ROBUST TRAIN-NORMAL CALIBRATION FOR FIRST-SHOT MACHINE SOUND ANOMALY DETECTION

Technical Report

Peihong Zhang *Shengchen Li*

Xi'an Jiaotong-Liverpool University
Suzhou, China
{Peihong.Zhang20@student.xjtlu.edu.cn, Shengchen.Li@xjtlu.edu.cn}

ABSTRACT

DCASE 2026 Task 2 evaluates first-shot unsupervised anomalous sound detection under source-domain, target-domain, and low-FPR criteria. In this setting, our experiments indicate that the dominant failure mode is not simply insufficient representation capacity, but unstable normal-score tails and fragile source/target operating conditions. We therefore submit a frozen normality-calibration portfolio: complementary pretrained audio scores are normalized only with train-normal references, guarded by conservative tail and domain-risk reliability checks, and evaluated with an exact replica of the official development metric. The primary system improves a strong proxy score from 0.611706 to 0.613941 official harmonic mean while preserving mean pAUC at 0.592030. Additional submitted slots provide an out-of-fold mirror, an independent Dasheng guarded variant, and a global-gating fallback for target false-positive risk. Negative stress tests are also informative: larger resource ensembles, adaptive local-density normalization, and localized temporal matching improved some AUC or target-risk indicators but degraded low-FPR pAUC, so they were excluded. The resulting submission emphasizes a practical conclusion for first-shot ASD: robust train-normal calibration and normal-tail preservation can be more valuable than adding another high-capacity detector.

***Index Terms*—** Anomalous sound detection, machine condition monitoring, DCASE 2026, train-normal calibration, score ensemble

1. INTRODUCTION

DCASE 2026 Task 2 addresses first-shot unsupervised anomalous sound detection (ASD) for machine condition monitoring in a noise-aware setting [1]. For each machine section, participants are given only normal training clips from source and target domains. The development dataset follows the first-shot protocol: each section contains 990 source-domain normal clips and 10 target-domain normal clips for training, while the corresponding test split contains normal and anomalous clips from both domains [2]. The central difficulty is therefore asymmetric: the system must model a target machine state from very few target-normal examples while remaining stable in both source and target domains and at low false-positive rates.

Industrial ASD has been studied in DCASE through datasets and tasks such as ToyADMOS, MIMII, ToyADMOS2, MIMII DG, and recent first-shot/domain-shift tracks [3, 4, 5, 6, 7, 8, 9]. Recent systems often rely on large pretrained audio representations, density scores, and ensembles. Our experiments support this direction, but also reveal a limit: many stronger-looking feature branches increase mean AUC while damaging the low-FPR tail that drives pAUC. We therefore treat ASD here as a calibrated normality-ranking problem rather than as a search for a single more expressive encoder.

The report makes three practical observations. First, difficult target-domain and low-FPR conditions, not the average score alone, determine whether an ASD system is useful. Second, train-normal tail calibration and independent reliability guards provide a small but reproducible gain over the proxy score without training on anomaly labels. Third, negative evidence matters: multi-channel relation learning, local-density normalization, large resource ensembles, and temporal localized matching did not beat the calibrated portfolio once low-FPR and source/target robustness were considered. The final submission freezes four systems that reflect these observations rather than the most complex models tried.

2. TASK METRIC AND COMPLIANCE PROTOCOL

Let $\mathcal{C}$ denote the set of metric cells used by the official Task 2 score. For each machine type and section, the cells include source-domain AUC, target-domain AUC, and pAUC over the low false-positive-rate region $[0, 0.1]$. The official development score is the harmonic mean

$$H = \frac{|\mathcal{C}|}{\sum_{c \in \mathcal{C}} 1/m_c}, \tag{1}$$

where m_c is the official AUC or pAUC value for cell c. Our implementation used a numerical safeguard only in unit tests and not in the reported values. We verified the evaluator by monotonic-transform tests, sign-inversion tests, tie-handling tests, and comparison with the trusted project metric implementation. All scores are oriented so that a larger value means more anomalous.

