VUI LABS SYSTEM FOR DCASE 2026 TASK 2: DUAL-CHANNEL BEATS FUSION FOR MACHINE SOUND ANOMALY DETECTION

Technical Report

Yanmin Qian[1], *Xinhu Zheng*[1,2], *Anbai Jiang*[3], *Wenrui Liang*[3], *Tianyu Liu*[3], *Shuwei Zhang*[3]
Xie Chen[2], *Cheng Lu*[4], *Wei-Qiang Zhang*[3], *Pingyi Fan*[3], *Jia Liu*[3]

[1] VUI Labs, Shanghai, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Tsinghua University, Beijing, China
[4] North China Electric Power University, Beijing, China
Email: yanminqian@sjtu.edu.cn, zhengxh24@sjtu.edu.cn

ABSTRACT

This report describes the VUI Labs submission to DCASE 2026 Challenge Task 2 on noise-aware anomalous sound detection. The submission focuses on dual-channel modeling for paired near-field and far-field machine recordings. We combine single-channel self-supervised representations with channel-interaction variants that use reference-channel information to stabilize noisy-channel anomaly scoring. Four systems are submitted, including two single-model systems and two score-level fusion systems. The best submitted system achieves a harmonic mean of 66.39% on the development dataset.

***Index Terms*—** DCASE Challenge, anomaly detection, dual-channel modeling, pre-trained models, noise-robust detection

1. INTRODUCTION

The DCASE 2026 Challenge Task 2 [1] focuses on noise-aware anomalous sound detection (ASD) under dual-channel recording conditions. The task builds on prior machine-sound ASD datasets and baselines, including ToyADMOS2 [2], MIMII DG [3], and first-shot ASD for machine condition monitoring [4]. Each machine provides paired near-field and far-field recordings, simulating practical deployment scenarios where microphones observe the same machine from different acoustic positions.

The key emphasis of this year's task is therefore not only domain generalization across machine sections, but also noise-aware scoring from paired channels. A robust system should exploit the relatively reliable near-field evidence and the noisier far-field evidence without allowing either channel to dominate the final anomaly decision.

Compared with a single-channel benchmark, this paired-channel setting introduces additional modeling issues:

- **Channel reliability mismatch**. The two channels describe the same machine event, but their reliability can differ from machine to machine and from recording to recording.
- **Far-field noise interference**. The far-field channel may contain environmental sound and propagation effects that are not directly related to machine condition.
- **Decision stability**. The anomaly detector must prevent channel-dependent artifacts from dominating the final anomaly decision.

The paired-channel setting changes the role of representation learning. A model must detect anomalies from both channels, but the two channels are not equally reliable. The near-field signal usually preserves machine-specific structure, while the far-field signal is more affected by environmental noise and propagation effects. A robust system should therefore use the two channels jointly without forcing them to be identical.

The VUI Labs submission emphasizes this channel-aware view. We use self-supervised audio representations [5, 6] as the base acoustic encoder and build several dual-channel variants on top of the encoder. The channel-interaction modules are kept lightweight and are used to exchange reference information between the paired recordings. The final systems combine single-channel and dual-channel evidence at the score level, so that channel-specific errors do not dominate the submitted score.

The remainder of this report is organized as follows. Section 2 outlines the dual-channel methodology. Section 3 describes the four submitted systems. Section 4 presents development-set results.

2. METHODS

2.1. Pre-trained Audio Backbones

We leverage self-supervised pre-trained audio models to extract acoustic representations from machine recordings. The submitted systems use BEATs [5] as the external pre-trained model. The audio waveform is segmented into fixed-duration clips and converted to spectral features. Spectral augmentation [7] is applied during training to reduce sensitivity to local time-frequency corruption.

The backbone outputs are pooled into utterance-level embeddings. These embeddings are optimized with proxy labels derived from machine identity, section, and available condition information. Prior ASD work has shown that adapting pre-trained audio models is an effective way to improve generalized anomaly detection [8, 9, 10, 11]; here, we extend this idea to a paired-channel benchmark.

Table 1: Results of four submitted systems on the development set

Machine	Metric	System 1	System 2	System 3	System 4
bearingEmu	AUC_s	60.28	63.88	65.98	64.46
	AUC_t	61.66	63.34	65.40	63.30
	pAUC	54.42	58.84	58.74	60.00
	hmean	58.61	61.94	63.20	62.53
fan	AUC_s	85.76	55.66	75.10	59.72
	AUC_t	71.18	66.20	67.54	62.42
	pAUC	61.63	54.05	64.16	57.21
	hmean	71.54	58.17	68.64	59.71
gearboxEmu	AUC_s	75.64	67.30	76.06	69.14
	AUC_t	74.22	64.20	75.02	65.06
	pAUC	62.05	55.26	63.32	54.58
	hmean	70.08	61.82	70.97	62.30
sliderEmu	AUC_s	57.96	57.40	62.06	58.32
	AUC_t	61.60	59.64	63.80	61.06
	pAUC	51.84	51.89	50.53	50.37
	hmean	56.84	56.12	58.16	56.20
ToyCar	AUC_s	67.50	70.68	72.10	70.88
	AUC_t	66.58	76.00	75.34	80.10
	pAUC	57.95	62.42	63.89	64.05
	hmean	63.71	69.24	70.10	71.08
ToyCarEmu	AUC_s	57.58	71.70	72.14	71.22
	AUC_t	85.44	74.40	83.60	74.36
	pAUC	49.21	55.47	57.42	53.68
	hmean	60.74	66.06	69.38	65.05
valveEmu	AUC_s	63.14	64.72	67.34	66.36
	AUC_t	70.34	74.36	77.58	76.68
	pAUC	50.47	56.84	57.11	61.21
	hmean	60.16	64.53	66.30	67.50
Overall	AUC_s	66.84	64.48	70.11	65.73
	AUC_t	70.15	68.31	72.61	69.00
	pAUC	55.37	56.40	59.31	57.30
	hmean	62.68	62.27	**66.39**	63.15

