EAT TOKEN-MFS/GBC BEAM ANOMALOUS SOUND DETECTION SYSTEM FOR DCASE2026 TASK 2

Technical Report

Hanbin Zhou[1], Shuchi Chen[1], Shenbin Chen[1], Zhifang Zheng[1], Shuzheng Tang[1]

School of Artificial Intelligence and Big Data
Hefei University, Hefei, China
2384883301@qq.com, 2929950149@qq.com, shbchen@hfuu.edu.cn
2813517120@qq.com, 895547788@qq.com

ABSTRACT

This report describes our anomalous sound detection system for DCASE 2026 Task 2. The system uses a pretrained Efficient Audio Transformer (EAT) backbone to extract audio token representations from mel-spectrograms. A token-level multi-frame-scale branch with gated bottleneck convolution (MFS/GBC) is used to enhance local time-frequency structures, and LoRA is adopted for parameter-efficient adaptation. At inference time, normal training embeddings are stored in a beam-subband memory bank, and anomaly scores are computed with a nearest-neighbor backend. On the development set, the system obtains 69.28% mean AUC, 75.21% source AUC, 63.35% target AUC, 58.05% pAUC, and a 63.48% official score.

Index Terms— Anomalous sound detection, EAT, multi-scale token fusion, LoRA, kNN

1. INTRODUCTION

Anomalous sound detection (ASD) aims to identify abnormal machine conditions from audio recordings. In DCASE 2026 Task 2, systems are trained mainly with normal samples and evaluated under source/target domain mismatch, machine-category diversity, and noisy recording conditions [1]. These conditions make the task difficult because abnormal examples are not directly available for supervised learning, while normal sounds can vary substantially across domains and machine types. The task setting is closely related to benchmark datasets and evaluation scenarios such as ToyADMOS2 [2], MIMII DG [3], and first-shot anomaly detection for machine condition monitoring [4].

Our submission focuses on robust feature representation and simple distance-based anomaly scoring. Instead of training a detector from scratch, we use a large-scale pretrained EAT model [6] as the acoustic backbone. EAT token maps provide local time-frequency representations that are useful for detecting subtle changes in machine sound. To better capture both short transient events and longer operating patterns, we add a token-level multi-scale fusion branch before anomaly scoring.

2. PROPOSED SYSTEM

The proposed system contains four components: audio preprocessing, EAT feature extraction, Token-MFS/GBC fusion, and memory-bank-based anomaly scoring. Each waveform is converted into a mel-spectrogram and passed to the pretrained EAT backbone. The backbone provides token maps at multiple temporal resolutions. These token maps are fused by the MFS/GBC branch, which combines multi-scale information and produces a compact embedding for each input recording.

The Token-MFS/GBC branch is designed to preserve the stable representation of the pretrained EAT model while adding local multi-scale information. Shorter temporal views are useful for impulsive or rapidly changing sounds, while longer views are more suitable for stationary machine operation. The gated bottleneck convolution module combines these views and outputs a residual correction to the reference token representation. The final embedding is obtained by pooling the fused token map.

For adaptation, the system uses LoRA [7] on the last EAT attention blocks. The main EAT backbone is kept frozen, while LoRA parameters, the fusion branch, and auxiliary heads are trained. Domain and attribute-related objectives are used to improve the structure of the embedding space. ArcFace loss [8], consistency regularization, and distillation are included during training. AdamW [9] is used for optimization, and the reported checkpoint is selected by an internal validation criterion based on normal training data.

During inference, embeddings from normal training samples are stored in a beam-subband memory bank. The backend keeps both global and subband-level information, which helps represent local frequency-band changes. For each test recording, the anomaly score is computed from the nearest-neighbor distance to the normal memory bank [10]. Larger distances indicate higher abnormality. The system outputs continuous anomaly scores for evaluation, and no abnormal training samples are used to set the detector boundary.

3. EXPERIMENTAL RESULTS

Table 1 shows the development-set results. All values are reported in percent. The official score is computed from source AUC, target AUC, and pAUC using the official harmonic-

mean aggregation. The average official score of the submitted EAT token-MFS/GBC beam system is 63.48%.

Table 1: Development-set results of the EAT token-MFS/GBC beam system.

Machine	AUC	AUC source	AUC target	pAUC	Official score
ToyCar	77.07	84.96	69.18	62.47	71.04
ToyCarEmu	66.05	69.54	62.56	53.00	60.93
bearingEmu	61.07	64.04	58.10	60.84	60.90
fan	62.94	77.06	48.82	53.68	57.60
gearboxEmu	72.92	79.64	66.20	57.79	66.72
sliderEmu	57.52	65.42	49.62	49.47	53.91
valveEmu	87.37	85.80	88.94	69.11	80.28
Average	69.28	75.21	63.35	58.05	63.48

The best performance is obtained on valveEmu, followed by ToyCar and gearboxEmu. These results indicate that the pretrained EAT token representation and the multi-scale fusion branch can separate normal and abnormal samples well for several machine types. The main weakness appears in fan and sliderEmu, where target-domain AUC and pAUC are relatively low. This suggests that target-domain mismatch and low-false-positive ranking remain the main bottlenecks for the current system.

4. CONCLUSION

We presented an EAT token-MFS/GBC beam system for DCASE 2026 Task 2. The system combines pretrained audio token representations, parameter-efficient LoRA adaptation, multi-scale token fusion, and a beam-subband nearest-neighbor memory bank. The development-set official score is 63.48%. Future work will focus on stronger domain calibration, improved score normalization, and more stable subband weighting for machine types with weak target-domain performance.

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 e-prints: 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 Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 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 Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), 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 Proceedings of the 31st European Signal Processing Conference (EUSIPCO), 2023, pp. 191–195.

[5] DCASE Challenge, "DCASE 2026 Challenge Website," 2026. https://dcase.community/challenge2026/.

[6] W.-T. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in Proceedings of the 33rd International Joint Conference on Artificial Intelligence, 2024.

[7] 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 International Conference on Learning Representations, 2022.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[9] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in International Conference on Learning Representations, 2019.

[10] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.