The following compliance rules were enforced throughout final selection. The evaluation set and submission outputs were not read during the audit or final engineering stages. Evaluation filenames, order, directory layout, score distributions, or batch statistics were not used for calibration. Development anomaly labels were used only for offline metric evaluation and for choosing among already-frozen candidate systems. The OOF, LOMO, and nested protocols fixed their splits and train-normal calibration before any metric was computed. No detector, normalizer, threshold, gating rule, or fusion weight was trained using development anomaly labels. Decision-result thresholds are generated from train-normal score scales and are diagnostic only; challenge ranking is determined by continuous anomaly scores.

3. RELIABILITY-AWARE NORMALITY PORTFOLIO

The submitted systems implement one idea: preserve a strong normality ranking while preventing unreliable normal-score tails from driving false alarms. Each branch maps an audio clip to one or more frozen normality scores. The audio is decoded deterministically and resampled to the native rate of the corresponding frontend or encoder. The first channel is treated as the near target-machine channel and the second channel as the farther reference channel when a branch uses two-channel views. Acoustically, the near channel is expected to contain stronger target-machine energy, while the far channel is more sensitive to propagation loss, background noise, reverberation, and operating-environment mismatch. Large disagreement between near-, far-, or residual-view scores is therefore treated as unreliable anomaly evidence rather than as a direct anomaly label. There is no test-batch channel adaptation.

Existing pretrained audio representations provide complementary acoustic cues: PANNs-style acoustic event embeddings [10], EAT self-supervised transformer embeddings [11], BEATs-style acoustic-token representations [12], CLAP-style audio-language embeddings [13], and Dasheng masked-audio-encoder embeddings [14]. These resources were audited in the project resource registry before being used; only cached checkpoints and feature dumps with recorded source and hash were allowed into final candidate scoring. The final freeze reads score CSVs rather than raw checkpoints, but the upstream cached feature families include `panns_cnn14`, `beats_audioset`, `laion_clap`, `EAT`, and `Dasheng`; the accepted score artifacts and feature inventories are SHA256-recorded. Earlier screening also considered other audio transformer families such as AST, PaSST, and AudioMAE [15, 16, 17]. In the final systems these representations are not trained on anomalies; they are used as frozen feature or score sources.

The resulting inference recipe is deliberately simple. First, audio is decoded and channel views are formed according to the frozen branch configuration. Second, each branch computes or reads a frozen clip score for the matching machine, section, and domain-aware reference setting. Third, branch scores are robustly normalized with train-normal statistics from the same fold or final train split. Fourth, fixed score transforms and fusion weights from the frozen configuration are applied. Fifth, a reliability guard subtracts only a small penalty when train-normal calibrated branch disagreement or tail risk is high. Finally, the continuous anomaly score is written with full precision, and the binary decision result is obtained by a fixed train-normal empirical-CDF threshold of 0.95.

3.1. Train-normal score calibration

For a branch score $s_b(x)$ and the corresponding normal training reference set $\mathcal{T}$, robust score calibration is performed using statistics estimated only from $\mathcal{T}$:

$$z_b(x) = \frac{s_b(x) - \mathrm{median}_{u\in\mathcal{T}} s_b(u)}{\mathrm{IQR}_{u\in\mathcal{T}} s_b(u) + \epsilon}. \tag{2}$$

When source and target normal banks are available, calibration keeps their statistics fold-safe and never estimates normalization parameters from test clips. Candidate variants may additionally use train-normal empirical tail references, branch disagreement, or target-bank proximity as reliability cues. These cues are used only as small conservative guards; they are not learned anomaly classifiers.

