SATLAB SYSTEM FOR DCASE 2026 TASK 2: GENERATIVE DATA AUGMENTATION FOR MACHINE SOUND ANOMALY DETECTION

Technical Report

Wenrui Liang[1], *Tianyu Liu*[1], *Xinhu Zheng*[2], *Anbai Jiang*[1], *Shuwei Zhang*[1], *Pingyi Fan*[1]
Cheng Lu[3], *Yanmin Qian*[2], *Xie Chen*[2], *Jia Liu*[1], *Wei-Qiang Zhang*[1]

[1]Tsinghua University, Beijing, China
[2]Shanghai Jiao Tong University, Shanghai, China
[3]North China Electric Power University, Beijing, China
E-mail: lwr24@mails.tsinghua.edu.cn

ABSTRACT

This report describes the SATLab submission to DCASE 2026 Challenge Task 2 on noise-aware anomalous sound detection. The central component of our system is a generative data augmentation pipeline: we train a diffusion-based audio generator from scratch on the challenge data to synthesize samples of rare working conditions, and combine the screened synthetic samples with the real recordings to alleviate the data imbalance and the source–target domain gap. On top of this augmented training set, the submitted systems use BEATs-based audio representations, Mahalanobis and nearest-neighbor anomaly scoring, and score-level ensemble selection. Four systems are submitted, including one single scoring system and three ensemble systems. The best submitted system achieves a development-set harmonic mean of 69.43%.

***Index Terms*—** Anomalous Sound Detection, Generative Data Augmentation, Diffusion Model, Ensembling

1. INTRODUCTION

The DCASE 2026 Challenge Task 2 [1] addresses noise-aware anomalous sound detection for machine condition monitoring. The task is related to prior machine-sound anomaly detection datasets and baselines, including ToyADMOS2 [2], MIMII DG [3], and first-shot anomaly detection for machine condition monitoring [4]. Prior systems for first-shot anomalous sound detection have shown the value of pre-trained audio representations, score fusion, and data augmentation [5, 6, 7, 8, 9]. In particular, self-supervised audio models fine-tuned on machine sounds have become a strong backbone for this task [10, 11, 12, 13, 14].

A persistent difficulty in this task is the imbalance of the training set across working conditions and the domain gap between the source and target domains: normal recordings for some operating conditions are scarce, which limits how well distance-based detectors can characterize the normal distribution. Motivated by recent progress on generative audio modeling, we make *generative data augmentation* the core of our system. We train a diffusion-based generator from scratch on the challenge data, use it to synthesize normal samples for under-represented working conditions, and screen the synthetic samples before adding them to the training set. This augmentation is the main factor distinguishing the SATLab submission from a plain BEATs [15] scoring baseline: all submitted systems are trained on the augmented set, while the rest of the pipeline (representation, scoring, and fusion) is kept deliberately standard so that generative augmentation remains the defining feature of the submission.

2. SYSTEM OVERVIEW

The SATLab pipeline has three stages. First, a generative augmentation module enriches the training set with synthetic samples of rare working conditions: a diffusion-based audio generator is trained on the challenge data, used to synthesize samples of under-represented working conditions, and the screened synthetic samples are combined with the real recordings to form the augmented training set (Section 3). Second, a single self-supervised backbone, BEATs [15], is fine-tuned on this augmented training set to provide acoustic representations (Section 4). Third, anomaly scores are computed from the BEATs embeddings with distance-based scoring methods, including Mahalanobis and nearest-neighbor variants, and the submitted ensemble systems combine multiple scoring branches by weighted score-level fusion selected on the development set (Sections 5 and 6).

A notable difference from our DCASE 2025 submission [5, 6], which adapted several self-supervised backbones in parallel (BEATs [15], EAT [16], and a self pre-trained model) and combined dozens of single models, is that the present system relies on a *single* BEATs backbone. All submitted systems are trained on the generatively augmented set, and the diversity of the ensemble comes from the scoring variants and the fine-tuning seeds rather than from the backbone. Accordingly, this report keeps the representation and scoring details at a high level and focuses on the generative augmentation stage.

3. GENERATIVE DATA AUGMENTATION

Generative data augmentation is the central component of the SATLab system. Recent works have shown that synthesizing rare samples with diffusion-based models is effective for anomalous sound detection [5, 9, 17, 18]. Following this line, we train a diffusion-based audio generator *from scratch* on the challenge data, rather than fine-tuning a large pre-trained text-to-audio model, so that the generator stays close to the machine-sound domain and can be conditioned directly on the working-condition labels of interest. Since this year's dataset provides both near-field and far-field recordings,

we train the generator on both channels.

3.1. Conditional Diffusion Generator

The generator operates on a time–frequency representation of machine audio and is trained with the standard forward-noising and reverse-denoising procedure of diffusion models [19, 20]. Generation is made controllable by conditioning the model on the working-condition attributes together with the recording channel (the near-field or far-field microphone), so that samples can be drawn for a chosen machine type, operating condition, and channel. The generator targets the under-represented working conditions most responsible for the source–target domain gap, and for each targeted condition we synthesize both near-field and far-field samples.

