MIMII-Agent: Leveraging LLMs with Function Calling for Relative Evaluation of Anomalous Sound Detection

Harsh Purohit, Tomoya Nishida, Kota Dohi, Takashi Endo, and Yohei Kawaguchi

Hitachi Ltd., R&D Group, Tokyo, Japan

***Abstract*—This paper proposes a method for generating machine-type-specific anomalies to evaluate the relative performance of unsupervised anomalous sound detection (UASD) systems across different machine types, even in the absence of real anomaly sound data. Conventional keyword-based data augmentation methods often produce unrealistic sounds due to their reliance on manually defined labels, limiting scalability as machine types and anomaly patterns diversify. Advanced audio generative models, such as MIMII-Gen, show promise but typically depend on anomalous training data, making them less effective when diverse anomalous examples are unavailable. To address these limitations, we propose a novel synthesis approach leveraging large language models (LLMs) to interpret textual descriptions of faults and automatically select audio transformation functions, converting normal machine sounds into diverse and plausible anomalous sounds. We validate this approach by evaluating a UASD system trained only on normal sounds from five machine types, using both real and synthetic anomaly data. Experimental results reveal consistent trends in relative detection difficulty across machine types between synthetic and real anomalies. This finding supports our hypothesis and highlights the effectiveness of the proposed LLM-based synthesis approach for relative evaluation of UASD systems.**

***Index Terms*—Large language model, Relative evaluation, Anomalous sound detection**

1. INTRODUCTION

Detecting anomalies in machine sounds is a critical aspect of predictive maintenance, aiming to prevent equipment failures and reduce downtime in industrial operations [1]. Machine sounds provide valuable insights into equipment health, where anomalies may indicate issues such as mechanical wear, misalignment, or impending component failure. Traditional anomaly detection methods [2]–[4] rely heavily on data, but collecting extensive datasets of real anomalous sounds is challenging due to the rarity and unpredictability of faults. Deliberately inducing faults is often impractical or hazardous, leading to a fundamental scarcity of real anomalous data.

This scarcity and lack of diversity in available real anomaly data significantly hinder not only training but also the evaluation of anomalous sound detection (ASD) systems across diverse fault conditions. Rigorous assessment of system capabilities requires suitable test data, which is often unavailable or fails to adequately represent real-world conditions. While conventional data augmentation [5]–[7] and text-to-audio (TTA)-based anomaly generation [8]–[10] attempt to address this, they face limitations: synthesized sounds may lack realism, or advanced generative models may require anomalous samples for training, further complicating evaluation when real-world anomaly data are sparse or absent. Consequently, reliably benchmarking ASD performance against diverse, realistic fault conditions across different machine types remains a significant challenge.

To address this issue, we first introduce the concept of **"relative evaluation"**, which complements conventional absolute evaluation. Relative evaluation assesses a system's comparative strengths and weaknesses by verifying **detection performance rankings across machine types**, where the system performs well for some machine types but poorly for others. This approach is essential in heterogeneous industrial environments, where maintenance engineers must know where a system is trustworthy and where it is not, enabling optimized sensor allocation, inspection scheduling, and risk management. Traditional absolute metrics, such as the Area Under the ROC Curve (AUC), can fluctuate with the severity of anomaly samples and become unreliable when real anomaly data are scarce. Specifically, severe anomalies in the test set lead to easy detection and high AUC scores, while mild anomalies make detection harder, resulting in lower AUC scores. In contrast, relative evaluation focuses on the comparative difficulty of detection tasks across machine types, providing consistent insights even when anomaly severity varies.

We further propose a novel method for synthesizing anomalous sounds to enable relative evaluation of ASD systems across machine types. The method leverages LLMs' implicit world knowledge—i.e., "common sense"—to interpret textual descriptions of faults and automatically apply appropriate acoustic transformations to normal machine sounds, thereby generating plausible acoustic characteristics for the described faults. This approach generates diverse, controllable synthetic anomalies without requiring prior anomalous data or manual intervention, facilitating relative evaluation of ASD systems' strengths and weaknesses across machine types.

