MULTI-ENCODER FUSION WITH LORA AND RANK NORMALIZATION FOR FIRST-SHOT ANOMALOUS SOUND DETECTION

Technical Report

Nehemiah Balozi

Sungkyunkwan University
Department of Intelligent Robotics Engineering
Suwon, South Korea
nehemiah@g.skku.edu

*Hyungpil Moon**

Sungkyunkwan University
Department of Mechanical Engineering
Suwon, South Korea
hyungpil@skku.edu

ABSTRACT

We present a multi-encoder fusion system for noise-aware unsupervised anomalous sound detection. Six pretrained audio encoders are combined: frozen CLAP, WavLM-Base, and AST with trainable projection heads, plus BEATs adapted via Low-Rank Adaptation (LoRA) at two ranks (r=16, r=32) and CLAP adapted via LoRA. Adapter layers and projection heads are trained on normal data using InfoNCE contrastive loss. Anomaly scores are computed via k-nearest neighbor distance in 256-dimensional embedding space, normalized using per-machine per-domain rank normalization, and fused via weighted score averaging. Key findings include: LoRA adaptation of BEATs and CLAP improved performance over frozen encoders in our development experiments, and per-(machine, domain) rank normalization improved development-set performance and was adopted in all submitted systems. On the DCASE 2026 Task 2 development dataset, our best six-stream system achieves a development-set macro(3) of 60.06% across the seven development machine types.

***Index Terms*—** Anomalous sound detection, LoRA fine-tuning, Multi-encoder score fusion, Rank normalization, Pretrained audio encoders

1. INTRODUCTION

DCASE 2026 Task 2 addresses first-shot anomalous sound detection (ASD): given only normal examples from a target machine at training time, systems must detect anomalous test clips without access to target anomalies [1]. The 2026 edition extends this setting with stronger domain shift between source and target operating conditions and with noise-aware evaluation, so models must generalize across machines, domains, and acoustic environments rather than memorizing a single training distribution. Effective systems therefore combine robust pretrained representations, adaptation to limited normal data, and scoring methods that remain stable when source and target score distributions differ.

Recent top-performing approaches illustrate several complementary directions. Wilkinghoff's work on sub-cluster AdaCos embeddings with careful score normalization has demonstrated strong single-model performance on DCASE ASD benchmarks [5]. Outlier-exposure methods pretrained on large external datasets such as AudioSet improve discrimination by exposing the model to diverse acoustic contexts beyond the development machines. Multi-encoder ensemble systems have recently emerged as a common strategy in DCASE submissions, fusing complementary backbones to reduce the failure modes of any single feature space. Our work follows this ensemble tradition but pushes it toward finer-grained adaptation and normalization.

We propose a six-stream encoder fusion framework that combines CLAP [6], WavLM [7], two BEATs-LoRA streams [8, 9], CLAP-LoRA, and AST [10]. Rather than relying on a single LoRA configuration, we treat BEATs fine-tuned at ranks 16 and 32 as complementary streams that capture different capacity vs. generalization trade-offs. At scoring time, we apply per-(machine, domain) rank normalization to each stream before fusion; this step consistently improved development-set performance during model selection. We submitted four systems that share the same inference pipeline and fusion architecture but differ in checkpoint training scope (7-machine vs. all12-machine LoRA/AST weights), to study how extra-machine pretraining affects generalization to unseen evaluation machines. Together, these design choices aim to combine the representational diversity of multi-encoder fusion with parameter-efficient domain adaptation and distribution-robust score fusion.

The remainder of this report is organized as follows. Section 2 describes our encoder architectures, contrastive pretraining, inference pipeline, and the exact configuration of each submitted system. Section 3 presents experimental results on the development dataset. Section 4 concludes.

