DECISION-CALIBRATED SEMANTIC ACOUSTIC IMAGING FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Junhyeong Kwon, Jongsuk Choi

Korea Institute of Science and Technology, Seoul, Korea
meoruu00@gmail.com, cjs@kist.re.kr

ABSTRACT

This report describes our system submitted to the audiovisual track of Task 3 of the DCASE 2026 Challenge on Semantic Acoustic Imaging for SELD. Building on the official UpLAM + Mask R-CNN baseline, we observe that it reconstructs the spatial energy field reasonably well, reaching an audiovisual Pearson correlation of about 0.44 on the development set, yet attains a near-zero detection mAP. We diagnose this as a *decision and calibration* failure rather than a representation failure: a single global confidence threshold cannot simultaneously retain under-confident true positives and suppress background false positives, and the rank-based scorer punishes both. Our system therefore runs the detector at a very low score threshold to surface the under-confident true positives; re-ranks the detections with a learned *matchability* model, a post-hoc calibration layer motivated by conformal and risk-controlled prediction that predicts whether each detection will match the ground truth; and emits a grid-sampled energy mask aligned with the challenge soft-IoU scorer. On the STAIRS26 development set this raises macro mAP from the official baseline's 0.0003 to 0.021 and EFRQ from 0.136 to 0.33–0.36. Absolute scores remain low—the task is hard for every system—but the gain over the baseline is large and consistent. We submit a four-system hedge portfolio that crosses matchability and raw ranking with cap-1 and cap-3 selection, and report development numbers transparently, including the recall ceiling imposed by the 4-channel spatial resolution.

***Index Terms*—** sound event localization and detection, acoustic imaging, calibration, instance segmentation, audiovisual

1. INTRODUCTION

Sound event localization and detection (SELD) jointly detects, classifies, and localizes sound events over time, with applications from robotics and surveillance to immersive media. DCASE 2026 Task 3 [1] reframes SELD as *semantic acoustic imaging*: from low-channel spatial audio, a 4-microphone tetrahedral subset of a 32-channel Eigenmike, the system must reconstruct, per 13-class category and per 10-FPS frame, a dense energy field over a 360×180 equirectangular canvas, separated into instances. The high-resolution 32-channel recordings serve only to generate the ground-truth maps; the model input is the low-channel signal. This departs sharply from prior SELD: the output is a dense mask rather than a sparse direction-of-arrival vector, and evaluation is mask-based—a soft intersection-over-union (IoU) mAP plus an energy-field reconstruction quality (EFRQ)—rather than the classic ER/F/LE/LR metrics.

The official baseline computes UpLAM acoustic features and feeds them, with the corresponding 360° video frame, to a Mask R-CNN–style instance detector. On the development set this baseline reaches a macro mAP of only ≈ 0.0003. Yet the same baseline reconstructs the energy field with a non-trivial Pearson correlation of about 0.44 with video, and its detector, run without confidence filtering, localizes a large fraction of ground-truth events at recall ≈ 0.71 for IoU 0.25 on a held-out subset. **The signal is present; it is the decision rule that fails.** A single global threshold either admits a flood of background false positives or discards the many under-confident true positives, and the rank-based mAP scorer penalizes both.

Our contributions, each targeting one facet of this decision failure, are:

- a *diagnosis* that the near-zero baseline mAP is a decision/calibration failure rather than a representation one, evidenced by recall ≈ 0.71 at IoU 0.25 alongside mAP ≈ 0.0003;
- *low-threshold detection* that surfaces the under-confident true positives a single global threshold discards;
- a learned *matchability reranker*—a post-hoc calibration layer that restores a calibrated detection ordering without modifying the acoustic front-end or retraining the energy heads; and
- a soft-IoU–matched grid-mask emission, packaged as a four-system hedge portfolio that is explicit about development→evaluation transfer risk.

Together, these post-hoc changes raise development macro mAP from the baseline's 0.0003 to 0.021 and EFRQ from 0.136 to 0.33–0.36 (Table 1), with no change to the trained representation. The remainder of the report describes the input and detector in Sec. 2 and the experiments, results, and limitations in Sec. 3.