2. RELATED RESEARCH

2.1. Unsupervised anomalous sound detection

Unsupervised Anomalous Sound Detection (UASD) is critical for ASD, especially when anomalous data is unavailable. The DCASE challenges [1], [11]–[14] have advanced this field by providing datasets like ToyADMOS series [15]–[18] and MIMII series [19]–[21], enabling benchmarking of techniques like autoencoder-based methods [22], Gaussian-Mixture-Model-based methods [23], embedding-similarity-based approaches [24], etc. Despite these advancements, evaluating detection systems across diverse fault conditions remains difficult due to the lack of representative anomalous test data, highlighting the need for synthetic datasets that simulate realistic anomalies.

2.2. Anomalous sound generation

Synthetic data generation addresses the shortage of real anomalous sounds for training and evaluation. Most conventional methods for anomalous sound generation focus on data augmentation to train UASD systems using anomalous examples.

One common approach is basic data augmentation, which includes pitch-shifting and time-stretching [5], Mixup-based augmentation [6], and statistics-exchange-based augmentation [7]. However, these methods often produce unrealistic sounds that are unsuitable for robust evaluation.

Another approach is TTA-based anomalous sound generation (See Table 1). Zahedi et al.'s method [8] generates anomalous sounds by randomly selecting prompts created by ChatGPT and feeding them into AudioLDM [25]. However, this method cannot achieve realistic, machine-type-specific synthesis because it does not consider machine type when selecting prompts. Zhang et al.'s method [9] converts metadata into captions, which are then input into AudioLDM. This approach can potentially achieve realistic synthesis by leveraging

Table 1: Conventional Approaches for TTA-based Anomalous Sound Generation

Characteristic	Zahedi et al. [8]	Zhang et al. [9]	MIMII-Gen [10]	**This work**
Machine-type-aware realistic synthesis	No	Yes (uses metadata)	Yes (uses metadata)	**Yes (uses metadata)**
Trainable with normal data alone	Yes	No	No	**Yes**
Purpose	Training	Training	Evaluation (absolute)	**Relative evaluation**

metadata such as machine type, but it requires anomalous samples for training. MIMII-Gen [10] has been proposed specifically for evaluating anomaly detection systems. Similar to Zhang et al. [9], MIMII-Gen converts metadata into captions and uses these captions as input for a diffusion model to generate anomalous sounds. While it can potentially achieve machine-type-aware realistic synthesis, it also requires anomalous samples for training, which limits its applicability when such data are sparse.

2.3. TTA models

This subsection highlights representative TTA models, which serve as the foundation for the above TTA-based anomaly-generation approaches [8]–[10]. TTA models leverage the contextual understanding of LLMs to generate speech, music, or environmental sounds directly from textual prompts. For high-fidelity generation, latent diffusion models have demonstrated exceptional effectiveness. AudioLDM [25] pioneered CLAP [26]-conditioned latent diffusion, enabling zero-shot audio generation. Building on similar approaches, recent works have achieved improvements in multi-domain synthesis quality and enhanced temporal coherence [27]–[29]. To address inference latency, some models compress diffusion into fewer steps [30]–[32]. TANGO 2 [33] employs preference optimization to align audio outputs with human-perceived prompt consistency, using reinforcement-style feedback to train more reliable generators. However, none of the above TTA models are trained on industrial machine recordings, while Zhang et al. [9] and MIMII-Gen [10] explicitly train their generators on machine-type-labeled industrial sound datasets, enabling type-specific fault synthesis.

2.4. Research gap and contribution

As mentioned earlier, traditional absolute metrics commonly used in anomaly detection benchmarks [1], [11]–[14] fluctuate with the severity of anomaly samples, making them unreliable when real anomaly data are scarce. Severe anomalies in the test set lead to easy detection and high AUC scores, while mild anomalies make detection harder, resulting in lower AUC scores.