2. SYSTEM DESCRIPTION

2.1. Overview

We submitted four anomaly-detection systems to DCASE 2026 Task 2. Each system fuses k-nearest-neighbor (k-NN) anomaly scores from multiple pretrained audio encoders. Six encoder types are used in total: frozen CLAP and WavLM with trainable projection heads; two BEATs streams with LoRA fine-tuning at ranks 16 and 32; CLAP with LoRA on the audio encoder; and frozen AST with a trainable projection head. Encoders are contrastively pretrained on development-set normal clips, then at inference time each stream builds a memory bank of normal embeddings, applies MemMixup augmentation, computes k-NN distances, rank-normalizes scores per machine and domain, and fuses the normalized stream

*Corresponding author: hyungpil@skku.edu

scores with fixed weights. Systems 1–3 are six-stream ensembles; System 4 is a six-stream variant in which the AST contribution is reduced to a negligible 0.01. The four submissions differ mainly in which checkpoints were trained on seven development machines versus twelve machines (development plus five evaluation machines).

2.2. Encoder Architectures

All encoders output 256-dimensional embeddings after a shared projection head design (linear layer plus ReLU).

CLAP (Stream 1). Pretrained model: `laion/clap-htsat-unfused`. The full `ClapAudioModelWithProjection` backbone is frozen. Only a trainable Linear(512, 256) + ReLU projection head is optimized. Trainable parameters: 131,328. Output dimension: 256. Audio is resampled from 16 kHz to 48 kHz before the CLAP feature extractor.

WavLM (Stream 2). Pretrained model: `microsoft/wavlm-base`. The WavLM transformer is fully frozen; frame hidden states are mean-pooled and passed through Linear(768, 256) + ReLU. Trainable parameters: 196,864.

BEATs LoRA r=16 (Stream 3). Pretrained backbone: Microsoft BEATs (`BEATs_iter3_plus_AS2M`). LoRA adapters are inserted into the last four transformer blocks on query, key, and value projections. LoRA rank $r = 16$, $\alpha = 32$. Frame features are mean-pooled and projected through Linear(768, 256) + ReLU. Trainable parameters: 491,776.

BEATs LoRA r=32 (Stream 4). Same architecture and checkpoint source as Stream 3, with LoRA rank $r = 32$ and $\alpha = 64$. Trainable parameters: 786,688.

CLAP-LoRA (Stream 5). Pretrained model: `laion/clap-htsat-unfused`. LoRA is applied to the query and value attention layers in the last four Swin-style blocks of the HTS-AT audio encoder. LoRA rank $r = 8$, $\alpha = 16$. All non-LoRA CLAP weights are frozen; the projection head Linear(512, 256) + ReLU is trainable. Trainable parameters: 205,056.

AST (Stream 6). Pretrained model: `MIT/ast-finetuned-audioset-10-10-0.4593`. The AST backbone is frozen; log-mel features are extracted internally, sequence hidden states are mean-pooled, and Linear(768, 256) + ReLU is trained. Trainable parameters: 196,864.

2.3. Contrastive Pretraining

Each encoder is fine-tuned on DCASE 2026 development-set normal training clips using a two-view contrastive objective.

Loss function. Training uses an InfoNCE / NT-Xent objective [11] with a small variance regularization term. Positive pairs are two augmented views of the same clip; negatives are all other clips in the batch. Temperature $\tau = 0.07$. Optimizer: AdamW with learning rate $1e{-4}$, batch size 64, weight decay $1e{-4}$. Training runs up to 20 epochs with early stopping based on validation embedding spread (pairwise L2 variance), intended to avoid representation collapse.

Training augmentations. Each clip yields two independently augmented views: additive Gaussian noise ($\sigma = 0.005$), random multiplicative gain uniformly drawn in ± 3 dB, and a random temporal crop of 0.5 s removed from a 10 s clip and zero-padded back to 10 s. Clips are mono, 16 kHz, fixed length 160,000 samples.

Checkpoint training data. "7-machine" checkpoints are trained on the seven development machine types only. "all12" checkpoints additionally include normal training data from the five evaluation machine types, which DCASE provides as additional training data for first-shot adaptation. Only unlabeled normal clips were used; no anomaly labels or test data from evaluation machines were accessed during training.