3.2. Tail calibration and guards

The primary family combines a frozen proxy normality score with an EAT-based reliability guard. The proxy score is a scalar normality score built from train-normal calibrated embedding-distance and density-style branches. Its reference set is the available normal training clips for the corresponding fold, machine, section, and domain setting. The EAT guard is an independent frozen representation check that identifies clips whose proxy score is in the upper train-normal tail but whose EAT-based normality evidence is weak or inconsistent. Tail calibration then applies a deterministic penalty on the train-normal empirical-tail scale. Operationally, the correction has the form

$$S(x) = \sum_b w_b\, \phi_b(z_b(x)) - \lambda R(x), \tag{3}$$

where ϕ_b is a fixed monotonic score transform, w_b are frozen branch weights, and $R(x)$ is a nonnegative reliability penalty. In the implemented guard, $R(x)$ is nonzero only when the base score lies in a high train-normal CDF tail and an auxiliary representation score does not support the same anomaly evidence; otherwise the base ranking is preserved. The weights, guard constants, and transforms were fixed by development validation before final packaging; per-section calibration statistics are computed only from normal training clips and were not fitted on evaluation data.

The Dasheng-guarded system adds an independent masked-audio representation guard. Its purpose is not to replace the main score, but to protect against target-domain false positives and acoustically unstable examples where the primary score and an independent SSL representation disagree. A separate global-gating fallback applies a fixed global reliability gate to the score mixture using fold-safe normal references, rather than a local tail guard. It has lower pAUC than the first three systems but the lowest target false-positive ratio among the selected systems.

The target FP ratio in Tables 2 and 3 is a diagnostic, not an official DCASE metric. It is the fraction of development normal clips marked as false positives by the same frozen train-normal empirical-CDF decision threshold used for decision-result generation. The diagnostic uses development normal/domain labels only after scores are produced; it does not fit thresholds, calibrators, or guards.

3.3. Decision result generation

The challenge requires both anomaly-score CSVs and binary decision-result CSVs. Our binary decisions are produced with a fixed train-normal threshold rule on the frozen score scale. They are not tuned with development anomaly labels and are not used to choose the ranking systems. They are included only because the task requires decision-result files.

4. EXPERIMENTAL EVIDENCE AND SELECTION

All candidate systems were evaluated with the official development evaluator. The final four were chosen after forced recompute and a pre-specified fallback audit. The fallback set for the fourth system contained only four already-frozen alternatives: the proxy baseline, the EAT-guarded baseline, the global-gating variant, and a low-correlation harmonic-repair variant. The audit illustrates the main empirical trade-off: the most orthogonal score strongly reduced target false positives but damaged pAUC, whereas the global-gating variant kept a better balance across AUC, pAUC, and target false alarms.

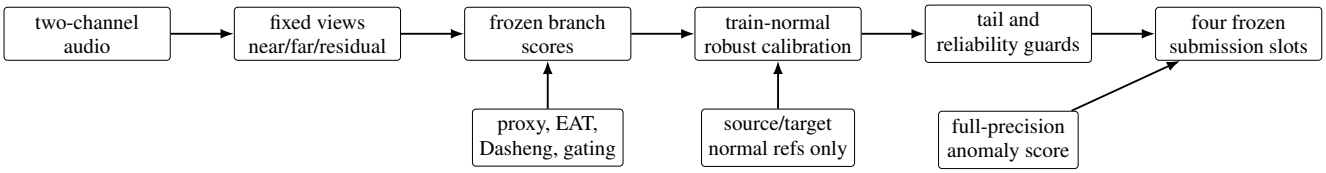

Figure 1: Final inference pipeline. All calibration and reliability controls are estimated from train-normal references; development anomaly labels are used only for offline metric evaluation and final candidate selection.

Table 1: Main accepted components and their leakage controls. "Frozen" means fixed before final evaluation inference.

Component	Input evidence	Role in the portfolio	Leakage control
Proxy normality score	Frozen distance/embedding scores	Stable base anomaly ranking	Train-normal robust scale only
EAT guard	Independent frozen EAT score	Detects unreliable high-tail proxy evidence	No anomaly-label fitting
Tail calibration	Train-normal score tails	Reduces extreme normal false alarms	Per-fold/train normal statistics
Dasheng guard	Independent masked-audio representation	Adds target-risk protection	Fixed branch, no eval statistics
Global gating	Fold-safe global reliability score	Fourth-system target-FP suppression fallback	Pre-specified fallback audit

The primary score family has high correlation across the first three submitted systems, but the systems serve different risk roles. System A is the best direct development system. System B is the OOF mirror and protects against direct-selection sensitivity. System C introduces the Dasheng guard with slightly lower harmonic score but lower target false-positive ratio. System D provides a more globally gated fallback, with the strongest target false-positive suppression among the four submitted systems. This is not an ensemble-for-ensemble's-sake design; each system preserves a different compromise between pAUC, source/target balance, and target-normal false alarms.