To address this limitation, our first contribution is the introduction of relative evaluation, which verifies detection performance rankings across machine types. Unlike absolute evaluation, which is sensitive to anomaly severity, relative evaluation identifies where the system performs better or worse.

Our second contribution is a scalable method for generating diverse and realistic anomalous sounds using LLMs to enable relative evaluation. Instead of directly generating synthetic anomalies or manually adding them, our approach leverages LLMs to interpret machine types and their corresponding anomalies, transforming normal machine audio into anomalous audio. As shown in Table 1, unlike the conventional TTA-based anomalous sound generation methods [8]–[10], our approach enables realistic synthesis tailored to specific machine types and operates with training only on normal data. This ensures reliable evaluation across different machines, even when real anomaly data are limited.

3. PROPOSED METHOD

Our proposed method introduces a novel approach to generating anomalous machine sounds by leveraging LLMs to intelligently select and apply appropriate sound effects to normal machine audio based on descriptive captions. The comprehensive workflow illustrated in Fig. 1 consists of two major parts: (a) synthetic anomalous sound generation and (b) relative evaluation of anomaly detection systems across different machine types.

3.1. Synthetic anomaly sound generation

3.1.1. Workflow: The synthetic anomaly generation process follows a systematic workflow as illustrated in the block diagram:

- **Input metadata:** The process begins with metadata that provides contextual information about machine types, operating and environmental conditions.
- **Generate caption:** Based on the input metadata, the system generates descriptive caption using Flan-T5 [34] that characterize the machine's operational state.
- **Generate normal sound:** Using the MIMII-Gen latent diffusion model [10], we generate high-fidelity normal machine audio that serve as the foundation for anomaly introduction.
- **Initialize prompt:** This step consists of carefully crafted system prompt, user prompt containing generated caption and anomalous sound effect functions described in section 3.1.2.
- **Request to LLM:** In this step, the initialized prompt is sent to a Large Language Model via an API call. The language model analyzes the captions and autonomously selects the most appropriate sound effect to simulate potential anomaly relevant to the operating condition present in the caption.
- **Receive and interpret answer from LLM:** The system parses the LLM's response, extracts the selected function name, and maps it to the corresponding audio processing function from a predefined library of anomalous sound effects.
- **Anomalous audio generation:** The selected function is applied to the normal sound obtained from MIMII-Gen, transforming it into anomalous audio with context-appropriate fault characteristics. The generated anomalous sounds are stored along with the applied anomaly effects.

3.1.2. Anomalous sound effect functions: We implement a comprehensive library of sound effect functions that simulate various machine fault conditions:

- **Squeaking or Squealing**: High-pitched sounds indicating faulty bearings or friction between components.
- **Rattling or Knocking**: Noises suggestive of loose parts or misalignment.
- **Grinding or Scraping**: Sounds indicative of severe mechanical wear or damage.
- **Humming or Buzzing**: Low-frequency sounds resulting from electrical issues or resonance.
- **Whistling or Hissing**: Sounds associated with air leaks or high-pressure flow.
- **Clicking or Tapping**: Noises caused by relay switches or intermittent contacts.