2.4. Inference and Scoring

Memory bank construction. For each stream, all development-set normal training clips (source and target domains) are embedded. Target-domain normals are oversampled so that the ratio of source to target bank entries per machine is at most 10:1. Embeddings are L2-normalized before distance computation.

MemMixup. After the memory bank is built, MemMixup doubles the number of bank rows per machine type. For each machine, n_m synthetic rows are added by convex-combining random pairs of existing bank embeddings with $\lambda \sim \text{Beta}(0.4, 0.4)$; mixed vectors are re-normalized to unit L2 norm.

k-NN scoring. Anomaly score for a test clip is the mean L2 distance to its $k = 5$ nearest neighbors in the memory bank. Multi-segment scoring and test-time adaptation are disabled in all submitted runs.

Rank normalization. Raw k-NN scores are converted to ranks using only the unlabeled test scores within each (machine type, domain) partition, without using anomaly labels. Ranks are linearly scaled to $[0, 1]$ by dividing by $(n - 1)$. Rank normalization is applied independently per stream before fusion.

Fusion. The final anomaly score is a weighted sum of rank-normalized stream scores:

$$s_{\text{fused}} = \sum_{i \in \text{streams}} \alpha_i \cdot s_i^{\text{rank}} \tag{1}$$

with all α values summing to 1.0.

2.5. Submitted Systems

All four systems use rank normalization, MemMixup, and disable multi-segment scoring and TTA. CLAP and WavLM checkpoints are shared across all submissions.

System 1 (MEF6-7m). Six-stream ensemble; all LoRA and AST checkpoints trained on 7 machines. Weights: CLAP=0.19, BEATs-r16=0.23, CLAP-LoRA=0.23, WavLM=0.10, BEATs-r32=0.15, AST=0.10.

System 2 (MEF6-12m). Six-stream; same weights as System 1; all LoRA and AST checkpoints trained on 12 machines.

System 3 (MEF6-hyb). Six-stream hybrid; BEATs LoRA from 7-machine training; CLAP-LoRA and AST from 12-machine training. Same weights as System 1.

System 4 (MEF5-7m). Six-stream system with reduced AST contribution (weight 0.01). All LoRA and AST checkpoints from 7-machine training. Weights: CLAP=0.22, BEATs-r16=0.26, CLAP-LoRA=0.26, WavLM=0.10, BEATs-r32=0.15, AST=0.01.

3. RESULTS

3.1. Experimental Setup

The development dataset comprises seven machine types (ToyCar, ToyCarEmu, bearingEmu, fan, gearboxEmu, sliderEmu, valveEmu) drawn from ToyADMOS2 [2] and MIMII DG [3], each with source and target domain recordings, following the first-shot evaluation

protocol of Harada et al. [4]. Training uses normal clips only; test clips include both normal and anomalous examples. The evaluation dataset adds five unseen machine types (BlowerDustCollector, Sander, SewingMachine, ToothBrush, ToyDrone). Performance is reported per machine using AUC_source, AUC_target, and pAUC at max FPR = 0.1. Macro averages are computed as unweighted means across machines; macro(3) is the arithmetic mean of the three macro metrics.

3.2. Development Set Results

Table 1 summarizes macro performance across the four systems. Per-machine breakdowns are shown in Tables 2–4.

Table 1: Macro-averaged development results (%).

System	AUC_s	AUC_t	pAUC	macro(3)
MEF6-7m	63.71	61.70	54.77	**60.06**
MEF6-12m	63.51	59.95	54.14	59.20
MEF6-hyb	63.62	61.52	54.56	59.90
MEF5-7m	63.48	61.59	54.93	60.00

Table 2: Per-machine AUC_source (%).

Machine	Sys 1	Sys 2	Sys 3	Sys 4
ToyCar	74.42	75.36	75.13	73.84
ToyCarEmu	57.05	57.44	57.14	56.60
bearingEmu	55.08	56.83	55.26	55.04
fan	51.16	49.30	50.60	50.40
gearboxEmu	68.22	65.30	67.42	68.12
sliderEmu	58.91	58.70	59.20	59.94
valveEmu	81.13	81.64	80.61	80.39

Table 3: Per-machine AUC_target (%).

