First-Shot Unsupervised Anomalous Sound Detection with Metadata-Assisted Audio Generation

Feiyang Xiao, Jian Guan, *Senior Member, IEEE*, Kejia Zhang, Qiaoxi Zhu, *Member, IEEE*, Wenbo Wang, Haohe Liu, Wenwu Wang, *Fellow, IEEE*

***Abstract*—First-shot unsupervised anomalous sound detection (ASD) aims to detect anomalies in new machine types using limited normal sounds. Its applicability is hindered by the lack of real anomalies for parameter tuning and the scarcity of target-domain normal sounds, both of which limit the robustness of detection models to domain shifts caused by variations in operating and environmental conditions. To address these challenges, we propose Gen-ASD, a metadata-assisted audio generation and detection framework that synthesizes anomalous sounds for parameter tuning during model validation and normal sounds for improved representation of target domain conditions. Machine metadata describe machine status, operating conditions, and environmental factors that shape acoustic characteristics. Leveraging these relationships, Gen-ASD employs a consistency-enhanced text-to-audio generator to learn metadata-acoustic relationships from reference machine types and synthesize normal and anomalous sounds for unseen machine types. Generated anomalies serve as surrogate validation data for tuning-dependent ASD methods, while generated normal sounds complement under-represented target-domain conditions to improve the robustness of detection models to domain shifts. These complementary capabilities enable detector training and anomaly-aware validation without real anomalies of new machine types. Extensive experiments demonstrate that Gen-ASD integrates seamlessly with both non-deep-learning and deep-learning ASD models in a plug-and-play manner. Further analyses validate the effectiveness of generated validation data for parameter tuning and the consistency of synthesised sounds with the under-represented target-domain conditions.**

***Index Terms*—Unsupervised learning, anomalous sound detection, metadata, audio generation**

I. Introduction

ANOMALOUS sound detection (ASD) aims to identify unexpected anomalous machine sounds from normal ones and is crucial for fault diagnosis, industrial monitoring, and predictive maintenance [1]–[6]. Early detection prevents costly failures and improves safety. However, collecting sufficiently diverse anomalous sounds that cover various conditions is inherently difficult because faults are rare and unpredictable [7], [8]. To address this issue, unsupervised ASD, which trains detection models using only normal machine sounds, has emerged as a promising approach [1].

Early unsupervised ASD methods usually model the feature distribution of normal sounds using feature reconstruction [9], [10] or self-supervised classification [7], [11], [12]. However, these methods often overlook domain shift, a common challenge in real-world applications. Domain shift arises from variations in machine operating and environmental conditions, such as operating speed, background noise, and recording position [13], leading to a distribution mismatch between the source domain used for training and the target domain encountered during deployment [14], [15]. Here, the source domain comprises machine sounds recorded under conditions adequately represented during training, whereas the target domain comprises sounds recorded in under-represented conditions that may arise in deployment. This mismatch can substantially degrade ASD performance when models are applied to machines operating under conditions that differ from those seen during training [16].

To mitigate domain shift, existing methods usually use metadata (i.e., machine types, operating conditions, and environmental noise) as self-supervised classification labels to model the distributions of audio features across domains [14], improving the generalization of ASD models under domain shift [17], [18]. However, although these ASD models are trained exclusively on normal samples, parameter tuning and model selection may still require validation data containing normal and anomalous samples [14], [19]. In practice, such anomalous samples are typically unavailable for new machine types, limiting the practical applicability of these methods in first-shot ASD [20].

This challenge is formalized in the first-shot unsupervised ASD task (i.e. Task 2) of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge [1], which aims to achieve anomaly detection in new machine types without access to anomalous samples during training [20]. Here, the training data comprise *reference* types (with both normal and anomalous sounds) and *new* types (with only normal sounds). Metadata, including machine type and operating conditions, are available in both groups. Importantly, the reference types (e.g., Fan and Gearbox) are completely disjoint from the new types (e.g., Vacuum and ToyTank) [21]. In this setting, the lack of anomalous samples from new

Corresponding author: Jian Guan.

Feiyang Xiao, Jian Guan, Kejia Zhang, and Wenbo Wang are with the College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China (emails: xiaofeiyang128@gmail.com; {j.guan, kejiazhang}@hrbeu.edu.cn; wwb1325864697@outlook.com).

Qiaoxi Zhu is with the University of Technology Sydney, Ultimo, NSW, Australia (email: qiaoxi.zhu@gmail.com).

Haohe Liu and Wenwu Wang are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (emails: {haohe.liu, w.wang}@surrey.ac.uk).

This work was partly supported by the Heilongjiang Provincial Natural Science Foundation of China under Grant No. BS2025F009, and the Greater Bay Area Institute of HPC-AI Co-Driven Innovation under Grant No. 03202604230020.

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

machine types prevents existing ASD methods from being tuned for first-shot scenarios [20], severely limiting their generalization capability [21].

Recent state-of-the-art methods for first-shot unsupervised ASD primarily leverage data augmentation or pretrained models to improve the generalization of acoustic representations. Data augmentation-based methods, including SSL4ASD [22], GLAM [23], and UnlabeledASD [24], use techniques including Mixup [25], feature exchange [26], and SpecAugment [27] to enrich the representations of normal sounds. However, these methods mainly perturb existing samples and are not designed to generate anomaly-aware validation data or synthesize normal sounds for under-represented target-domain conditions [1], [12], [28]. Pretrained model-based methods, including AnoPatch [29], Wav2Vec2ASD [30], SSAMG-ASD [31], and FMAQP [32], leverage large-scale pretrained audio models to learn discriminative acoustic representations, often incorporating metadata into representation learning. For example, AnoPatch [29] converts metadata into discrete labels to train an audio encoder to distinguish the corresponding metadata classes. Thus, metadata mainly provides discriminative supervision, rather than serving as a semantic condition for explicitly modeling how operating and environmental conditions shape acoustic characteristics.

This distinction motivates us to exploit the semantic relationship between machine metadata and acoustic characteristics. Metadata describe machine type, operating conditions (e.g., rotational speed and mode), and environmental factors (e.g., environmental noise and microphone location) [13], all of which shape the physical processes underlying machine sounds [12], [18], [22]. Therefore, metadata serve not merely as domain labels [17], but as semantic cues for the expected acoustic characteristics under specific conditions [33], [34]. Since reference machine types provide both normal and anomalous sounds with corresponding metadata, models can learn the relationships between machine states, operating conditions, and acoustic variations associated with anomalies, and transfer this knowledge to unseen machine types. Such transfer relies on the relevance of learned acoustic patterns to unseen machines and the informativeness of their metadata, without requiring identical machine types or recording conditions.

Recent advances in text-to-audio (TTA) generation [35]–[37] enable sound synthesis from semantic descriptions. This raises a key question: Can we use TTA generation to exploit the transferable relationship between machine metadata and acoustic characteristics for first-shot unsupervised ASD? To investigate this, we fine-tune pretrained TTA models on reference machine recordings and their metadata to adapt them to industrial machine acoustics. Given the diversity and unpredictability of machine faults, the goal is not to reproduce specific anomalies faithfully, but to generate transferable anomaly-related acoustic variations that serve as surrogate validation signals for model parameter tuning.

Several studies have investigated the use of synthetic anomalous sounds for ASD. Early methods, such as [38], use augmented normal sounds as surrogate anomalies for representation learning, but do not leverage anomalous sounds from reference machine types or exploit metadata–acoustic relationships. More recent methods, including MIMII-Gen [39] and MIMII-Agent [40], employ latent diffusion models to generate proxy-anomalous sounds for ASD evaluation. However, they are not designed to transfer anomaly-related acoustic knowledge across machine types to previously unseen machines in first-shot scenarios. Therefore, despite demonstrating the potential of synthetic anomalies, existing methods do not address the two interrelated challenges of first-shot ASD under domain shift, i.e., the lack of anomalous validation data and the limited representation of target-domain normal conditions.