3.2. Sample Screening

Generated samples are not added to the training set directly. We apply a screening step in which an auxiliary classifier, trained on the original data, predicts the attribute label and confidence of each generated clip. A sample is retained only when its predicted label is consistent with the intended condition and its confidence falls in a moderate range, so that the retained samples are both label-consistent and sufficiently diverse rather than near-duplicates of the real data. This screening trades off fidelity against diversity and is important for the synthetic data to act as a useful supplement to the real recordings.

3.3. Augmented Training

The screened synthetic samples are merged with the real recordings to form the augmented training set on which the BEATs backbone is fine-tuned (Section 4). The augmentation chiefly benefits machine types and conditions whose target-domain data are scarce, where the additional synthetic samples help the distance-based detectors form a more complete model of the normal distribution. All submitted systems use this augmented training set, so the augmentation underlies every scoring branch in the ensemble rather than being an optional variant.

4. BEATS REPRESENTATION

4.1. Self-Supervised Backbone

Following the success of self-supervised learning (SSL) models on machine anomalous sound detection [10, 11, 12, 13], we adopt BEATs [15] as the acoustic backbone of the system. BEATs is a Transformer audio encoder pre-trained on large-scale general audio with acoustic tokenizers; its pre-trained representation transfers well to machine sounds and serves as the SSL baseline on which the rest of our pipeline is built. Whereas our previous systems fused multiple SSL backbones to gain diversity, in this submission we deliberately keep BEATs as the only backbone and instead obtain diversity from the scoring stage and the data augmentation, so that the representation is a stable reference across all submitted systems.

4.2. Fine-tuning

The pre-trained backbone is general-purpose, so we adapt it to machine sounds by fine-tuning it as an attribute classifier on the challenge data, where the classification label is formed from the machine type, section, the available operating-condition attributes, and the recording channel, so that the near-field and far-field versions of the same condition form distinct classes and the backbone is encouraged to encode the channel difference rather than collapse it. An attentive statistical pooling layer is appended to the backbone to produce an utterance-level embedding, and the backbone is fine-tuned end-to-end on the generatively augmented training set described in Section 3.

5. ANOMALY SCORING

Given a fine-tuned BEATs backbone, anomaly scores are computed in the embedding space with distance-based detectors. For each machine type and section we extract the embeddings of the normal training samples and store them in memory banks, keeping the source-domain and target-domain samples separate so that the scarcity of target-domain data does not bias the source-domain statistics. Two scoring variants are used. The nearest-neighbor (KNN) variant scores a test clip by its distance to the nearest stored normal embedding, while the Mahalanobis variant scores it against a normal distribution estimated from the stored embeddings, which accounts for the covariance of the normal class. The two variants are complementary, and both are kept as branches for the ensemble.

6. SUBMITTED SYSTEMS

All submitted systems are built on BEATs backbones fine-tuned on the generatively augmented training set. They differ along two axes: the scoring variant (Mahalanobis or KNN) and the fine-tuning seed, together with the number of branches that are fused. Score-level fusion linearly combines the anomaly scores of the selected branches, and the fusion coefficients are chosen on the development set. We submit four systems:

- **System 1**: a single BEATs/Mahalanobis scoring system, serving as the single-branch baseline.
- **System 2**: a seven-subsystem KNN-based ensemble system.
- **System 3**: a nine-subsystem Mahalanobis ensemble system.
- **System 4**: a twelve-subsystem ensemble combining Mahalanobis and KNN-based scoring branches.

System 1 uses a single scoring branch, whereas Systems 2–4 fuse increasingly many branches and combine the two scoring variants, so that the comparison across the four systems reflects the effect of the scoring fusion and the ensemble size on top of the shared augmented backbone.

7. EXPERIMENT RESULTS

The submitted systems are trained on the generatively augmented training set and evaluated on the DCASE 2026 Task 2 development set. Table 1 presents the per-machine and overall results of the four submitted systems, where the overall performance is summarized by the harmonic mean (hmean) over all machine-level AUC and pAUC values. The best performing system, System 4, which fuses the Mahalanobis and KNN scoring branches, achieves an overall hmean of 69.43% on the development set.

Table 1: Development-set results (%) of the SATLab submitted systems.