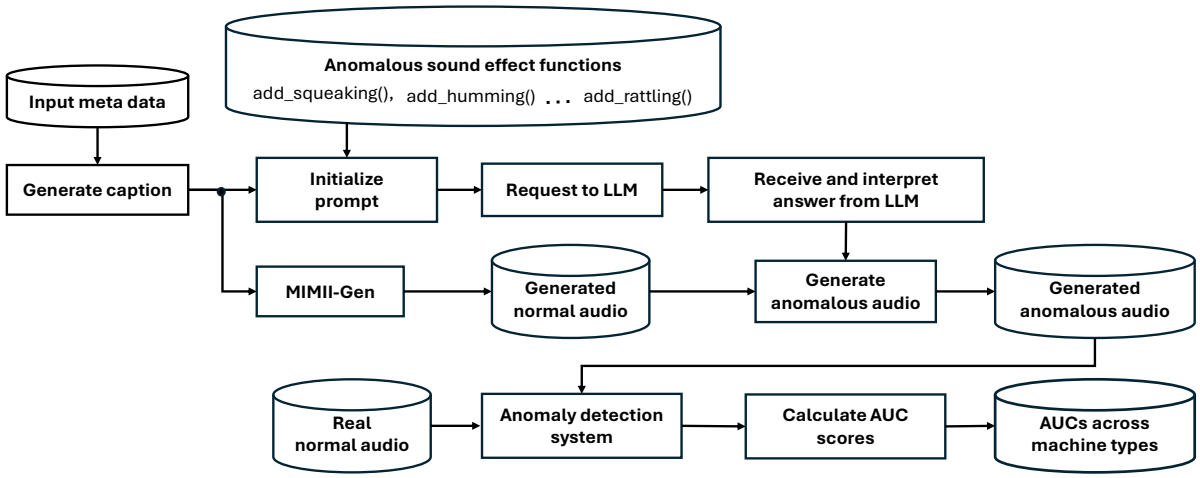

Fig. 1: Workflow of Our Approach

- **Pulsing or Chattering**: Effects indicating fluctuating power supply or control issues.
- **Pop or Bang**: Sudden sounds simulating abrupt failures or explosive events.
- **Changes in Tonal Quality or Frequency**: Alterations representing shifts in machine operation.
- **Broadband Noise Increases**: Overall noise level increases to simulate general degradation.

Each function uses specific digital signal processing techniques to modify normal sound waveforms, creating realistic fault signatures.

Example: The sound effect functions are implemented using Python libraries such as NumPy, Librosa, and SoundFile, manipulating the audio waveforms to introduce the selected anomalies.

```
def add_squeaking(audio, sr,
    duration=2.0, freq=4000, intensity=0.3):
    # Adds a high-pitched squeaking sound
    # to the audio
    # Function implementation
    return audio_with_squeaking
```

3.2. Anomaly detection system evaluation

3.2.1. Unsupervised anomaly detection system: The UASD system is trained exclusively on real normal sound and calculates anomaly scores, such as reconstruction errors in autoencoders, to identify anomalies. This unsupervised approach aligns with real-world scenarios where anomalies are rare and underrepresented in training data. The UASD system can utilize methods such as autoencoder-based [22], Gaussian Mixture Model-based [23], or embedding-similarity-based approaches [24]. The system processes all clips from real normal sound and synthetic anomaly datasets and computes anomaly scores for each clip.

3.2.2. AUC calculation and relative evaluation: The AUC score for each machine type is calculated as:

$$\mathrm{AUC}_m = \frac{1}{N_m^- N_m^+} \sum_{i=1}^{N_m^-} \sum_{j=1}^{N_m^+} \mathcal{H}(\mathcal{A}(x_j^+) - \mathcal{A}(x_i^-)), \qquad (1)$$

where m represents the machine type index, $\mathcal{H}(x)$ returns 1 if $x > 0$ and 0 otherwise, $\{x_i^-\}_{i=1}^{N_m^-}$ are normal test clips, and $\{x_j^+\}_{j=1}^{N_m^+}$ are anomalous test clips for machine type m. N_m^- and N_m^+ indicate the number of normal and anomalous test clips, respectively.

By comparing AUC scores across machine types, users can identify the system's relative strengths and weaknesses based on detection performance rankings.

4. EXPERIMENTATION

The experiments aim to validate two key objectives. The first objective is to confirm the correlation between synthetic and real anomaly detection performance rankings across different machine types, demonstrating that synthetic anomaly generation using our proposed approach enables users to identify the system's relative strengths and weaknesses. The second objective is to validate the effectiveness of LLM-based approaches in generating contextually appropriate synthetic anomalies compared to manual and random methods, through an ablation study. This section provides details on the dataset, anomaly detection system, results, and an ablation study.