2. SYSTEM DESCRIPTION

2.1. Input and acoustic features

Figure 1 summarizes the pipeline. The model input is the 4-channel tetrahedral microphone subset at 24 kHz, using channels 5, 9, 25, and 21 of the Eigenmike, which the evaluation set provides directly. Per frame we compute the cross-spectral visibility matrix and pass it through **UpLAM**, the Upsampled Latent Acoustic Mapping model [2], a self-supervised model that learns to upsample a low-channel covariance matrix into a latent acoustic map for direction-of-arrival estimation; we use the organizer-provided frozen weights. This yields a 9-band latent of shape 9×484, which is projected onto the equirectangular grid to form a $9 \times 180 \times 360$ acoustic feature

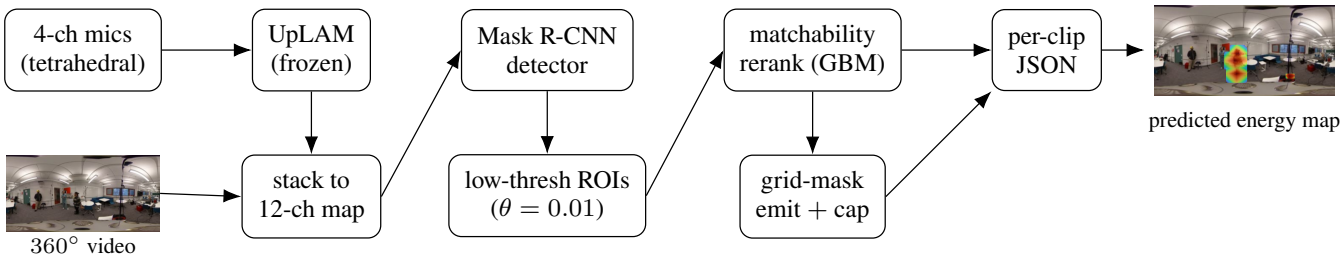

Figure 1: System pipeline. The frozen UpLAM front-end maps the 4-channel covariance to a 9-band acoustic latent, stacked with the RGB frame into a 12-channel map; a Mask R-CNN detector run at a low threshold proposes many candidate detections; a learned matchability model reranks them, and a grid-mask stage emits the per-clip energy mask. Only the detector and reranker are trained; UpLAM is frozen.

map. For the audiovisual track we stack the corresponding 360×180 equirectangular RGB frame, giving a 12-channel input of 3 RGB and 9 acoustic channels. Frames absent from the video stream are filled with the ImageNet mean. The audiovisual input is essential: an audio-only field is effectively noise on the development set, with Pearson ≈ 0.08–0.10, whereas the audiovisual field carries usable signal at ≈ 0.44.

2.2. Detector

The detector is a Mask R-CNN [3] variant on the 12-channel input with about 44.9M parameters; the UpLAM acoustic front-end is a separate frozen module of about 2.8M. For each instance it predicts a bounding box, one of 13 class labels, a confidence score, and a 28×28 per-instance energy map. It is trained on the full development-train split following the organizer recipe. The natural-image backbone is *not* frozen, since freezing collapses the acoustic field to noise; instead a progressive-unfreeze schedule keeps layer1 trainable throughout, unfreezes layer4 from epoch 3 and layer3 from epoch 7, and keeps layer2 frozen, all with per-group learning rates, RGB Dropout for modality robustness, and energy-map supervision. Training used mixed precision and batch size 4. We select epoch 4 by realized development mAP; the energy-field correlation keeps rising at later epochs but box localization over-fits after about epoch 2.

2.3. Low-threshold detection

At inference the region-of-interest head is run with `score_thresh` $= 0.01$ and up to 100 detections per image—far below the baseline operating point. This surfaces the under-confident true positives a global threshold would discard, at the cost of many false positives, which the reranking stage and the rank-based scorer tolerate.