Machine	Sys 1	Sys 2	Sys 3	Sys 4
ToyCar	69.36	66.96	68.92	67.74
ToyCarEmu	64.84	65.80	66.03	65.42
bearingEmu	51.58	49.12	51.04	51.04
fan	50.86	49.48	51.73	50.85
gearboxEmu	63.01	60.92	63.41	63.48
sliderEmu	60.42	57.06	58.26	61.22
valveEmu	71.82	70.30	71.24	71.36

3.3. Analysis

System 1 (MEF6-7m) ranks first on macro(3) at 60.06%. Its strongest advantage is in AUC_target (61.70%), which is 1.75–2.75 percentage points above Systems 2 and 4. Under our training configuration, checkpoints trained exclusively on development machines produced better target-domain performance than checkpoints trained on the expanded 12-machine set.

System 2 (all12 checkpoints) performs worst overall, despite achieving the highest AUC_source on ToyCar and valveEmu. The all12 weights show reduced transfer to target-domain conditions, despite achieving higher source-domain AUC on some machines. System 3 (hybrid) nearly matches System 1, indicating that the 12-machine CLAP-LoRA and AST checkpoints contribute useful diversity without the target-domain penalty of all12 BEATs. System 4 (reduced-AST variant) is competitive and achieves the best gearboxEmu pAUC and sliderEmu AUC_target, showing that redistributing weight from AST to the LoRA streams can help specific machines.

Table 4: Per-machine pAUC (%).

Machine	Sys 1	Sys 2	Sys 3	Sys 4
ToyCar	63.03	58.84	61.32	62.37
ToyCarEmu	52.26	51.95	50.63	52.16
bearingEmu	52.11	52.11	51.79	51.89
fan	48.95	49.89	48.74	49.32
gearboxEmu	58.11	56.95	60.58	60.26
sliderEmu	52.63	53.89	53.95	52.84
valveEmu	56.32	55.32	54.95	55.68

Per-machine, valveEmu is consistently the easiest (AUC_source 80–82%) while fan is the hardest (AUC near chance). gearboxEmu shows the largest spread across systems. Evaluation-set results were not available at the time of writing.

4. CONCLUSION

We presented four multi-encoder fusion systems for DCASE 2026 Task 2 first-shot anomalous sound detection. Our approach combines six pretrained audio encoders with parameter-efficient adaptation, k-nearest-neighbor scoring, per-(machine, domain) rank normalization, and weighted score fusion. On the development set, our best system (System 1, MEF6-7m) achieves macro(3) = 60.06%.

Among our four submissions, the 7-machine LoRA/AST checkpoints (System 1) generalized best to the development distribution, while the all12-machine variant (System 2) traded source-domain gains for reduced target-domain robustness. The hybrid configuration (System 3) demonstrated that selectively mixing 7-machine and all12-machine checkpoints can recover most of System 1's performance.

Several promising directions remain unexplored. Outlier exposure with external audio data such as AudioSet, proven effective by past DCASE winners, was not used in our submissions. Test-time adaptation on the 10 target-domain normal clips per machine could further reduce source-target gaps. Inlier-modeling approaches via auxiliary machine-identification tasks may complement our contrastive embeddings.

5. REFERENCES

[1] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and Discussion on DCASE 2026 Challenge Task 2: Noise-aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," arXiv:2606.01578, 2026.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proc. DCASE Workshop*, Barcelona, Spain, Nov. 2021, pp. 1–5.

[3] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proc. 7th DCASE Workshop*, Nancy, France, Nov. 2022.

[4] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," in *Proc. 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191–195.

[5] K. Wilkinghoff, "Sub-cluster AdaCos: Learning representations for anomalous sound detection," in *Proc. IJCNN*, 2021.

[6] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE ICASSP*, 2023.

[7] S. Chen, C. Wang, Z. Chen, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, 2023, pp. 5178–5193.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[10] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[11] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv:1807.03748, 2018.