To address this gap, we propose Gen-ASD, a metadata-assisted audio generation and detection framework for first-shot unsupervised ASD. Unlike existing anomaly-synthesis methods, Gen-ASD learns transferable metadata-acoustic relationships from reference machine types with both normal and anomalous sounds and transfers this knowledge to unseen machine types for which only normal sounds and metadata are available. Specifically, metadata describing machine operating status and environmental context are formulated as semantic conditions to guide a consistency-enhanced text-to-audio (CE-TTA) generator [37]. The resulting synthetic sounds provide anomaly-aware validation signals for parameter tuning, while condition-aware normal sounds enrich under-represented target-domain conditions, thereby mitigating domain shift. To the best of our knowledge, Gen-ASD is the first framework to explicitly exploit transferable metadata-acoustic relationships through generative modeling for first-shot ASD under domain shift. Moreover, its model-agnostic and plug-and-play design enables integration with both deep and non-deep ASD models without requiring architectural modification.

Extensive experiments conducted on the datasets from Task 2 of DCASE 2023 Challenge illustrate the superiority and effectiveness of our Gen-ASD, compared with state-of-the-art ASD methods. The initial version of our study [41] formed the core of our system ranked 7th in the DCASE 2023 Challenge. In this paper, the improved version with lightweight TWFR-GMM [19] surpasses the 1st-place system by 1.05% in AUC metric while using less than 1.1% of its parameters. Moreover, integrating Gen-ASD with both deep-learning (AnoPatch [29]) and non-deep-learning (TWFR-GMM [19]) detectors without any architecture changes also highlights the strong generalization ability of our framework.

It should be noted that the proposed Gen-ASD is fundamentally different from conventional data augmentation techniques. Instead of introducing augmentations to existing audio samples, Gen-ASD explicitly models the metadata–acoustic relationship through a metadata-assisted generative framework. This enables the synthesis of anomalous sound characteristics for unseen machine types, facilitating anomaly-aware parameter optimization and target-domain acoustic alignment beyond conventional data augmentation.

Contributions of this work can be summarized as follows:

1) We propose Gen-ASD, a metadata-assisted framework for first-shot unsupervised ASD under domain shift that addresses the lack of anomalous validation data and the limited representation of target-domain conditions.
2) Gen-ASD learns the relationships between machine status, operating conditions, environmental factors, and

acoustic characteristics from reference machine types, and then transfers this knowledge to unseen machine types through metadata-guided generative modelling.
3) Gen-ASD generates anomalous sounds for parameter tuning during model validation in the presence of unseen machine types, and synthetic normal sounds for improving the representation of target-domain conditions.
4) Extensive experiments with deep and non-deep ASD detectors, together with controlled ablations and repeated runs, validate the effectiveness, robustness, and model-agnostic applicability of Gen-ASD.

This study substantially extends our preliminary conference paper [28] in four key aspects. First, we replace the original TTA model AudioLDM [36] with the faster, consistency-enhanced AudioLCM [37], achieving a significant reduction in generation time (as shown in Section IV-C3). Second, beyond addressing the lack of anomalies in first-shot scenarios, we mitigate domain shift by generating target-domain normal sounds to enrich audio representations and improve detection performance. Third, we demonstrate the generalizability of Gen-ASD across different detector architectures by extending the evaluation from the non-deep-learning-based TWFR-GMM [19] to the deep learning-based AnoPatch [29]. Finally, we conduct extensive additional analyses to investigate the rationale behind generation-based parameter tuning (Section IV-D1), the impact of the number of generated sounds under domain shift (Section IV-D2), the effects of alternative metadata formulations (Section IV-C4), and the efficiency of different generators (Section IV-C3). Together, these extensions provide a more comprehensive evaluation of Gen-ASD for first-shot unsupervised ASD.

II. Preliminary

To demonstrate the versatility of the proposed Gen-ASD, we integrate it with two state-of-the-art ASD methods, i.e., the non-deep-learning-based TWFR-GMM [19] and the deep learning-based AnoPatch [29]. For clarity, this section first provides a concise overview of each method's workflow before detailing how our Gen-ASD solution works.

A. TWFR-GMM

TWFR-GMM [19] is a state-of-the-art non-deep-learning-based ASD method with only 33K parameters, making it highly suitable for resource-constrained environments. It obtains the time-weighted frequency domain audio representation (TWFR), by incorporating a hyperparameter r for each machine type, as follows

$$R(\mathbf{X}) = \mathrm{Ranking}(\mathbf{X}) \cdot \left[\frac{r^0}{z(r)}, \frac{r^1}{z(r)}, ..., \frac{r^{N-1}}{z(r)}\right]^\top, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{M \times N}$ is the log-mel spectrogram of an audio signal with M mel-bins and N time frames, and $R(\mathbf{X}) \in \mathbb{R}^M$ denotes the TWFR feature. $\mathrm{Ranking}(\cdot)$ denotes the operation of re-arranging $\mathbf{X}$ in descending order for the energy values over time frames for the weight calculation, following [19]. Here, the hyper-parameter r determines the weight assigned to each time frame, and $z(r) = \sum_{n=1}^{N} r^{n-1}$ is for normalization. $\top$ denotes the transpose operation. Then, the TWFR feature is sent to a lightweight Gaussian mixture model (GMM) detector to distinguish anomalies from normal sounds [19].

The TWFR representation relies on a machine-specific hyper-parameter r, typically selected using anomalous validation samples, limiting the applicability of TWFR-GMM to first-shot ASD. This makes it a suitable testbed for evaluating whether generated anomalies provide informative cues for effective parameter tuning in the absence of real anomalies.

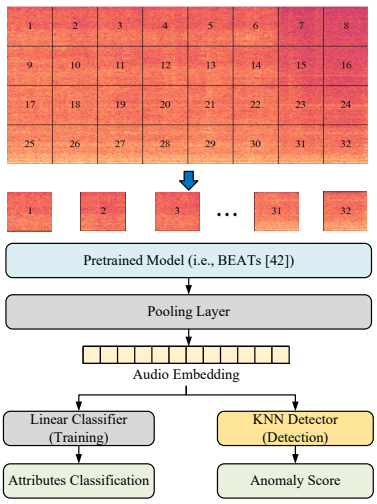


Fig. 1. Workflow of the state-of-the-art deep learning based method AnoPatch [29]. Here, the numbers (e.g. 1, 2, ..., 32) in the spectrogram index the input audio feature patches.

B. AnoPatch

AnoPatch is the current top-performing deep learning method for first-shot ASD [29]. It uses the feature representation from the pretrained audio model, i.e., BEATs [42], to improve its performance for new machine types.

Specifically, the spectrogram of machine sound is partitioned into a sequence of local time-frequency patches, each serving as an individual feature unit that is sequentially fed into a pretrained model for representation learning. As illustrated in Figure 1, AnoPatch fine-tunes the pretrained BEATs model for attribute classification, enabling it to learn detailed and discriminative audio representations that better distinguish normal from anomalous sounds. Then, following [22], a detector based on K-Nearest Neighbor (KNN) [43] is used to calculate the cosine distances between the representation of a test machine sound and those of the training samples, with the minimum distance adopted as the anomaly score.

However, in first-shot scenarios, AnoPatch's audio representations remain constrained by the lack of anomalous samples, making parameter tuning heavily reliant on empirical judgment and thus less interpretable. In addition, imbalances in the number of available sound samples across domains can further limit its detection performance in the target domain.