Here "best direct" denotes global development-set model selection after all train-normal calibration was fixed; it does not mean that anomaly labels were used to train a detector, threshold, gate, or normalizer. This selection can still be optimistic, so System B keeps the OOF mirror and LOMO checks were used to identify candidates that improved only through direct-development overfitting. After the final freeze, evaluation-set inference is a single pass: no evaluation statistics, labels, filenames, ordering, or score distributions enter any component.

4.1. What the rejected branches show

The rejected branches give a clearer scientific message than a list of failed scores. A multi-channel normal relation model tried to learn near/far joint acoustic structure from normal stereo audio, but nested/fusion evaluation did not exceed the calibrated score portfolio. Other trials were motivated by one-class density estimation and memory-based industrial anomaly detection [18, 19, 20, 21, 22]; however, their local-density corrections often changed the rank order of difficult source or pAUC conditions too aggressively. Additional allowed-resource ensembles reached only 0.598230 development score with best pAUC 0.571504, below the submitted systems. The final temporal/localized matching test improved target AUC and target false-positive ratio, but reduced pAUC, obtaining $H = 0.609381$ and mean pAUC $= 0.580526$; the best ALDN result was $H = 0.555701$. These results suggest that for this task, representation diversity alone is not the bottleneck unless it is coupled with careful tail preservation.

This negative evidence is important for the final design. Local density corrections and aggressive temporal pooling can produce attractive AUC or target-domain behavior, but low-FPR pAUC exposes whether they raise the highest-scoring normal clips. The final portfolio therefore keeps the calibrated tail systems rather than adding a new branch that damages pAUC. In practical terms, we found that a method which rescues a difficult target domain is still harmful if it raises the top normal tail in another machine.

The final selection should therefore be read as a robustness-oriented system choice rather than as a claim that every submitted slot is statistically independent. The first three slots deliberately preserve the strongest pAUC behavior, while the fourth slot trades some pAUC for a different target false-positive profile. This portfolio construction is useful because unseen machine conditions may emphasize different source/target mixtures from the development set.

5. REPRODUCIBILITY AND FINAL ENGINEERING

The final engineering package freezes score files, configurations, evaluator scripts, manifests, software versions, and artifact hashes. Run manifests record random seeds, deterministic settings where applicable, feature manifests, and the Python/PyTorch/scikit-learn/CUDA environment used on NVIDIA RTX A6000 GPUs. The selected score files were forcibly recomputed where scripts and frozen configs were available. The global-gating score matched exactly after recomputation; the tail-calibrated files were accepted after a guarded recompute comparison with maximum absolute difference zero. The final dry-run package validator reported a valid four-system structure, no headers in CSV files, finite scores, no duplicate filenames, and the required anomaly-score and decision-result files for each machine type and section.

Final evaluation inference is performed once from the frozen manifest after all systems, calibrators, and thresholds are fixed. No further model search, calibration, score normalization, or threshold tuning is performed after the freeze. This separation between research selection and final inference is intended to make the submission reproducible and to avoid accidental use of evaluation distribution information.

Table 2: Submitted systems and exact development metrics. The official harmonic mean is computed over AUC and pAUC cells as in (1); pAUC uses $p = 0.1$.

System	Description	H	mean AUC	mean pAUC	source AUC	target AUC	target FP ratio
A	Primary proxy+EAT tail-calibrated score	0.613941	0.628100	0.592030	0.670229	0.628857	0.914286
B	Out-of-fold mirror of System A	0.613916	0.628186	0.591805	0.670229	0.629086	0.914286
C	Dasheng-guarded reliability variant	0.612939	0.627086	0.591353	0.670686	0.624914	0.885714
D	Global-gating target-risk fallback	0.612051	0.628300	0.586015	0.673143	0.625771	0.771429

Table 3: Fourth-system fallback audit. All candidates satisfy the project compliance rules.

