 90$ spherical feature map.

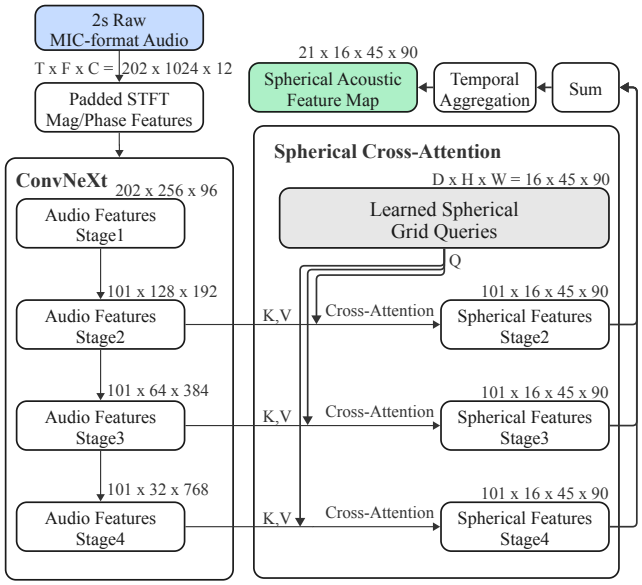

Figure 2: Audio-to-Spherical Feature Backbone. Dimensions are shown for a 2-s input segment. Time-frequency tensors are written as $T \times F \times C$, and spherical tensors are written as $T \times D \times H \times W$. Learned spherical grid queries provide the queries, while stage-2, stage-3, and stage-4 time-frequency features provide keys and values. The cross-attention outputs are summed and temporally aggregated into the spherical acoustic feature map used by the Mask Decoder.

The three stage-wise spherical maps are summed at the 101-frame resolution and then temporally aggregated to form $\mathbf{F}_{\text{sph}} \in \mathbb{R}^{21\times16\times45\times90}$, which is passed to the Mask Decoder.

2.2.3. Mask Decoder

The Mask Decoder follows the query-based mask prediction paradigm of DETR, MaskFormer, and Mask2Former [7, 8, 9]. It maintains $Q = 16$ learned mask queries, each acting as a candidate slot. Across 10 decoder layers, the query states attend to $\mathbf{F}_{\text{sph}}$ and produce intermediate mask predictions; these masks guide the

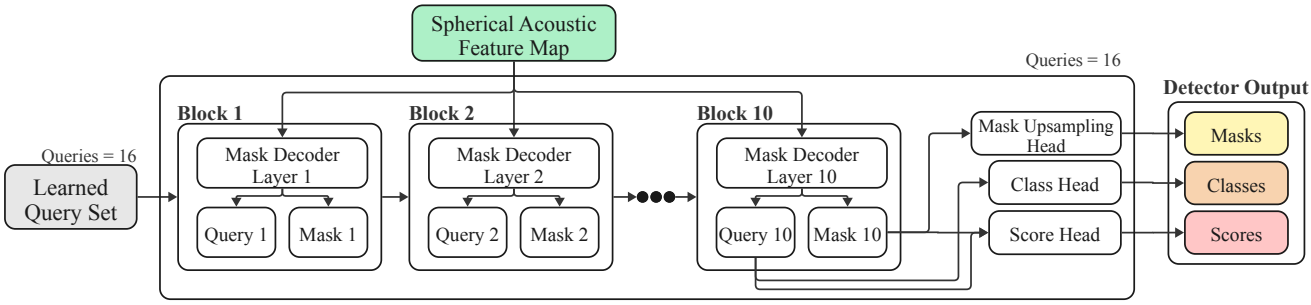

Figure 3: Mask Decoder. The detector uses 16 learned query slots and 10 decoder layers. The labels 1, 2, and 10 indicate decoder layers, and each layer maintains 16 query slots. The spherical acoustic feature map is provided to every decoder layer, while the query states and mask predictions are updated across layers. The final query states are passed to class and score heads, and the final mask prediction is converted to the exported acoustic mask.

next masked-attention step, so query states and masks are refined together. After the final layer, the query states are passed to the class and score heads, while the final mask embedding is combined with mask features and upsampled to form the exported acoustic mask. The score head outputs the detector score s_d for each candidate, which is later calibrated by the Fusion Module.

2.3. AudioMAE Prior Module

The AudioMAE Prior Module receives the same raw MIC-format audio and estimates class activity for two-second windows. It averages the four MIC channels into a mono waveform, resamples the waveform to 16 kHz, and converts it to filterbank features for a frozen AudioMAE encoder. A trained lightweight MLP head maps the pooled AudioMAE embedding for each window to the 13 DCASE event classes. During fusion, the recording is divided on the 10 Hz output timeline into non-overlapping 2-s windows, corresponding to 20 output frames. The class probabilities from each window are assigned to the frames covered by that window as $p(c, t)$; the last window is clipped at the recording boundary.