III. Proposed Metadata-assisted Audio Generation-based ASD Framework

To address the challenge of first-shot unsupervised ASD, we propose a metadata-assisted audio Gen-ASD framework,

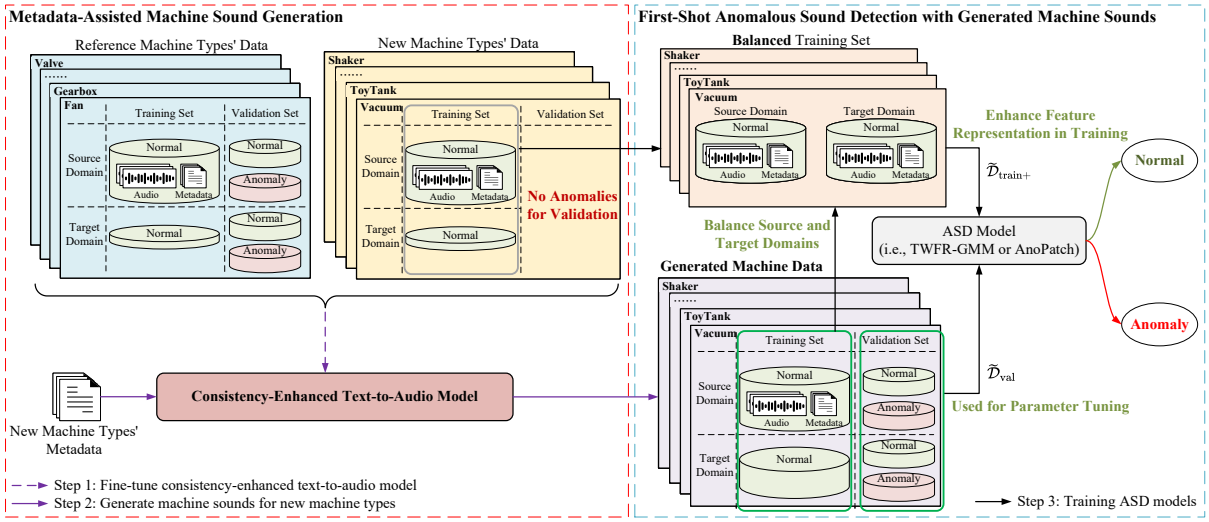


Fig. 2. The overview of the proposed metadata-assisted Gen-ASD framework. The proposed framework explores relationship between acoustic characteristics and metadata to generate machine sounds for ASD model training, which addresses the lack of anomalous sounds in first-shot unsupervised ASD and enhances the detection performance of ASD models by expanding the training data.

illustrated in Figure 2. In this framework, machine sounds for new machine types are generated under metadata guidance by fine-tuning the CE-TTA model [37], which learns both normal and anomalous acoustic patterns while enforcing the consistency between metadata and audio features. Building on this capability, Gen-ASD enables efficient detection with the non-deep-learning method TWFR-GMM [19] by generating both normal and anomalous sounds to support parameter tuning. It also generalizes effectively to deep-learning-based approaches, where generating additional normal sounds helps alleviate data imbalance under domain shift, thereby improving the detection performance of AnoPatch [29].

A. Metadata-Assisted Machine Sound Generation

Unlike methods (e.g., SpecAugment [27] and mixup [25]), which introduce augmentations to existing audio features, the proposed Gen-ASD framework performs metadata-assisted generative modeling. Specifically, Gen-ASD learns the relationship between machine metadata (e.g., operating status and environmental factors) and acoustic characteristics by generating audio samples conditioned on semantic information derived from the metadata, rather than relying on conventional data augmentation. Details are described below.

1) Metadata Caption Extraction: Metadata captures critical physical and operational characteristics of machine sounds, such as machine type, status, operating conditions and environmental noise, which directly influence acoustic properties [12], [17]. Metadata is usually embedded in filenames [1], [44], [45]. For example, the machine type (ToyCar), status (Normal), and attributes (model B2, speed 31V, mic position 1) information are encoded in the filename "ToyCar/Section_00_Source_test_Normal_0001_Car_B2_Spd_31V_Mic_1.wav".

Rather than using these discrete metadata fields independently, we organize them into a complete semantic description of the machine state, operating condition, and acoustic environment using the template: "*This is the **[machine status]** sound of a **[machine type]** with **[attributes]**.*" For instance, the above metadata is converted into "*This is the **normal** sound of a **toy car** with **model B2 and speed 31V, recorded by a microphone placed at position 1**.*" Although the template words are shared across samples, they integrate multiple discrete attributes into a unified contextual description, allowing the text-conditioned generator to interpret their joint operating context rather than treating them as isolated metadata tokens. Table I illustrates examples of this conversion. This caption-based representation is an auxiliary metadata-organization strategy rather than a core contribution of Gen-ASD, and its effectiveness is empirically evaluated against direct discrete metadata input in Section IV-C4.

2) Consistency-Enhanced Text-to-Audio Model Fine-tuning: In the proposed framework, we fine-tune the CE-TTA model [37] to generate machine sounds for new machine types, conditioned on metadata captions. This builds the relationship between the metadata caption and the machine sound by enhancing the consistency of their latent features during generation. The fine-tuning process can be represented as

$$\mathcal{G} \leftarrow P(\boldsymbol{A}|\boldsymbol{C}), \tag{2}$$

where P denotes the pretrained AudioLCM model [37], $\mathcal{G}$ denotes the model fine-tuned by a set of available machine sounds $\boldsymbol{A}$ with the corresponding set of metadata captions $\boldsymbol{C}$ as a condition, as shown in the left part of Figure 2.

Unlike conventional augmentation methods that perturb existing samples, Gen-ASD uses metadata to learn the relationship between machine conditions and acoustic characteristics, generating condition-aware normal sounds to mitigate domain shift and proxy anomalous sounds to support parameter tuning when real target-domain anomalies are unavailable.

In the CE-TTA model [37], the text encoder consists of the contrastive language-audio pretraining (CLAP) module [46] and the T5 language module [47] to capture detailed textual content from a metadata caption c, as

$$\mathbf{c} = [\mathrm{CLAP}(c); \mathrm{T5}(c)], \tag{3}$$

TABLE I

EXAMPLES OF METADATA CAPTIONS FOR MACHINE SOUNDS FROM ORIGINAL FILENAMES, INCLUDING THE MACHINE TYPE, STATUS, OPERATING AND ENVIRONMENTAL CONDITIONS. HERE, BOTH THE NORMAL AND ANOMALOUS SOUND OF THE REFERENCE MACHINE TYPE, E.G., TOYCAR, AND THE NORMAL SOUND OF THE NEW MACHINE TYPE, E.G., GRINDER, CAN BE USED TO EXTRACT THE METADATA CAPTIONS.

Machine Type	Original Filename	Metadata Caption
ToyCar	ToyCar/section_00_source_test_***normal***_0001_car_***B2***_spd_***31V***_mic_***1***.wav	This is the ***normal*** sound of a ***toy car*** with model ***B2*** and speed ***31V***, recorded by a microphone placed at the position ***1***.
ToyCar	ToyCar/section_00_source_test_***anomaly***_0001_car_***B2***_spd_***31V***_mic_***1***.wav	This is the ***anomaly*** sound of a ***toy car*** with model ***B2*** and speed ***31V***, recorded by a microphone placed at the position ***1***.
Grinder	Grinder/section_00_source_train_***normal***_0000_grindstone_***2***_plate_***2***.wav	This is the ***normal*** sound of a ***grinding*** machine with grindstones ***2*** and metal plates ***2***.

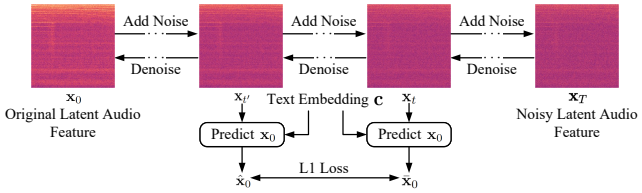


Fig. 3. Illustration of the fine-tuning process of the CE-TTA generation model in our method.

where $\mathbf{c}$ denotes the text embedding extracted from the text encoder. The CLAP module aligns text and audio embeddings in a shared latent space, while the T5 language model enriches text embeddings by capturing detailed semantic information from metadata captions. Following [37], [48], we concatenate the two embeddings directly, combining the cross-modal alignment offered by CLAP with rich semantic representations from T5 without introducing additional fusion parameters. The resulting text embedding guides audio generation in the latent feature space.

As illustrated in Figure 3, in the generation process, the CE-TTA model [37] first adds random Gaussian noise to the latent audio feature $\mathbf{x}_0$ of a machine sound in $\boldsymbol{A}$ to transform the distribution of audio feature into the standard Gaussian, as follows,

$$\mathbf{x}_t = \alpha(t)\mathbf{x}_0 + \beta(t)\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4}$$

where $\mathbf{x}_t$ denotes the obtained noisy feature at the t-th time step, $\alpha(t)$ and $\beta(t)$ denote the functions for controlling noise levels, satisfying a normalization condition (i.e., $\alpha(t)^2 + \beta(t)^2 = 1$ and $\alpha(0) = 1, \beta(0) = 0$) to ensure that $\mathbf{x}_t = \mathbf{x}_0$ when $t = 0$. The maximum time step T is set empirically as 1000 in our experiments.