4.1. Dataset preparation

Table 2 summarizes the datasets used in this study. We prepared three distinct datasets to support our evaluation: normal sounds, synthetic anomalous sounds, and real anomalous sounds. Each dataset was tailored to the five machine types under study.

Table 2 summarizes the datasets used in this study. We prepared three distinct datasets tailored to five machine types: bearings, gearboxes, fans, valves, and slide rails.

- **Normal Sounds:** We collected 900 normal sound recordings per machine type (bearings, gearboxes, fans, valves, and slide rails), each 10 seconds long, sampled at 16 kHz. These were sourced from industrial environments and public datasets like MIMII-DG [21] ensuring a comprehensive representation of typical operating conditions.
- **Synthetic Anomalous Sounds:** For each machine type, we generated 50 synthetic anomalies by applying sound effects to normal sounds. The language model (GPT-4) selected effects (e.g., squeaking, rattling) based on captions describing operating conditions (e.g., "Bearing operating at 24 krpm").
- **Real Anomalous Sounds:** We acquired 50 real anomalous recordings per machine type. These anomalies represent actual faults, such as mechanical wear or misalignment, and vary in severity.

Table 2: Dataset Summary

Category	Samples per Machine	Duration	Sample Rate
Normal Sounds	900	10 s	16 kHz
Synthetic Anomalies	50	10 s	16 kHz
Real Anomalies	50	10 s	16 kHz

Table 3: Detection Performance on Synthetic vs. Real Anomalies

Machine Type	Synthetic			Real		
	AUC		Rank	AUC		Rank
	MSE	MAHALA		MSE	MAHALA	
Bearing	0.85	0.82	3	0.57	0.61	3
Gearbox	0.88	0.86	2	0.62	0.67	2
Fan	0.92	0.95	1	0.90	0.93	1
Slide rail	0.80	0.79	4	0.55	0.57	4
Valve	0.78	0.72	5	0.53	0.52	5

4.2. Anomaly detection system design

We employed an unsupervised anomaly detection system based on an autoencoder, trained exclusively on normal sounds to detect deviations indicative of anomalies.

Autoencoder Architecture:

- **Input Layer:** Log-mel spectrograms with 128 mel-bins, extracted from 64-ms frame windows with 50% hop size.
- **Encoder:** Three layers (128, 64, 32 filters, kernel size 3), each followed by ReLU activation and max-pooling (2x2).
- **Decoder:** Mirrored layers with upsampling, reconstructing the input spectrogram.

Training Details: The autoencoder was trained for 100 epochs with a batch size of 32, using the Adam optimizer (learning rate 0.001) and mean squared error (MSE) loss. A single model was trained across all machine types to generalize normal patterns, reflecting real-world scenarios with diverse equipment.

4.3. Results

Table 3 summarizes the detection performance for synthetic and real anomalies using AUC scores, derived from two distinct metrics: Mean Squared Error (MSE) and Mahalanobis distance (MAHALA). Employing both scoring methods provides a more robust validation of the relative evaluation on synthetic data. Synthetic anomalies consistently achieved higher AUC scores compared to real anomalies, indicating they are easier to detect. Notably, the ranking of AUC scores, and thus the relative anomaly detection difficulty, is consistent across machine types for both synthetic and real data. These results demonstrate that synthetic anomaly generation using our proposed approach effectively enables users to evaluate the system's relative strengths and weaknesses.

Differences in AUC scores reflect the unique operational behaviors and fault characteristics of each machine type. For example, lower AUC scores for valves and slide rails suggest that anomalies in these machines may be subtler or involve features that are harder for the anomaly detection system to capture.