Candidate	H	AUC	pAUC	target FP
Proxy baseline	0.611706	0.626486	0.589774	0.971429
EAT-guarded baseline	0.611706	0.626486	0.589774	0.971429
Global-gating variant	0.612051	0.628300	0.586015	0.771429
Low-correlation repair variant	0.607260	0.630743	0.569699	0.714286

Table 4: Most difficult primary-system development conditions. These cases motivated conservative pAUC/tail handling and explain why several higher-AUC variants were rejected.

Cell	Machine	Metric	Value
target AUC	sliderEmu	AUC	0.5056
target AUC	fan	AUC	0.5076
target AUC	bearingEmu	AUC	0.5188
pAUC	sliderEmu	pAUC@0.1	0.5316
pAUC	ToyCar	pAUC@0.1	0.5584
pAUC	bearingEmu	pAUC@0.1	0.5605

6. CONCLUSION

Our DCASE 2026 Task 2 submission is a frozen, train-normal calibrated normality portfolio. The main result is a reliability lesson: in first-shot machine ASD, source/target robustness and normal-tail control can dominate the benefit of adding more encoders or more flexible density models. The best development harmonic score is 0.613941 with mean AUC 0.628100 and mean pAUC 0.592030. The submitted systems therefore emphasize calibrated normality evidence, conservative reliability guards, and exact offline evaluation. Extensive later experiments, including two-channel relation learning, large allowed-resource ensembles, local-density normalization, and temporal localized matching, were excluded because their gains in target AUC or diversity did not survive low-FPR pAUC evaluation. This provides a reproducible and practically useful conclusion: for this task, protecting the ranking of normal tails is at least as important as improving average anomaly separation.

7. ACKNOWLEDGMENT

This work was conducted as an independent DCASE Challenge submission at Xi'an Jiaotong-Liverpool University.

Table 5: Selected candidate comparison. The rows are not a nested training ablation; they compare the main accepted score families under the same development evaluator.

System family	H	mean AUC	mean pAUC
Proxy baseline	0.611706	0.626486	0.589774
Primary calibrated system	0.613941	0.628100	0.592030
Out-of-fold mirror	0.613916	0.628186	0.591805
Dasheng-guarded variant	0.612939	0.627086	0.591353
Global-gating variant	0.612051	0.628300	0.586015

Table 6: Negative final stress-test evidence. These systems were not submitted.

Branch	H	mean AUC	mean pAUC
Best temporal/localized fusion	0.609381	0.628886	0.580526
Best adaptive local-density variant	0.555701	–	–
Best large-resource component	–	–	0.571504

8. 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] ——, "Development Dataset for DCASE 2026 Challenge Task 2," https://zenodo.org/records/19336329, 2026.

[3] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 313–317.

[4] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019, pp. 209–213.

[5] N. Harada, Y. Nishida, T. Komatsu, K. Imoto, and Y. Koizumi, "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.

[6] 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.

[7] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2023, pp. 31–35.

[8] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and Discussion on DCASE 2024 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," *arXiv preprint arXiv:2406.07250*, 2024.

[9] ——, "Description and Discussion on DCASE 2025 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," *arXiv preprint arXiv:2506.10097*, 2025.

[10] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[11] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-Supervised Pre-Training with Efficient Audio Transformer," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[12] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio Pre-Training with Acoustic Tokenizers," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[13] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[14] H. Dinkel, Z. Yan, Y. Wang, J. Wang, Y. Zhang, and B. Wang, "Scaling Up Masked Audio Encoder Learning for General Audio Classification," in *Proceedings of Interspeech*, 2024.

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

[16] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," in *Proceedings of Interspeech*, 2022.

[17] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked Autoencoders that Listen," in *Advances in Neural Information Processing Systems*, 2022.

[18] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[19] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proceedings of the IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[20] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep One-Class Classification," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[21] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards Total Recall in Industrial Anomaly Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2021.