With the text embedding $\mathbf{c}$ of the metadata caption, the CE-TTA model can predict the original latent audio feature from the noisy feature at any time step, as shown in Figure 3. For instance, the original latent audio feature can be predicted from noisy feature $\mathbf{x}_t$ at the t-th time step, as

$$\hat{\mathbf{x}}_0 = \lambda_{\text{skip}}(t)\mathbf{x}_t + \lambda_{\text{out}}(t)\mathcal{G}(\mathbf{x}_t, t, \mathbf{c}), \tag{5}$$

where $\hat{\mathbf{x}}_0$ denotes the predicted audio feature from the CE-TTA model $\mathcal{G}$, $\lambda_{\text{skip}}(t)$ and $\lambda_{\text{out}}(t)$ are differentiable functions defined in consistency-based generation models [37], [48] to balance the input noisy audio feature and the model output feature for stable prediction of $\mathbf{x}_0$. With a similar process, $\mathbf{x}_0$ can also be predicted from another time step $t' \neq t$, as

$$\bar{\mathbf{x}}_0 = \lambda_{\text{skip}}(t')\mathbf{x}_{t'} + \lambda_{\text{out}}(t')\mathcal{G}(\mathbf{x}_{t'}, t', \mathbf{c}), \tag{6}$$

where $\bar{\mathbf{x}}_0$ denotes the predicted latent audio feature from the time step t'. Then, to enhance the consistency of the latent audio feature and the text embedding of metadata, the predicted audio features from different time steps are compared via a loss function characterizing the latent consistency as follows

$$\mathcal{L}_{\mathcal{G}} = \mathbb{E}_{\mathcal{G}} \|\hat{\mathbf{x}}_0 - \bar{\mathbf{x}}_0\|_1^1. \tag{7}$$

This loss function minimizes the prediction errors with the CE-TTA generation model, enabling it to learn the relationship between metadata and acoustic features and leverage metadata to guide machine-sound generation. Furthermore, compared to the standard TTA model (e.g., AudioLDM [36]) that generates sounds with hundreds of time steps [36], the CE-TTA model can generate sounds with only a few time steps, due to the enhanced prediction capability, significantly improving the generation efficiency, as discussed in Section IV-C3.

During fine-tuning, the CE-TTA generator is trained on paired metadata-audio examples, which comprise normal and anomalous recordings from reference machine types, and normal recordings from new machine types, without accessing real anomalies from the new machine types. The constraint on latent-space consistency encourages stable learning of the relationship between metadata and acoustic characteristics across different noise levels, allowing anomaly-related variations to be transferred to new machine types. The generated anomalous sounds are not intended to replicate specific fault patterns exactly, but rather to capture informative anomaly-related deviations that support parameter tuning.

3) Machine Sound Generation: After fine-tuning, the CE-TTA model can be used to generate normal and anomalous machine sounds, based on input metadata captions. For new machine types, only normal metadata captions can be directly obtained. To simulate metadata captions of anomalous sounds, normal metadata captions are modified by replacing "normal" with "anomaly". This allows our framework to generate both normal and anomalous machine sounds without requiring actual anomalous data of new machine types. Then, generated normal sounds support detector training, while generated normal and anomalous sounds provide validation data for parameter tuning.

This capability arises from the learned metadata–acoustic relationship, enabling CE-TTA to transfer discriminative differences between normal and anomalous sounds from reference to unseen machine types through metadata conditioning. Rather than synthesizing physically accurate faults, the goal is to capture anomaly-related deviations that distinguish normal from anomalous acoustic distributions, providing anomaly-aware cues for parameter tuning and decision-boundary calibration beyond conventional augmentation.

Gen-ASD generates anomaly-conditioned sounds as surrogate validation signals for parameter tuning without requiring

exact fault reproduction. Unlike conventional methods that perturb normal sounds through frequency shifting or noise injection [38], Gen-ASD learns normal–anomalous acoustic relationships from reference machines and transfers them to unseen machines via metadata conditioning, enabling parameter selection without real anomalies. Section IV-D1 evaluates the effectiveness of this validation strategy.

B. First-Shot Unsupervised ASD with Generated Sounds

In this section, we apply the proposed Gen-ASD framework to two representative methods, i.e., the non-deep-learning-based TWFR-GMM [19] and the deep learning-based AnoPatch [29], resulting in the Gen-TWFR-GMM and Gen-AnoPatch models with improved detection performance in first-shot scenarios.

The machine sounds generated by Gen-ASD are partitioned into two subsets: $\widetilde{\mathcal{D}}_{\text{train}}$, containing only generated normal sounds for the expansion of the training set, and $\widetilde{\mathcal{D}}_{\text{val}}$, containing generated normal and anomalous sounds for validation and parameter tuning. The generated anomalous sounds are not used to train the ASD detectors. Real anomalous sounds from reference machine types are used solely to fine-tune the CE-TTA generator (Section III-A), and the ASD detectors for unseen machine types are trained without anomalous samples.

On the one hand, $\widetilde{\mathcal{D}}_{\text{train}}$ is central to our domain adaptation solution. It expands the training set with generated normal sounds for under-represented target-domain conditions. We construct an expanded training set $\widetilde{\mathcal{D}}_{\text{train+}}$ as:

$$\widetilde{\mathcal{D}}_{\text{train+}} = \mathcal{D}_{\text{train}} \cup \widetilde{\mathcal{D}}_{\text{train}}, \tag{8}$$

where $\mathcal{D}_{\text{train}}$ denotes the set of original real normal sounds used for training. This improves audio representation for the target domain and mitigates source-target imbalance.

On the other hand, $\widetilde{\mathcal{D}}_{\text{val}}$ addresses the scarcity of anomalous samples in first-shot scenarios. It serves only as a proxy validation set for anomaly-aware parameter tuning when real anomalies of unseen machine types are unavailable. This parameter tuning process is performed during the construction of our Gen-TWFR-GMM, Gen-AnoPatch and Gen-UnlabeledASD models, as described below.

1) Gen-TWFR-GMM: In our Gen-TWFR-GMM model, we provide a solution to estimate the hyperparameter r with the greedy algorithms for fine-grained audio feature representation based on the generated validation set $\widetilde{\mathcal{D}}_{\text{val}}$ rather than real anomalous data, as

$$\hat{r} = \arg\max_{r} \{\text{Eval}(r, \widetilde{\mathcal{D}}_{\text{val}})\}, \tag{9}$$

where $\hat{r}$ denotes the estimated value of r, and $\text{Eval}(r, \widetilde{\mathcal{D}}_{\text{val}})$ denotes the evaluation metric on the generated validation set. The hyper-parameter r is searched over a predefined fixed range and each candidate value is evaluated on the generated validation set $\widetilde{\mathcal{D}}_{\text{val}}$.

This allows TWFR-GMM to perform first-shot unsupervised ASD without changing its original architecture. This integration results in improved feature representation and improved anomaly detection performance, while maintaining the lightweight property of TWFR-GMM with only 33K parameters. Here, to avoid potential bias between the generated machine sound and the real machine sound, only generated data $\widetilde{\mathcal{D}}_{\text{train}}$ is used for model training.

2) Gen-AnoPatch: With the generated normal sounds, our Gen-AnoPatch benefits from the balanced amount of sounds from different domains for model training to learn more fine-grained and robust audio representations under domain shift. In addition, the generated validation set enables parameter tuning for our Gen-AnoPatch model, as follows

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \{\text{Eval}(\mathcal{M}, \widetilde{\mathcal{D}}_{\text{val}})\}, \tag{10}$$

where $\hat{\mathcal{M}}$ denotes the parameters of the AnoPatch detection model, i.e., the complete set of model weights rather than a single scalar hyper-parameter. This decouples parameter tuning operation from empirical manual selection while anomalies are unavailable in first-shot scenarios, thereby improving performance and interpretability of the detection model.