Machine	Metric	System 1	System 2	System 3	System 4
bearingEmu	AUC_s	66.18	68.12	65.32	67.24
	AUC_t	60.94	66.32	64.24	64.20
	pAUC	57.16	61.89	56.79	60.84
	hmean	61.21	65.34	61.88	63.99
fan	AUC_s	72.26	88.60	79.98	90.16
	AUC_t	60.86	69.74	66.26	71.62
	pAUC	57.05	62.84	59.58	65.79
	hmean	62.76	72.22	67.60	74.53
gearboxEmu	AUC_s	80.30	78.40	78.80	79.84
	AUC_t	86.08	80.52	83.08	82.32
	pAUC	72.42	66.79	71.42	68.74
	hmean	79.20	74.73	77.46	76.49
sliderEmu	AUC_s	59.28	61.34	60.90	62.02
	AUC_t	57.78	66.86	62.06	66.76
	pAUC	50.37	51.53	50.95	51.84
	hmean	55.53	59.21	57.51	59.53
ToyCar	AUC_s	69.42	68.08	74.38	72.16
	AUC_t	80.04	82.68	78.72	80.34
	pAUC	62.68	64.05	62.79	63.58
	hmean	70.01	70.76	71.30	71.37
ToyCarEmu	AUC_s	81.60	61.90	82.02	71.38
	AUC_t	90.86	91.38	92.26	94.08
	pAUC	66.37	51.53	67.79	54.21
	hmean	78.27	64.51	79.40	69.63
valveEmu	AUC_s	68.00	74.14	70.14	73.32
	AUC_t	71.42	85.50	77.88	85.54
	pAUC	53.00	67.53	54.11	65.58
	hmean	63.06	75.01	65.82	73.93
Overall	AUC_s	71.01	71.51	73.08	73.73
	AUC_t	72.57	77.57	74.93	77.84
	pAUC	59.86	60.88	60.49	61.51
	hmean	66.15	68.37	67.92	**69.43**

Overall AUC_s, AUC_t, and pAUC are arithmetic averages over machine types. Overall hmean is computed as the harmonic mean over all machine-level AUC and pAUC values.

8. CONCLUSION

This report presented the SATLab submission to DCASE 2026 Task 2. The core of the system is a generative data augmentation pipeline built around a diffusion-based generator trained from scratch, which synthesizes and screens samples of rare working conditions to alleviate the training-set imbalance and the source–target domain gap. On top of the augmented training set, the systems combine BEATs representations, Mahalanobis and KNN-based scoring, and score-level fusion. The best submitted system achieves a development-set harmonic mean of 69.43%.

9. 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 preprint 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 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)*, Nancy, France, November 2022.

[4] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Ya-

suda, "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] A. Jiang, W. Liang, S. Feng, Y. Qiu, Y. Zhao, J. Li, P. Fan, W.-Q. Zhang, C. Lu, X. Chen, Y. Qian, and J. Liu, "Thuee system for dcase 2025 anomalous sound detection challenge," DCASE2025 Challenge, Tech. Rep., June 2025.

[6] X. Zheng, A. Jiang, B. Han, S. Zhang, W.-Q. Zhang, X. Chen, C. Lu, P. Fan, J. Liu, and Y. Qian, "Sjtu-aithu system for dcase 2025 anomalous sound detection challenge," DCASE2025 Challenge, Tech. Rep., June 2025.

[7] Z. Lv, A. Jiang, B. Han, Y. Liang, Y. Qian, X. Chen, J. Liu, and P. Fan, "Aithu system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[8] A. Jiang, X. Zheng, Y. Qiu, W. Zhang, B. Chen, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Thuee system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[9] W. Liang, Y. Qiu, A. Jiang, B. Han, T. Liu, X. Zheng, P. Fan, C. Lu, J. Liu, and W.-Q. Zhang, "Refgen: Reference-guided synthetic data generation for anomalous sound detection," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 15 877–15 881.

[10] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen, Y. Qian, J. Liu, and P. Fan, "Anopatch: Towards better consistency in machine anomalous sound detection," in *Interspeech 2024*, 2024, pp. 107–111.

[11] X. Zheng, A. Jiang, B. Han, Y. Qian, P. Fan, J. Liu, and W.-Q. Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 969–974.

[12] A. Jiang, X. Zheng, B. Han, Y. Qiu, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Adaptive prototype learning for anomalous sound detection with partially known attributes," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[13] B. Han, A. Jiang, X. Zheng, W.-Q. Zhang, J. Liu, P. Fan, and Y. Qian, "Exploring self-supervised audio models for generalized anomalous sound detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 4126–4141, 2025.

[14] P. Fan, A. Jiang, S. Zhang, X. Zheng, Z. Lv, B. Han, W. Liang, J. Li, W.-Q. Zhang, Y. Qian, X. Chen, and J. Liu, "FISHER: A foundation model for multimodal industrial signal comprehensive representation," *IEEE Transactions on Industrial Informatics*, pp. 1–12, 2026.

[15] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 5178–5193.

[16] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 3807–3815.

[17] 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)*. IEEE, 2024, pp. 1271–1275.

[18] J. Yin, Y. Gao, W. Zhang, T. Wang, and M. Zhang, "Diffusion augmentation sub-center modeling for unsupervised anomalous sound detection with partially attribute-unavailable conditions," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[20] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations (ICLR)*, 2021.