2.4. Matchability reranking

We train a gradient-boosted classifier [4] on development detections to predict *matchability*, defined as whether a detection's best IoU with ground truth is ≥ 0.25. The model is a `HistGradientBoosting` classifier with 120 iterations, depth 6, learning rate 0.1, $L_2 = 1.0$, at least 100 samples per leaf, and early stopping. Each detection is described by a 16-dimensional feature vector. Fourteen features are detection-intrinsic: the raw score; box area, elevation, and aspect ratio; co-detection counts per frame and per frame-and-class; energy-map max, mean, std, prominence, top-10% concentration, normalized entropy, and peak offset; and a local

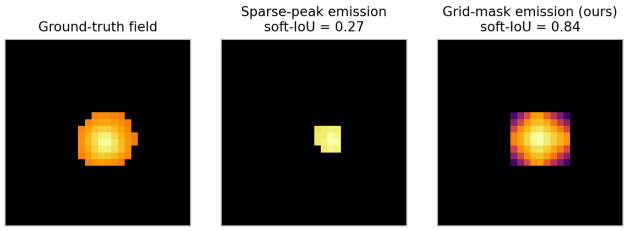

Figure 2: Mask emission. The ground-truth energy field, a sparse-peak emission, and our grid-mask emission, all rendered by the official scorer. The grid mask renders the detector's own predicted energy over its full extent, rather than a few peak pixels, so it matches the dense ground-truth field; this is a faithful rendering, not mask inflation, and attains a far higher soft-IoU.

non-maximum-suppression dominance ratio. The remaining two are frame-level acoustic gate features, closure and diffuseness, that are zeroed at inference, so no ground-truth-derived information leaks at test time. Only a few percent of the low-threshold detections are matchable, so the reranker must surface a sparse positive set from a large candidate pool; this fraction also upper-bounds the achievable gain, as discussed in Sec. 3. The classifier's positive-class probability, *pmcp*, replaces the raw score in the submitted `score` field, which drives the mAP ranking. As a post-hoc calibration fitted on the development distribution, it is motivated by conformal and risk-controlled prediction [5, 6]; on development it roughly doubles macro mAP over raw ranking. Instance *selection*, described in Sec. 2.5, remains by raw score, since the secondary EFRQ metric is insensitive to reranking.

2.5. Grid-mask emission and selection

Each kept detection is converted to a dense mask by sampling its 28×28 energy map on a 3×3 grid inside the box, which we call "grid3": nine energy-weighted $[x, y, \text{intensity}]$ points in the 360×180 output, with intensities min–max normalized per instance. Because the ground-truth field is a dense blob rather than a point, a sparse-peak emission under-renders it and is structurally penalized by soft-IoU; the grid mask instead renders the detector's own energy over its predicted extent (Figure 2), changing neither the detections nor their scores. The 3×3 grid reproduces the IoU of a denser development-time grid while keeping every per-clip JSON below the 20 MB submission limit. Selection keeps the top detection per

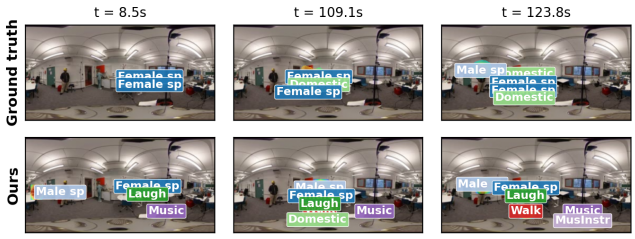

Figure 3: Qualitative comparison on a development clip: ground-truth acoustic energy on the top row and our prediction on the bottom row, overlaid on the 360° frame with class labels.

frame and class at cap = 1, or the top three at cap = 3, and at most 6 detections per frame, ordered by raw score.

2.6. Submission portfolio

The official scorer ranks systems by cumulative rank over macro mAP, AP50, and EFRQ, where the mAP is a soft-IoU mAP over thresholds 0.25, 0.5, and 0.75. Because the matchability reranker is fit on development data and may transfer imperfectly to the different-room evaluation set, we submit a four-system hedge, shown in Table 1, organized as a 2×2 factorial that crosses matchability and raw ranking with cap-1 and cap-3 selection. The ranking axis isolates the calibration gain and its transfer risk: the raw-ranking systems are insurance against reranker distribution shift. The selection axis isolates the mAP–EFRQ trade-off, as cap 3 gives up a little mAP for higher EFRQ.