3) Gen-UnlabeledASD: We further integrate Gen-ASD with the recent open-source UnlabeledASD [24]. Similar to Gen-AnoPatch, generated normal sounds complement under-represented target-domain conditions during training, while the generated validation set enables anomaly-aware parameter tuning following Eq. (10). Generated anomalous sounds are used only for validation. This integration further verifies the model-agnostic applicability of Gen-ASD to recent deep-learning ASD methods.

Overall, Gen-ASD enables both conventional and deep-learning ASD models to operate effectively in first-shot scenarios by leveraging generated sounds for training and validation. It also extends lightweight ASD methods to these scenarios, enhancing their practicality. The following section presents the experimental setup and performance comparisons.

IV. Experiments and Results

In this section, we first introduce the experimental setup and evaluation metrics. Then, we compare our Gen-ASD-based method with state-of-the-art approaches to demonstrate its effectiveness. Furthermore, we conduct extensive ablation studies to analyze the impact of key components in the proposed framework. Finally, we provide detailed analyses of the generated machine sounds, including parameter tuning validation, domain balancing effects, comparisons with data augmentation, and qualitative visualizations, to gain insights into how our Gen-ASD improves anomalous sound detection in first-shot scenarios.

A. Experimental Setup

1) Dataset: Experiments are conducted on the dataset of Task 2 in the DCASE 2023 Challenge [1], which includes seven reference machine types (Fan, Gearbox, Bearing, Slider, ToyCar, ToyTrain, Valve) and additional seven machine types (Vacuum, ToyTank, ToyNscale, ToyDrone, Bandsaw, Grinder, Shaker) as new machine types. For reference machine types, there is a validation set with 100 normal and 100 anomalous sounds of each reference machine type. The training set includes 1000 normal sounds of each machine type in both

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON DCASE 2023 CHALLENGE TASK 2 EVALUATION SET.

Method	DCASE Ranking	ToyDrone		ToyNscale		ToyTank		Vacuum		Bandsaw		Grinder		Shaker		Overall	
		AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
State-of-the-art Methods:																	
Jie_IESEFPT [49]	1	58.03	51.58	89.03	77.74	60.33	61.53	96.18	85.32	65.66	53.35	66.63	62.45	68.08	55.97	69.75	62.03
Lv_HUAKONG [50]	2	54.84	49.37	82.71	57.00	**74.80**	63.79	93.66	87.42	58.48	50.30	66.69	61.22	74.24	65.24	70.05	60.11
Jiang_THUEE [51]	3	55.83	49.74	73.44	61.63	63.03	59.74	81.98	76.42	71.10	56.64	62.18	62.41	75.99	64.68	68.03	60.71
Wilkinghoff_FKIE [52]	4	53.90	50.21	87.14	76.58	63.43	62.21	83.26	74.00	66.06	52.87	67.10	62.11	65.91	50.24	67.95	59.58
Bai_JLESS [53]	5	51.44	50.89	59.85	51.16	70.05	59.58	81.46	69.47	74.51	55.65	67.07	**63.03**	78.30	63.37	67.44	58.34
Zhou_SHNU [54]	6	61.10	55.74	62.23	52.11	68.66	59.53	77.05	63.53	69.13	51.99	69.04	61.51	68.83	55.94	67.66	56.89
Guan_HEU [41]	7	62.93	52.05	68.94	54.21	66.41	60.63	79.47	72.47	57.22	50.76	62.38	54.96	78.46	61.47	67.12	57.32
Du_NERCSLIP [55]	8	58.06	51.47	86.84	65.37	61.29	57.58	82.05	60.84	47.58	49.92	49.06	49.21	93.24	80.78	64.00	57.73
Official Baseline [44]	9	58.93	51.42	50.73	50.89	57.89	53.84	86.84	65.32	69.10	57.54	60.19	59.55	72.28	62.33	63.41	56.82
SSL4ASD [22]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	72.60	61.60
AnoPatch [29]	–	61.61±2.87	51.93±1.77	77.65±3.46	65.20±5.41	65.28±3.80	56.33±4.04	77.86±1.51	67.12±5.07	63.96±3.51	54.71±2.03	70.04±3.28	59.41±2.80	80.87±2.44	77.42±3.49	70.21±1.16	60.58±1.35
UnlabeledASD [24]	–	56.23±1.95	**56.48**±1.37	90.83±1.89	77.97±3.28	61.00±2.21	60.17±1.31	94.27±1.55	80.33±2.34	60.55±1.72	55.17±0.93	59.05±0.81	61.99±2.17	90.09±2.93	75.70±2.47	69.72±0.47	65.34±0.19
Our Models:																	
Gen-TWFR-GMM	–	56.50	52.21	57.35	50.26	65.89	59.37	**97.62**	89.05	**74.60**	53.79	64.29	57.25	79.33	72.68	70.80	62.09
Gen-AnoPatch	–	**63.81**±3.80	52.65±2.88	82.09±4.57	65.43±5.68	63.77±4.74	58.66±2.99	84.09±3.19	71.83±5.25	66.07±3.34	**59.17**±2.30	**70.34**±2.86	59.97±1.37	80.72±1.47	78.25±1.69	71.90±1.16	62.59±1.24
Gen-UnlabeledASD	–	57.55±3.83	55.86±0.61	**93.11**±0.83	**82.95**±2.48	66.39±2.64	**64.16**±1.82	94.79±2.04	**89.24**±4.33	59.14±3.58	54.04±1.15	62.79±1.65	59.01±0.96	**95.83**±1.70	**87.88**±2.90	**72.07**±1.08	**67.54**±0.54

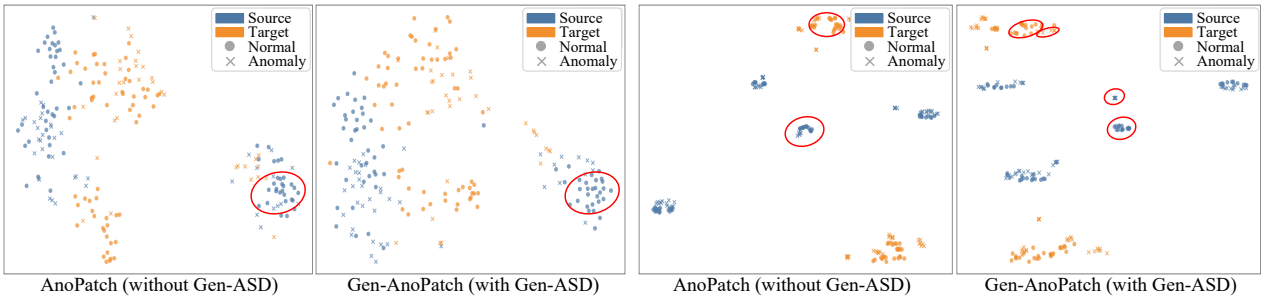


Fig. 4. The t-SNE visualization of the audio embeddings from the pooling layer before applying the KNN detector of AnoPatch (without our Gen-ASD) and Gen-AnoPatch (with our Gen-ASD) on new machine types, e.g., Bandsaw and Vacuum. Here, (a) represents the t-SNE visualizations of Bandsaw, and (b) denotes the t-SNE visualizations of Vacuum.

reference and new machine types. In the training set, the ratio of source to target domains is 99:1. The evaluation set consists of 200 sound samples per new machine type, whose status (normal or anomalous) and operating conditions are unknown.

2) Evaluation Metric: Detection performance is evaluated using the area under the receiver operating characteristic curve (AUC) and partial AUC (pAUC), where pAUC measures performance over a low false-positive rate range like $[0, 0.1]$ as in [1], [56]. Meanwhile, the AUC values of the source domain and the target domain are calculated respectively to evaluate the performance under domain shift, termed as source AUC and target AUC [13]. Following [29], the overall performance across all machine types is quantified using the harmonic mean of source AUC, target AUC, AUC and pAUC, referred to as the total score. A larger value for all metrics indicates better ASD performance.