Results are reported for the submitted file-level scores. Overall AUC_s, AUC_t, and pAUC are arithmetic means over the listed machines, while Overall hmean is the official harmonic mean. AUC_s and AUC_t denote source and target domain AUC, respectively.

2.2. Dual-Channel Representation

For each recording pair, the model receives a near-field channel and a far-field channel. We process the two channels with shared or partially shared encoders, depending on the branch. Shared encoding keeps the representation space aligned across channels, while channel-interaction branches model channel-dependent noise patterns.

The dual-channel branches use lightweight interaction modules after the backbone feature extraction stage. In these branches, one channel provides a reference for the other channel. The interaction can be viewed as a feature-space conditioning step rather than a waveform enhancement step. This choice avoids reconstructing clean audio and keeps the anomaly detector focused on representation quality.

We treat the near-field channel as a stable reference rather than as a replacement for the far-field channel. The goal is not to remove all channel differences, because some differences are useful for judging whether a recording is reliable. Instead, the dual-channel representation is encouraged to preserve shared machine-condition information and suppress channel-specific disturbances. This design also keeps the output compatible with the same distance-based scoring backend used by the single-channel system.

2.3. Gating and Attention Variants

We explore multiple channel-interaction variants. One group uses soft gating to regulate how much information from each channel contributes to the final embedding. Another group uses attention-like interaction [12, 13] to exchange contextual information between channel features.

At the system level, these variants serve the same purpose: they reduce the effect of unreliable channel evidence while preserving

machine-specific cues. This is especially important for low false-positive-rate pAUC, where a small number of noisy normal recordings can strongly affect the final metric.

The gating variants are used to adjust the contribution of each channel at the representation level. The attention-style variants are used to align channel evidence before score computation. Both groups are designed as small additions on top of the pre-trained encoder, so the main acoustic representation remains driven by self-supervised pre-training and task adaptation. This keeps the channel modeling focused on reliability estimation rather than on learning a new acoustic model from scratch.

2.4. Anomaly Detection and Fusion

After training, anomaly scores are computed through nearest-neighbor distance-based methods in the learned embedding space. The normal training samples form a machine-wise reference set. For each test recording, the detector compares its embedding with the reference set and converts the distance evidence into an anomaly score.

The final fusion systems combine selected single-channel and/or dual-channel branches by weighted score averaging. The reported fusion composition follows the branches retained in the submitted weighted average. Since the branches make different errors across machine types, score-level fusion provides a simple mechanism to preserve useful single-channel behavior while adding channel-aware robustness.

All branches produce file-level anomaly scores, allowing the same evaluation interface to be used for single-channel, dual-channel, and mixed fusion systems.

3. SUBMITTED SYSTEMS

We submit four systems, consisting of two single-model systems and two score-level fusion systems.

- **System 1**: a single BEATs backbone system with ArcFace fine-tuning and KNN-style anomaly scoring.
- **System 2**: a single dual-channel Gumbel-gating system [14].
- **System 3**: a fusion system composed of a single-channel BEATs ArcFace reference, a dual-channel co-attention branch, and a dual-channel Gumbel-gating branch.
- **System 4**: a compact fusion system composed of the dual-channel co-attention and Gumbel-gating BEATs branches.

The four systems are organized as a progression. System 1 is the single-channel reference system and fixes the basic representation and scoring interface. System 2 introduces explicit dual-channel gating while keeping a single-model submission. System 3 combines the single-channel reference with co-attention and Gumbel-gating branches. System 4 removes the single-channel reference and focuses on the two retained dual-channel BEATs branches, which isolates the contribution of channel-aware modeling in a compact fusion setting.

4. EXPERIMENT RESULTS

Performance is measured using source-domain AUC (AUC_s), target-domain AUC (AUC_t), partial AUC at 0–0.1 false positive rate (pAUC), and their harmonic mean. Table 1 presents the per-machine breakdown for all four systems.

The result table is intended to compare channel-use strategies rather than individual architectural details. System-level differences should be interpreted together with the submitted-system descriptions: the single-channel reference establishes the baseline setting, while the dual-channel and mixed systems test whether paired recordings provide more stable anomaly evidence under the same evaluation metrics.

5. CONCLUSION

This report presented the VUI Labs submission to DCASE 2026 Task 2. The submission emphasizes dual-channel machine sound modeling, combining single-channel self-supervised representations with channel-interaction branches. The best score-fusion system attains a harmonic mean of 66.39% on the development set.

6. 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] 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. 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.

[5] 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," *arXiv preprint arXiv:2212.09058*, 2022.

[6] 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.

[7] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of Interspeech 2019*, 2019, pp. 2613–2617.

[8] 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.

[9] 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.

[10] 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.

[11] 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*, 2025.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[13] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[14] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *International Conference on Learning Representations (ICLR)*, 2017.