4.4. Ablation study

To validate the reliability of the LLM-based approach, we compared three configurations: (1) our approach with GPT-4o-based anomaly function calling (with values reproduced from Table 3), (2) a keyword-based manual mapping of anomalies created through human labeling based on common knowledge (e.g., adding a "squeaking" anomaly if the caption includes "bearing"), and (3) random selection of possible anomalies without contextual understanding.

Table 4: Ablation Study: Common Knowledge Impact on MSE-AUC

Machine Type	GPT-4o		Manual-mapping (w/ Knowledge)		Random (w/o Knowledge)	
	AUC	rank	AUC	rank	AUC	rank
Bearing	0.85	3	0.70	3	0.81	5
Gearbox	0.88	2	0.72	2	0.85	3
Fan	0.92	1	0.78	1	0.82	4
Slide rail	0.80	4	0.67	4	0.86	2
Valve	0.78	5	0.65	5	0.89	1

Table 4 presents the results of the ablation study. The AUC-score rankings across machine types produced using GPT-4o and the keyword-based manual mapping approach formed by human labels were both closely aligned with those of real anomalies. In contrast, random anomaly selection showed no correlation with the AUC-score rankings of real anomalies, highlighting the importance of contextual understanding in anomaly generation. These results confirm that the ability of LLMs to interpret machine-specific characteristics and fault descriptions, leveraging common knowledge embedded within them, enables the creation of realistic and relevant anomalies. Furthermore, the observation that sound effects impact machines differently emphasizes the need to tailor the selection process to each machine type. The consistent performance of LLM-based anomaly generation demonstrates its potential as a scalable and efficient alternative to human labeling, capable of supporting diverse machine types and anomaly scenarios. Improving prompt design and incorporating domain-specific constraints could further enhance the realism of generated anomalies, thereby increasing the approach's utility for UASD evaluation.

5. CONCLUSION

This paper addressed the challenge of evaluating UASD systems in the absence of sufficient and diverse real anomaly data. To tackle this, we proposed two key contributions: (1) the introduction of relative evaluation, which verifies detection performance rankings across machine types. Unlike absolute evaluation, which is sensitive to anomaly severity, relative evaluation identifies where the system performs better or worse. (2) a novel synthesis approach using LLMs with function-calling capabilities. Our method leverages the world knowledge of LLMs to interpret textual descriptions of machine conditions and automatically apply audio transformations, generating diverse and plausible synthetic anomalies for evaluation, without requiring prior anomalous examples.

Experiments showed that AUC-score rankings are consistent across machine types as well as different anomaly detection systems for both synthetic and real data. Also, rankings produced using GPT-4o and keyword-based manual mapping closely aligned with those of real anomalies, while random anomaly selection showed no correlation, highlighting the importance of contextual understanding in anomaly generation. These findings validate the ability of LLMs to generate realistic and relevant anomalies by interpreting machine-specific characteristics and fault descriptions. Our approach offers a reliable and scalable tool for benchmarking UASD performance and understanding relative detection difficulty across machine types, especially in scenarios lacking sufficient real anomaly data.

REFERENCES

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. Workshop*

on Detection and Classification of Acoustic Scenes and Events (DCASE), 2020, pp. 81–85.

[2] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 1, pp. 279–288, 2015.

[3] G. Coelho, L. M. Matos, P. J. Pereira, A. Ferreira, A. Pilastri, and P. Cortez, "Deep autoencoders for acoustic anomaly detection: Experiments with working machine and in-vehicle audio," *Neural Computing and Applications*, vol. 34, no. 22, pp. 19 485–19 499, 2022.

[4] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 271–275.

[5] T. Inoue, P. Vinayavekhin, S. Morikuni, S. Wang, T. H. Trong, D. Wood, M. Tatsubori, and R. Tachibana, "Detection of anomalous sounds for machine condition monitoring using classification confidence," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 66—-70.

[6] K. Wilkinghoff, "Fraunhofer FKIE submission for task2: First-shot unsupervised anomalous sound detection for machine condition monitoring," DCASE Challenge, Tech. Rep., 2023.