3) Implementation Details: In our proposed framework, the CE-TTA model is initialized by the pretrained parameters of AudioLCM [37]. Following the official AudioLCM training pipeline, the entire model is fine-tuned in our implementation. Specifically, the model is fine-tuned using the constructed machine sound-metadata caption pairs, with a learning rate of 3e-6, batch size of 8, and 100 epochs. The optimizer is AdamW, and the experiments are conducted on 8 GPUs. No parameter-efficient modules such as low rank adaptation (LoRA) [57] or adapter tuning [58] are used in the current implementation. The generation operation is based on the denoising strategy of the denoising diffusion implicit models (DDIM) [59] and the DDIM time step is empirically set as 10, to balance the time cost of machine sound generation and the enhancement for detection performance.

To balance the number of sound signals in the source domain and target domain, we employ the generated normal sounds to adjust the ratio of the source and target domain from 99:1 to 1:1. During unsupervised ASD model training, the settings for TWFR-GMM in our Gen-TWFR-GMM and AnoPatch in our Gen-AnoPatch follow their original configurations [28], [29].

B. Performance Analysis for First-Shot ASD

1) Comparison with State-of-the-art Methods: To evaluate the effectiveness and generalizability of the proposed Gen-ASD framework for first-shot ASD, we compare Gen-TWFR-GMM, Gen-AnoPatch, and Gen-UnlabeledASD with the official baseline [44], top-ranked submissions from the DCASE 2023 Challenge Task 2 [41], [49]–[55], and representative first-shot ASD methods, including SSL4ASD [22], AnoPatch [29], and the recent open-source UnlabeledASD [24]. For the reproduced deep-learning methods, we conduct five independent runs with different random seeds and report the mean and standard deviation of AUC and pAUC in Table II.

The compared methods incorporate various augmentation strategies to improve training diversity and representation robustness. Mixup is used in SSL4ASD [22] and several top-ranked DCASE systems [49], [52], [53], [55]; SpecAugment is adopted by AnoPatch [29] and the third-place system [51]; temporal and frequency shifts are employed in [50], [54]; and SMOTE is used in our TWFR-GMM-based challenge submission [41]. Gen-AnoPatch retains the original SpecAugment configuration, allowing us to assess the additional performance benefit of integrating Gen-ASD into an augmentation-equipped ASD baseline.

Overall, Gen-ASD improves the mean performance of both AnoPatch and the recent open-source UnlabeledASD. Specifically, Gen-AnoPatch raises the overall AUC and pAUC from $70.21\% \pm 1.16\%$ and $60.58\% \pm 1.35\%$ to $71.90\% \pm 1.16\%$ and

TABLE III
COMPUTATIONAL COMPLEXITY AND PERFORMANCE COMPARISON OF GEN-ASD-BASED MODELS WITH OTHER STATE-OF-THE-ART METHODS.

Method	DCASE Ranking	Parameters for Training (one-off cost)	Parameters for Detection	Average AUC
Jie_IESEFPT [49]	1	3M	3M	69.75
Lv_HUAKONG [50]	2	300M	300M	70.05
Jiang_THUEE [51]	3	6M	6M	68.03
Wilkinghoff_FKIE [52]	4	34M	34M	67.95
Guan_HEU [41]	7	33K (TWFR-GMM) + 792M (AudioLDM)	**33K**	67.12
AnoPatch [29]	–	91M	91M	71.96
SSL4ASD [22]	–	34M	34M	72.60
Gen-TWFR-GMM (Ours)	–	33K (TWFR-GMM) + 1.1B (AudioLCM)	**33K**	70.80
Gen-AnoPatch (Ours)	–	91M (AnoPatch) + 1.1B (AudioLCM)	91M	**73.88**

TABLE IV
ABLATION STUDY OF THE PROPOSED GEN-ASD ON THE DCASE 2023 CHALLENGE TASK 2 EVALUATION SET. "+GEN VAL" DENOTES GENERATED VALIDATION DATA ONLY, "+GEN NORMAL" DENOTES GENERATED TARGET-DOMAIN NORMAL SOUNDS ONLY, AND "+GEN-ASD (FULL)" COMBINES BOTH COMPONENTS.

Method	Source AUC	Target AUC	AUC	pAUC	Total Score
Baseline (UnlabeledASD [24])	79.30±1.33	62.21±0.85	69.72±0.47	65.34±0.19	68.57±0.37
+Gen Val	80.57±1.59	63.57±1.03	71.06±0.95	64.28±0.86	69.23±0.50
+Gen Normal	**80.80±2.46**	65.58±1.45	**72.37±0.94**	65.87±1.02	70.62±0.68
+Gen-ASD (Full)	79.75±2.20	**65.77±1.43**	72.07±1.08	**67.54±0.54**	**70.88±0.83**

$62.59\% \pm 1.24\%$, while Gen-UnlabeledASD improves them from $69.72\% \pm 0.47\%$ and $65.34\% \pm 0.19\%$ to $72.07\% \pm 1.08\%$ and $67.54\% \pm 0.54\%$, respectively, achieving the highest overall pAUC. Despite minor metric-wise variations on individual machine types, the repeated-run results support the robustness and cross-backbone applicability of Gen-ASD.

To further understand these improvements, we visualize the learned audio representations using the t-SNE visualization for both AnoPatch [29] and our Gen-AnoPatch on Bandsaw and Vacuum types, shown in Figure 4. On Bandsaw, our Gen-AnoPatch enhances the separation between normal and anomalous samples, as evident from the distinct clusters, compared to the more entangled patterns in the red circle of Figure 4(a). Similarly, for Vacuum, our Gen-AnoPatch reduces overlap between normal and anomalous sounds in both source and target domains and yields clearer anomaly clustering in Figure 4(b), which validates the improved audio representation ability enabled by our Gen-ASD.

Meanwhile, our Gen-TWFR-GMM also achieves an overall AUC of 70.80% and pAUC of 62.09% with a lightweight non-deep-learning detector. Combined with the consistent gains observed for AnoPatch and UnlabeledASD, this demonstrates that Gen-ASD can be applied to both deep and non-deep ASD detectors without architectural modification, supporting its general applicability to first-shot ASD under domain shift.

2) Efficiency and Complexity Analysis: To further assess the efficiency and model complexity of the proposed Gen-ASD framework, we compare the parameter counts and detection performance of Gen-TWFR-GMM and Gen-AnoPatch with other state-of-the-art methods, as summarized in Table III. The AnoPatch and Gen-AnoPatch performance values in Table III are selected single-run results with relatively high detection performance, rather than the five-run means reported in Table II. The comparison shows higher AUC scores for the evaluated GMM-based and AnoPatch-based systems without increasing the parameter count of the detector. These results support the applicability of Gen-ASD to the evaluated detection backbones and datasets.

TABLE V
PERFORMANCE COMPARISON UNDER DOMAIN SHIFT BETWEEN STATE-OF-THE-ART METHODS AND OUR GEN-ASD-BASED MODELS.

Method	with Gen-ASD	Source AUC	Target AUC	Average AUC	Average pAUC	Total Score
Max-GMM [19]	✗	70.30	43.67	57.38	54.34	54.83
Mean-GMM [19]	✗	**83.12**	53.41	68.80	58.81	64.21
Gen-TWFR-GMM	✓	83.11	**57.72**	**70.80**	**62.09**	**67.13**
AnoPatch [29]	✗	78.74	69.60	71.96	61.42	69.87
Gen-AnoPatch	✓	**79.10**	**72.52**	**73.88**	**63.07**	**71.66**

TABLE VI
ABLATION OF THE CONSISTENCY-ENHANCED GENERATION STRATEGY IN OUR GEN-ASD FRAMEWORK.

Method	Audio Generator Model in our Gen-ASD	Average AUC	Average pAUC	Total Score
Mean-GMM [19]	–	68.80	58.81	64.21
FS-TWFR-GMM [28]	AudioLDM [36]	68.41	59.76	66.02
Gen-TWFR-GMM	AudioLCM [37]	**70.80**	**62.09**	**67.13**

Although Gen-ASD introduces additional parameters during the sound-generation stage, both Gen-TWFR-GMM and Gen-AnoPatch retain the same model size and inference complexity as their original ASD models. For example, Gen-TWFR-GMM requires only 33K parameters during inference, which is less than 1.1% of the parameter count of the DCASE top-1 system [49], while still achieving competitive detection performance. These results highlight the advantage of improving detection performance while retaining the size of the original detector during inference.

