# GISP@HEU'S SUBMISSION FOR DCASE 2026 TASK 2: TRAINING-FREE ANOMALOUS SOUND DETECTION WITH ADAPTIVE LAYER SELECTION

Technical Report

*Tong Ye*[1], *Yao Xiao*[1], *Wenbo Wang*[1], *Qiaoxi Zhu*[2], *Jian Guan*[1*]

[1]Group of Intelligent Signal Processing, Harbin Engineering University, Harbin, China
[2]University of Technology Sydney, Ultimo, Australia

## ABSTRACT

This report presents our submission to Task 2 of the DCASE 2026 Challenge, which focuses on developing noise-robust unsupervised anomalous sound detection (UASD) systems for first-shot machine condition monitoring under domain-shift conditions. To address this challenge, we propose a fully training-free system consisting of three frozen audio pre-trained models (i.e., BEATs, CED, and EAT) together with an adaptive layer selection strategy to obtain generalizable and robust machine sound representations for anomalous detection. Experimental results on the DCASE 2026 Task 2 development set demonstrate the effectiveness of the proposed training-free system, with the harmonic-mean AUC and harmonic-mean pAUC reaching 53.7% and 59.4%, respectively.

***Index Terms*—** Anomalous sound detection, Domain Generalization, Audio Pretrained models, Hyperbolic Embeddings, Memory Bank

## 1. INTRODUCTION

Unsupervised anomalous sound detection (ASD) aims to identify anomalous sounds emitted by a target machine using only normal sound for training. Building on previous DCASE Task 2 settings, this task copes with several practical challenges that arise in real-world deployment [1]. One major difficulty is the presence of domain shifts between source and target domains, which are caused by variations in operating conditions and recording environments [2, 3]. Moreover, the task adopts a first-shot setting, meaning that the machine types used for evaluation are not available during development [4, 5]. Harada et al. introduced a domain-generalization baseline for first-shot machine condition monitoring [6], highlighting the importance of generalization to unseen machine types. As a result, machine-specific optimization strategies cannot be relied upon, requiring models to generalize effectively across unseen machine types.

In DCASE 2026, the Task 2 is further extended to a noise-aware setting with two-channel recordings captured by near and far microphones [1]. Although the far-channel signal may provide useful noise-reference information, the use of multi-distance recordings also introduces additional distributional variability. Together with the limited target-domain normal samples, these factors make robust and generalizable anomaly scoring particularly challenging.

To address these constraints, we design a training-free system with adaptive layer selection, built upon frozen audio pre-trained models, including BEATs [7], CED [8], and EAT [9]. The system extracts complementary layer-wise representations from multiple backbones and stores them as memory banks [10]. By comparing input embeddings with normal reference embeddings, the system enables robust first-shot anomaly detection without gradient updates, anomaly labels, or machine-specific hyperparameter tuning.

*Corresponding author.

## 2. PROPOSED SYSTEM

We propose a fully training-free ASD system based on multiple frozen audio pre-trained models (i.e., BEATs [7], CED [8], and EAT [9]). Instead of training a task-specific detector, the system exploits complementary representations from these frozen audio pre-trained models. Since different Transformer layers capture different levels of acoustic information, we leverage an adaptive layer selection strategy to select informative layer-wise representations. The selected representations are then used for memory-bank-based anomaly scoring to achieve anomalous sounds detection.

## 3. EXPERIMENTAL RESULTS

### 3.1. Dataset

The DCASE 2026 Task 2 development set used in this work is constructed from machine-sound datasets designed for anomalous sound detection under domain-shift and domain-generalization conditions, including ToyADMOS2 and MIMII DG [11, 12]. It contains seven machine types: ToyCar (Emu), ToyCar, Fan, Gearbox (Emu), Bearing (Emu), Slide rail (Emu), and Valve (Emu). For each machine type, the dataset provides $1{,}000$ normal training clips and a labeled test split containing both source- and target-domain clips. All audio clips are single-channel recordings sampled at 16,kHz. Among the seven machine types, Fan, Gearbox (Emu), and ToyCar (Emu) include attribute metadata, such as operating speed or product configuration, whereas the remaining four machine types are released without attribute information.

Following the challenge protocol, performance is reported as the area under the ROC curve (AUC) overall and per domain, the partial AUC (pAUC) at a false-positive rate of $0.1$, and the official score, defined as the harmonic mean of the source AUC, target AUC, and pAUC.

### 3.2. Experiment Setup

All backbones are initialized with publicly available pre-trained weights and kept frozen throughout the entire system. The input

audio is processed using a fixed duration of 10s. For each backbone, we extract multi-layer Transformer representations with adaptive layer selection strategy to obtain clip-level embeddings.

Table 1: Development-set results of the proposed system. Columns are AUC, partial AUC, source/target AUC, and the official Hmean score. The bottom row is the harmonic mean across machine types.

| Machine type | AUC | pAUC | $AUC_s$ | $AUC_t$ | Hmean |
|---|---|---|---|---|---|
| ToyCar (Emu) | 0.699 | 0.513 | 0.561 | 0.877 | 0.616 |
| ToyCar | 0.768 | 0.561 | 0.730 | 0.818 | 0.685 |
| Fan | 0.540 | 0.542 | 0.641 | 0.415 | 0.516 |
| Gearbox (Emu) | 0.603 | 0.561 | 0.600 | 0.605 | 0.588 |
| Bearing (Emu) | 0.608 | 0.589 | 0.641 | 0.569 | 0.598 |
| Slider (Emu) | 0.592 | 0.500 | 0.624 | 0.569 | 0.560 |
| Valve (Emu) | 0.690 | 0.506 | 0.746 | 0.673 | 0.625 |
| Overall | 0.635 | 0.537 | 0.643 | 0.613 | **0.594** |

### 3.3. Results

Table 1 reports the per-machine and aggregate results on the DCASE 2026 Task 2 development set. For each machine type, Hmean denotes the official score, computed as the harmonic mean of source AUC, target AUC, and pAUC. The bottom row reports the harmonic mean across the seven machine types.

## 4. CONCLUSION

We presented a fully training-free system which consisits of three frozen audio pre-trained models and an adaptive layer selection strategy for DCASE 2026 Task 2. The design contains no learnable parameters, which makes it well suited to the first-shot, domain-generalized setting of the challenge. On the development set it reaches an overall official score of $0.594$ without any training.

## 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," *In arXiv e-prints: 2606.01578*, 2026.

[2] H. Lan, Q. Zhu, J. Guan, Y. Wei, and W. Wang, "Hierarchical metadata information constrained self-supervised learning for anomalous sound detection under domain shift," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 7670–7674.

[3] J. Guan, J. Tian, Q. Zhu, F. Xiao, H. Zhang, and X. Liu, "Disentangling hierarchical features for anomalous sound detection under domain shift," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5.

[4] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2305.07828*, 2023.

[5] H. Zhang, Q. Zhu, J. Guan, H. Liu, F. Xiao, J. Tian, X. Mei, X. Liu, and W. Wang, "First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1271–1275.

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

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

[8] H. Dinkel, Z. Wang, Y. Yan, M. Niu, J. Zhang, Y. Wang, and B. Wang, "CED: Consistent ensemble distillation for audio tagging," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 291–295.

[9] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[10] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 298–14 308.

[11] 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. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[12] 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 Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.