2.4. Fusion Module

Each detector candidate has a class label, output-frame index, and detector score. The aligned AudioMAE prior supplies $p(c, t)$ for the candidate class c and output frame t. We compute the fused score as

$$s_f = s_d \, p(c, t)^{\alpha}, \tag{1}$$

where s_d is the detector score and s_f is the fused score. The parameter α controls the strength of the recognition prior, and the system uses $\alpha = 0.5$. Fusion changes only the candidate score used for ranking; the predicted class and acoustic mask are kept from the Acoustic Mask Detector. The ranked candidates are then passed to the DCASE export step.

3. EXPERIMENTS

3.1. Data preparation

We construct two datasets for the audio-only mask detector. The synthetic class-agnostic dataset provides acoustic-region masks from simulated room acoustics, while the DCASE augmentation dataset converts official polygon annotations into class-aware spherical mask targets.

3.1.1. Synthetic class-agnostic data

The synthetic class-agnostic dataset is generated from 2 s VCTK speech clips [13]. For each item, source clips are placed in a randomly sampled shoebox room and rendered as 48 kHz four-channel MIC-format audio. The room dimensions are sampled from 2–10 m along the horizontal axes and 2–4 m along height. We use image-source room acoustics implemented with pyroomacoustics [14], with an image source order of five.

The synthetic scenes contain up to six simultaneous sources. To avoid limiting the target to point sources, the renderer represents an extended source with multiple sub-sources sampled inside an angular support. The resulting target is a class-agnostic spherical acoustic mask that describes the source region.

3.1.2. DCASE augmentation data

The DCASE augmentation dataset is built from the STAIRS26/DCASE development recordings and official polygon annotations [15]. Each 2 s MIC-format segment is paired with the annotations active in the segment. The polygon annotations are rasterized into class-aware acoustic mask targets at 180×360 resolution, with 90×180 and 45×90 targets used as lower-resolution mask targets. Each target records an event class and the corresponding spherical acoustic region.

The official MIC recordings contain 32 channels, whereas our detector uses a four-channel MIC-format input. We therefore construct four azimuth views for each segment by selecting four-channel subsets from the array. The views correspond to yaw angles of 0°, 90°, 180°, and 270°. For each view, the rasterized mask is shifted along the azimuth axis by the same yaw angle, keeping the selected audio view and mask target aligned. The four-view augmentation gives 30,024 two-second segments for detector training.

3.2. Training

Table 1 summarizes the training settings for the submitted system. The Acoustic Mask Detector is trained with synthetic class-agnostic masks and then DCASE class-aware masks, while the AudioMAE Prior is trained separately for two-second window class prediction.

3.2.1. Detector training

The first detector stage uses the synthetic class-agnostic masks described in Section 3.1. These masks are generated from randomly sampled simulated scenes and contain acoustic regions with-

Table 1: Training settings for the submitted audio-only system.

Stage	Supervision	Main Loss	Extra Loss	Optimization
Detector (Synthetic Data)	Class-agnostic Mask	Hungarian + BCE + Dice	Mask Refinement	AdamW, lr=1×10^{-4}, batch size=1, steps=500k
Detector (DCASE)	13-class Mask	Hungarian + CE + BCE + Dice	Class Calibration, Candidate Scoring	AdamW, lr=3×10^{-5}, batch size=1, steps=60k
AudioMAE Prior	2-s Class Activity	Multi-label BCE	None	AdamW, lr=3×10^{-4}, batch size=8, steps=30k

out DCASE event classes. This stage trains the Acoustic Mask Detector to predict class-agnostic spherical masks from multichannel audio. The predicted candidates are matched to the target masks with Hungarian matching and optimized with binary cross entropy and Dice losses.

The second detector stage uses the DCASE class-aware mask targets. For each frame, the detector outputs a fixed set of mask candidates, each with a class label, mask, and detector score. We match these candidates to the available targets with Hungarian matching, using a cost based on class prediction, mask binary cross entropy, and mask Dice agreement. Matched candidates receive class and mask losses, while unmatched candidates contribute non-event score supervision. The mask-refinement loss supervises the high-resolution mask output for matched candidates. The class-calibration loss applies additional class-logit supervision, and the candidate-scoring loss trains the detector confidence used for ranking.

3.2.2. AudioMAE Prior Training

The AudioMAE Prior is trained separately from the detector. We keep the pretrained AudioMAE encoder frozen and optimize only the lightweight MLP classifier. The supervision is 13-class activity for two-second windows derived from the DCASE annotations, and the training loss is multi-label binary cross entropy.