[7] H. Chen, Y. Song, Z. Zhuo, Y. Zhou, Y.-H. Li, H. Xue, and I. McLoughlin, "An effective anomalous sound detection method based on representation learning with simulated anomalies," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[8] E. Zahedi, M. Saraee, F. S. Masoumi, and M. Yazdinejad, "Regularized contrastive masked autoencoder model for machinery anomaly detection using diffusion-based data augmentation," *Algorithms*, vol. 16, no. 9, p. 431, 2023.

[9] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1271–1275.

[10] H. Purohit, T. Nishida, K. Dohi, T. Endo, and Y. Kawaguchi, "MIMII-Gen: Generative modeling approach for simulated evaluation of anomalous sound detection system," *arXiv preprint arXiv:2409.18542*, 2024.

[11] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021, pp. 186–190.

[12] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2022.

[13] 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 *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023, pp. 31–35.

[14] 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," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2024, pp. 111–115.

[15] Y. Koizumi, S. Saito, N. Harada, H. Uematsu, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 313–317.

[16] 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. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021, pp. 1–5.

[17] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "ToyADMOS2+: New ToyADMOS data and benchmark results of the first-shot anomalous sound event detection baseline," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023, pp. 41–45.

[18] D. Niizumi, N. Harada, Y. Ohishi, D. Takeuchi, and M. Yasuda, "ToyADMOS2: Yet another dataset for the DCASE2024 challenge task 2 first-shot anomalous sound detection," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2024, pp. 106–110.

[19] 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 *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019, pp. 209–213.

[20] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 21–25.

[21] 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. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2022.

[22] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the Neyman-Pearson lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2019.

[23] Z. Mnasri, S. Rovetta, and F. Masulli, "Anomalous sound event detection: A survey of machine learning based methods and applications," *Multimedia Tools and Applications*, vol. 81, pp. 1–35, 2021.

[24] Y. Wang, Y. Zheng, Y. Zhang, Y. Xie, S. Xu, Y. Hu, and L. He, "Unsupervised anomalous sound detection for machine condition monitoring using classification-based methods," *Applied Sciences*, vol. 11, no. 23, p. 11128, 2021.

[25] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. International Conference on Machine Learning (ICML)*, 2023, pp. 21 450–21 474.

[26] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[27] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-An-Audio: Text-to-audio generation with prompt-enhanced diffusion models," in *Proc. International Conference on Machine Learning (ICML)*, 2023, pp. 13 916–13 932.

[28] J. Huang, Y. Ren, R. Huang, D. Yang, Z. Ye, C. Zhang, J. Liu, X. Yin, Z. Ma, and Z. Zhao, "Make-An-Audio 2: Temporal-enhanced text-to-audio generation," *arXiv preprint arXiv:2305.18474*, 2023.

[29] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[30] H. Liu, R. Huang, Y. Liu, H. Cao, J. Wang, X. Cheng, S. Zheng, and Z. Zhao, "AudioLCM: Text-to-audio generation with latent consistency models," in *Proc. ACM International Conference on Multimedia (MM)*, 2024, pp. 7008—7017.

[31] Y. Bai, T. Dang, D. Tran, K. Koishida, and S. Sojoudi, "ConsistencyTTA: Accelerating diffusion-based text-to-audio generation with consistency distillation," in *Proc. INTERSPEECH*, 2024, pp. 3285–3289.

[32] K. Saito, D. Kim, T. Shibuya, C.-H. Lai, Z. Zhong, Y. Takida, and Y. Mitsufuji, "SoundCTM: Unifying score-based and consistency models for full-band text-to-sound generation," in *Proc. International Conference on Learning Representations (ICLR)*, 2025.

[33] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," in *Proc. ACM International Conference on Multimedia (MM)*, 2024, pp. 564—-572.

[34] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 1, pp. 70:1–70:53, 2024.