C. Ablation Studies

To evaluate the contributions of different components in Gen-ASD, we conduct a series of ablation studies. Specifically, we first disentangle the effects of generated anomalous validation data and generated target-domain normal sounds, and then analyze the robustness of Gen-ASD under domain shift, the impact of the consistency-enhanced generation strategy, and the influence of different metadata representations.

1) Contribution of Generated Normal and Anomalous Sounds: To disentangle the contributions of the two generated-data splits in Gen-ASD, we conduct a controlled ablation study on the UnlabeledASD backbone under the same experimental protocol. Four settings are compared: the original UnlabeledASD baseline without Gen-ASD, using only the generated validation set for anomaly-aware parameter tuning (i.e., +Gen Val), using only generated target-domain normal sounds for training (i.e., +Gen Normal), and the full Gen-ASD combining both components.

For settings using generated-anomaly-based validation, the model checkpoints obtained during normal-data training are evaluated on the generated validation set, and the checkpoint with the best validation performance is selected. For settings without parameter tuning, we adopt the final-epoch model weights as a predefined selection rule, following SSL4ASD [22] and UnlabeledASD [24].

As shown in Table IV, +Gen Val improves overall AUC and total score over the baseline, supporting its utility for anomaly-aware tuning without real anomalies of new machine types. +Gen Normal yields larger gains in target and

TABLE VII
COMPARISON ON GENERATION STEPS AND TOTAL GENERATION TIME COST BETWEEN USING THE CONSISTENCY-ENHANCED GENERATION STRATEGY (I.E., AUDIOLCM [37]) WITHOUT IT (I.E., AUDIOLDM [36]) IN THE PROPOSED FRAMEWORK.

Generator Model	Using Consistency-Enhanced Generation Strategy	Generation Step	Total Generation Time (Hours)
AudioLDM [36]	✗	150	58.83
AudioLCM [37]	✓	**10**	**2.48**

overall AUC, supporting improved representation of under-represented target-domain conditions. Full Gen-ASD achieves the highest mean target AUC, pAUC, and total score, although its mean source and overall AUC are lower than those of +Gen Normal. This indicates improved target-domain detection and low false-positive-rate (FPR) performance over $[0, 0.1]$, yielding a modest aggregate gain without consistent improvements across all metrics.

These results indicate metric-dependent benefits from combining the improved representation of target-domain conditions with anomaly-aware parameter tuning for model validation. However, the standard deviations reflect variability within each setting, not uncertainty in paired differences. Thus, the mean improvements do not establish statistically significant complementary gains. The following ablation studies further examine these functions from the perspectives of robustness under domain shift, audio generator choice, and metadata conditioning.

2) Effect for ASD under Domain Shift: We compare the performance on both source and target domains with and without Gen-ASD, as summarized in Table V. For TWFR-GMM, the corresponding baselines without Gen-ASD are Mean-GMM and Max-GMM, since TWFR-GMM cannot perform anomaly-aware parameter tuning in first-shot scenarios [19]. For Gen-AnoPatch, the baseline is the original AnoPatch [29]. The results for AnoPatch and Gen-AnoPatch in Table V are selected from single runs with relatively high performance and therefore differ from their five-run means in Table II.

The results in Table V demonstrate that integrating the proposed framework substantially enhances the generalization capability of ASD models to the target domain. Specifically, Gen-TWFR-GMM outperforms Max-GMM and Mean-GMM on target AUC by 14.05% and 4.31%, respectively, leading to a significantly higher total score compared to both of them. Similarly, Gen-AnoPatch achieves a 2.92% improvement in target AUC over AnoPatch, resulting in a clear advantage in overall performance. These results confirm that the proposed framework effectively improves both the generalization ability and robustness of ASD models under domain shift. In practical scenarios, domain shift often arises from uncertainties such as variations in operating conditions, recording positions, and environmental noise. The consistent improvement on the target AUC indicates that our Gen-ASD framework enhances robustness against such operational uncertainties. Importantly, our Gen-ASD framework delivers this improvement without altering the original model architecture of ASD models, and is effective for both non-deep-learning and deep-learning-based ASD models, further highlighting its broad applicability.

3) Impact of Different Generators for Machine Sound Generation: To evaluate the impact of different generators, we compare three variants: our Gen-TWFR-GMM using the consistency-enhanced generator (i.e., AudioLCM [37]), its previous version FS-TWFR-GMM [28] which employs the standard audio generator AudioLDM [36], and the baseline Mean-GMM [19] without generative data. As shown in Table VI, both generative model-based versions outperform the baseline in terms of pAUC and the total score. This clearly demonstrates that the effectiveness of our framework is not tied to a specific generator architecture. Notably, the performance gain with AudioLCM is more substantial, indicating that the choice of generator influences the degree of improvement, with AudioLCM providing a more noticeable benefit.

We further compare the generation efficiency of AudioLCM and AudioLDM in Table VII. For each setting, we generate 1,000 normal and 1,000 anomalous sounds for each machine type, resulting in 14,000 generated samples in total. AudioLCM reduces the generation steps per sample from 150 to 10 and decreases the total generation time from 58.83 hours to 2.48 hours. These results demonstrate that the consistency-enhanced generation strategy substantially improves generation efficiency, making Gen-ASD more practical for first-shot ASD deployment.

4) Impact of Different Forms of Metadata for Machine Sound Generation: To further explore the impact of different forms of metadata information used in the proposed framework, we compare the detection performance of our Gen-TWFR-GMM and Gen-AnoPatch models with different forms of metadata, e.g., original discrete metadata labels and fluent metadata captions for machine sound generation. Table VIII illustrates these two formats. The discrete labels list attribute keywords individually (e.g., "toy car", "model B2", "microphone position 1"), while captions express them in coherent natural language sentences. We apply each format as a text prompt for audio generation and evaluate the detection performance of resulting ASD models.

Table IX is specifically designed to compare different forms of metadata conditioning. Both metadata formats are evaluated using generated-validation-based parameter tuning, without accessing real anomalous validation data from the new machine types. As shown in Table IX, metadata captions consistently outperform discrete labels across both models. For Gen-TWFR-GMM, captions improve the overall AUC from 62.98% to 70.80% and pAUC from 55.07% to 62.09%. For Gen-AnoPatch, the overall AUC increases from 72.42% to 73.65% and pAUC from 63.65% to 64.85%. These results suggest that metadata captions provide more effective conditioning information for the CE-TTA generator by better preserving the contextual relationships among machine status, operating conditions, and environmental factors.

Nevertheless, it should be emphasized that the use of captions derived from metadata is not the core novelty of this work. The main contribution of our Gen-ASD lies in modeling the relationship between machine metadata and acoustic characteristics through metadata-assisted generative learning for first-shot ASD. In our framework, metadata captions mainly serve as an effective interface for expressing machine condi-

TABLE VIII
ILLUSTRATION OF DIFFERENT FORMS OF METADATA INFORMATION, INCLUDING THE METADATA CAPTION AND THE DISCRETE LABELS.

Metadata Caption	Discrete Labels
This is the ***normal*** sound of a ***toy car*** with model ***B2*** and speed ***31V***, recorded by a microphone placed at the position ***1***.	normal, toy car, model B2, speed 31V, microphone position 1
This is the ***anomaly*** sound of a ***toy car*** with model ***B2*** and speed ***31V***, recorded by a microphone placed at the position ***1***.	anomaly, toy car, model B2, speed 31V, microphone position 1
This is the ***normal*** sound of a ***grinding*** machine with grindstones ***2*** and metal plates ***2***.	normal, grinding, grindstones 2, metal plates 2

TABLE IX
PERFORMANCE WHEN USING DISCRETE METADATA LABELS AND METADATA CAPTIONS FOR AUDIO GENERATION IN OUR MODELS.