3.3. Inference and Export

3.3.1. Inference

At inference time, the Acoustic Mask Detector and AudioMAE Prior Module are applied to the same recording. The Fusion Module re-ranks each detector candidate with the prior of its predicted class; the class label and mask remain those predicted by the detector.

The DCASE output is written at 10 Hz. After ranking, we apply a score threshold of 0.05 and retain at most six detections per frame. Each retained detection contains an event class, fused score, and acoustic mask on the 180×360 spherical grid, serialized with a mask-energy threshold of 0.10.

3.3.2. Prediction Compression for Export

Each predicted mask is stored in the DCASE JSON file as a set of (x, y, e) points, where x and y are the azimuth and elevation grid indices of the 180×360 spherical image, and e is the mask intensity. A dense 180×360 mask may contain many such points, which makes the full-recording JSON files large. We therefore reduce the point set after inference. This reduction is designed around the official evaluation process [1]: the evaluator renders the submitted points with a spherical Gaussian kernel ($\sigma = 6^\circ$) and then thresholds the rendered map at 10% of its peak. The goal is to keep the points that are most important for this rendered mask, rather than to store the dense mask itself.

For each mask, we first remove points below 10% of the mask peak. We then divide the spherical image into local grid cells and keep the strongest remaining point in each cell. A small number of boundary support points is also kept to reduce shrinkage after Gaussian rendering. The submitted setting uses a 2-pixel grid for detections with fused score at least 0.20, and a 6-pixel grid with 2-pixel boundary support for lower-score detections. The compressed file keeps the standard DCASE fields: frame index, class label, score, and mask points.

3.4. Evaluation and Results

We evaluate the original predictions on 78 full recordings from the development test set, using the 10 Hz and 180×360 export setting in Section 3.3. mAP is the main metric, AP50 is AP at an IoU threshold of 0.50, and Pearson r is the average correlation between rendered prediction and reference mask-energy maps for matched pairs.

Table 2: Results on the full-recording development test set.

System	mAP	AP50	Pearson r
Proposed system	0.1017	0.2378	0.7904

Table 3: Prediction compression on the full-recording development test set.

Setting	mAP	Max JSON	Avg. JSON
Original prediction	0.1017	148.98 MB	58.84 MB
Compressed prediction	0.1009	19.03 MB	7.90 MB

The compression result greatly reduces the file size while keeping mAP close to the original prediction. The absolute mAP remains low, showing that robust class-aware acoustic mask prediction remains challenging. The low mAP together with the relatively high Pearson r suggests that the predictions preserve part of the frame-level acoustic energy structure but do not consistently produce correctly ranked class-aware detections. This gap indicates that candidate ranking and class-aware mask selection remain the main bottlenecks of the current audio-only system.

4. CONCLUSION

We presented an audio-only semantic acoustic imaging system for DCASE2026 Task 3. The system connects multichannel acoustic cues to class-aware mask prediction through an Acoustic Mask Detector and an AudioMAE-based score-fusion prior. On the full-recording development test set, the system obtains 0.1017 mAP, 0.2378 AP50, and 0.7904 Pearson r. The prediction-compression step reduces the JSON size for submission with only a small change in mAP. Overall, the results show that forming useful acoustic mask evidence is possible in the audio-only setting, while reliable class-aware candidate ranking remains the main challenge.

5. REFERENCES

[1] "DCASE2026 challenge task 3: Semantic acoustic imaging for sound event localization and detection from spatial audio and audiovisual scenes," https://dcase.community/challenge2026/, 2026.

[2] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[3] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[4] K. Shimada, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *ICASSP*, 2022.

[5] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 72 931–72 957.

[6] Y. Wang, J. Du, *et al.*, "The NERC-SLIP system for sound event localization and detection with source distance estimation of DCASE2024 challenge," DCASE2024 Challenge Technical Report, Tech. Rep., 2024.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.

[8] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 17 864–17 875.

[9] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022.

[10] N. Carion, L. Gustafson, Y.-T. Hu, *et al.*, "SAM 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2026.

[11] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in *Advances in Neural Information Processing Systems*, 2022.

[12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[13] C. Veaux, J. Yamagishi, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," University of Edinburgh, The Centre for Speech Technology Research, 2019.

[14] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[15] I. R. Roman, A. Politis, K. Shimada, H. Cheston, P. Sudarsanam, D. Díaz-Guerra, Y. Sun, T. Shibuya, S. Takahashi, and Y. Mitsufuji, "STAIRS26: Sony-tau acoustic images of real-world scapes 2026," Zenodo, 2026.