3. EXPERIMENTS

All experiments use the STAIRS26 development set [7], comprising 168 clips and about 7.5 h across 13 classes at 10 FPS, with the official dev-train split of 90 clips and dev-test split of 78 clips; the evaluation set comprises 79 clips. STAIRS26 builds on the STARSS23 [8] and STARSS22 [9] real-scene spatial-audio recordings; audiovisual frames are extracted from the corresponding perspective videos at 10 FPS and 360×180. Scoring uses the official baseline `evaluate.py`, which reports macro soft-IoU mAP at thresholds 0.25, 0.5, and 0.75, together with AP50 and EFRQ. The submitted systems use no external datasets and no data augmentation.

Results. Table 1 reports development results on the dev-test split with audiovisual inference. Our low-threshold detector with a sparse-peak emission scores macro mAP 0.0013, already above the official baseline. Replacing the sparse emission with our grid mask (Fig. 2) lifts this to 0.011, about $37\times$ the baseline, and matchability reranking then roughly doubles it to 0.021, about $70\times$; the grid emission also raises EFRQ from 0.136 to 0.33–0.36. All of these gains come from post-hoc emission and ranking choices, with no change to the trained representation. Figure 3 shows a qualitative comparison.

Limitations. *All quoted numbers are development-set.* The matchability reranker is fit on development detections and re-applied to evaluation detections, so its real evaluation gain is uncertain: the +88% development gain is an optimistic ceiling, and the model is exposed to a development-to-evaluation room shift. The raw-ranking systems are submitted as insurance against this shift.

Table 1: Development-set results, audiovisual dev-test split. Grid: grid-mask emission; MCP: matchability rerank; best per column in bold.

System	Grid	MCP	Cap	mAP↑	EFRQ↑
Baseline			—	0.0003	0.136
Sparse peaks			1	0.0013	0.293
TASK3B_3	✓		1	0.0111	0.329
TASK3B_4	✓		3	0.0110	**0.360**
TASK3B_1	✓	✓	1	**0.0209**	0.329
TASK3B_2	✓	✓	3	0.0205	**0.360**

Recall ceiling. Even with unfiltered detection that keeps every proposal, a large fraction of ground-truth events is never localized well enough to count. The recall ceiling—the best achievable recall if every proposed box were accepted—falls from 0.82 at IoU 0.1 to 0.71 at IoU 0.25 and just 0.48 at IoU 0.5, so roughly half of all events go unlocalized at the stricter threshold. This is a 4-channel spatial-resolution limit that constrains all participants. Rare classes such as bell, knock, and door have near-zero AP. No post-hoc reranking recovers events the detector never proposes; breaking this ceiling needs a stronger detector or representation, not a better decision rule. **Over-prediction is largely benign** under the rank-based scorer; aggressive suppression through lower caps or temporal filtering measurably *hurt* EFRQ, so we keep a permissive per-frame budget.

4. CONCLUSION

On the audiovisual track of DCASE 2026 Task 3, we raise development macro mAP from the official baseline's 0.0003 to 0.021 using only post-hoc, non-retraining changes to the official UpLAM + Mask R-CNN system: a soft-IoU–matched grid-mask emission, low-threshold detection, and learned matchability reranking. Absolute scores stay low, as the task is hard for every system, but the gains are consistent and require no change to the trained representation. The remaining bottleneck is detection recall, bounded by the 4-channel input. Future work targets the representation rather than the decision: stronger acoustic-spatial encoders, spatial-audio augmentation and simulated room impulse responses to densify rare classes, learned temporal context, and direct dense energy-map decoders.

5. REFERENCES

[1] DCASE 2026 Challenge, Task 3: Semantic Acoustic Imaging for Sound Event Localization and Detection, `https://dcase.community/challenge2026/`.

[2] A. S. Roman, I. R. Roman, and J. P. Bello, "Latent acoustic mapping for direction of arrival estimation: a self-supervised approach," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025, pp. 1–5.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE ICCV*, 2017, pp. 2961–2969.

[4] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. NeurIPS*, 2017, pp. 3146–3154.

[5] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023.

[6] V. Rozenfeld and B. Laufer-Goldshtein, "Uncertainty quantification and risk control for multi-speaker sound source localization," *arXiv:2603.17377*, 2026.

[7] 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, doi:10.5281/zenodo.18171005.

[8] K. Shimada, A. Politis *et al.*, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. NeurIPS Datasets and Benchmarks*, 2023.

[9] A. Politis *et al.*, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2022.