Method	Text Prompt	ToyDrone		ToyNscale		ToyTank		Vacuum		Bandsaw		Grinder		Shaker		Overall	
		AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
Gen-TWFR-GMM	Discrete Labels	55.49	50.89	52.54	**50.68**	**70.02**	55.68	78.81	72.26	66.11	52.01	63.05	54.92	54.82	49.05	62.98	55.07
	Metadata Captions	**56.50**	**52.21**	**57.35**	50.26	65.89	**59.37**	**97.62**	**89.05**	**74.60**	**53.79**	**64.29**	**57.25**	**79.33**	**72.68**	**70.80**	**62.09**
Gen-AnoPatch	Discrete Labels	**68.56**	**54.26**	75.89	66.89	**67.79**	**63.42**	78.41	72.11	64.83	57.83	**75.72**	59.97	78.37	76.97	72.42	63.65
	Metadata Captions	66.84	51.26	**84.82**	**68.95**	66.22	61.16	**80.08**	**76.63**	**68.65**	**64.13**	72.16	**62.24**	**81.41**	**77.74**	**73.65**	**64.85**

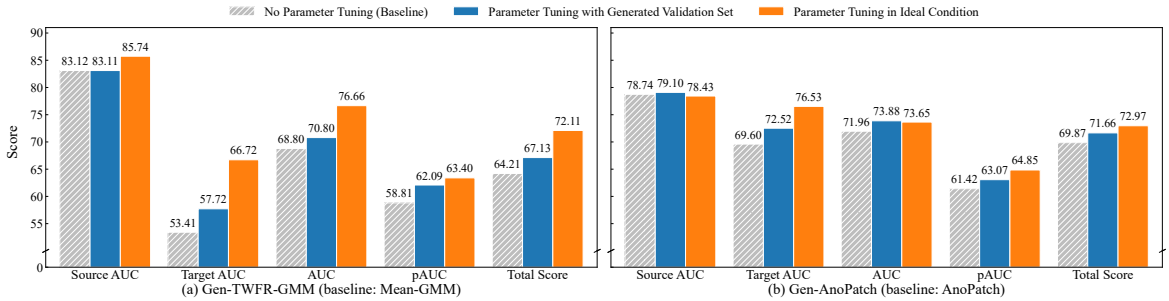


Fig. 5. Performance comparison of Gen-TWFR-GMM and Gen-AnoPatch under three settings: no parameter tuning (using Mean-GMM and AnoPatch as the respective baselines), parameter tuning with the generated validation set, and parameter tuning under the ideal condition where real normal and anomalous validation sounds are available. Performance is evaluated in terms of source AUC, target AUC, AUC, pAUC, and total score.

TABLE X
PERFORMANCE COMPARISON BETWEEN TWFR-GMM [19] TUNED WITH REAL ANOMALIES AND GEN-TWFR-GMM TUNED WITH GENERATED ANOMALIES, EVALUATED ON THE DCASE 2022 TASK 2 BENCHMARK.

Machine	Source AUC		Target AUC		pAUC	
	TWFR-GMM	Gen-TWFR-GMM	TWFR-GMM	Gen-TWFR-GMM	TWFR-GMM	Gen-TWFR-GMM
ToyCar	**93.35**	89.33	**73.64**	71.85	59.51	**60.25**
ToyTrain	**91.41**	89.88	**51.27**	36.98	**60.35**	53.53
Bearing	62.09	**65.78**	**85.37**	82.27	55.32	**59.11**
Fan	78.65	**81.86**	66.16	**67.61**	63.39	**64.58**
Gearbox	81.66	**82.61**	76.91	**77.10**	65.79	**65.98**
Slider	**90.38**	90.09	**77.70**	74.06	**68.54**	63.88
Valve	**96.13**	94.86	**90.47**	90.08	**72.88**	68.37
Average	84.81	**84.92**	**74.50**	71.42	**63.68**	62.24

tions in a form compatible with text-to-audio generators. Although captions obtained from metadata are empirically more effective than discrete labels in our experiments, they primarily function as a practical semantic representation for conditioning the text-to-audio generator, rather than constituting the central contribution of our Gen-ASD.

D. Analysis of Machine Sound Generation for First-Shot ASD

In this section, to illustrate the advantages of the machine sound generation in our framework for first-shot ASD task, we first compare parameter tuning using real data v.s. generated data to demonstrate the effectiveness of the parameter tuning with synthetic samples, as illustrated in Figure 5 and Table X. Furthermore, we analyze the influence of the amount of generated data on the ASD model's capability to mitigate domain shift in Table XI. Finally, we show the effectiveness of metadata-assisted sound generation based on the correlation between physical metadata description and acoustic characteristics with a t-SNE visualization of both generated and original machine sounds in Figure 6.

1) Validation of Parameter Tuning with Generated Sounds: This experiment evaluates whether anomalous sounds generated by Gen-ASD can serve as surrogate validation data for anomaly-aware parameter tuning when real anomalies of new machine types are unavailable. We conduct two complementary analyses using Figure 5 and Table X.

Figure 5 compares three settings: no parameter tuning, parameter tuning with the generated validation set, and parameter tuning under the ideal conditions using real normal and anomalous validation sounds of the new machine types. Mean-GMM and AnoPatch serve as the respective baselines without parameter tuning. The AnoPatch-based results in Figure 5 are selected single-run observations with relatively high detection performance. The AnoPatch baseline uses the final-epoch weights, whereas generated-validation and ideal tuning select the checkpoint with the best performance on the generated and real validation sets, respectively. The ideal setting serves only as an oracle reference and is not available under first-shot constraints. For both frameworks, tuning with the generated validation set improves the target AUC, pAUC, and total score over the untuned baselines, bringing overall performance closer to that achieved under ideal tuning conditions. Although improvements are not consistent across all metrics, these results demonstrate that generated validation sounds provide useful guidance for parameter selection and improve overall detection performance when real anomalous validation data are unavailable.

To further validate this effect, we conduct experiments on the DCASE 2022 Task 2 dataset [13], where real anomalous samples are available for all machine types. Table X compares

TABLE XI

IMPACT OF DIFFERENT AMOUNTS OF GENERATED NORMAL SOUNDS ON ASD PERFORMANCE. DIFFERENT SOURCE-TO-TARGET SAMPLE RATIOS ARE EVALUATED, RANGING FROM THE ORIGINAL IMBALANCED SETTING (SOURCE:TARGET = 99:1) TO THE BALANCED SETTING (SOURCE:TARGET = 99:99).

Method	Source:Target	ToyDrone		ToyNscale		ToyTank		Vacuum		Bandsaw		Grinder		Shaker		Overall	
		AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
AnoPatch [29]	99:1	61.72±4.35	52.66±2.20	77.51±4.54	64.34±7.10	64.57±4.02	55.93±4.05	78.39±2.22	68.33±3.43	64.16±3.11	58.21±3.87	**72.23±1.51**	61.27±1.21	80.03±1.88	77.66±2.43	70.40±1.73	61.52±1.17
Gen-AnoPatch	99:11	62.79±2.24	53.10±2.34	**82.75±4.14**	69.82±4.58	64.25±4.00	56.23±4.28	79.45±3.03	68.45±3.96	65.53±2.21	57.67±1.60	70.33±1.59	60.40±1.75	80.53±1.77	76.53±3.27	71.30±0.79	62.09±0.74
	99:22	62.00±3.08	52.38±2.17	75.58±3.95	66.61±6.43	**65.38±4.19**	56.22±3.45	79.57±3.98	68.25±3.26	66.41±2.77	56.78±2.57	69.61±2.38	59.54±2.89	80.64±1.89	76.66±2.86	70.80±0.86	61.26±1.27
	99:33	63.96±3.49	54.15±1.50	81.00±2.41	67.00±4.63	64.89±2.43	56.21±3.62	79.36±2.29	65.75±5.71	64.90±3.52	57.19±2.25	70.24±3.21	59.93±1.99	81.35±1.62	78.48±1.89	71.39±1.22	61.65±1.18
	99:44	64.50±4.51	53.29±1.67	81.62±2.54	**69.89±3.20**	64.35±4.44	56.60±4.18	80.94±3.79	65.80±5.74	65.73±2.66	55.61±3.00	70.32±4.31	61.62±3.12	80.97±2.00	79.00±1.68	71.70±1.50	61.94±1.59
	99:55	62.85±3.27	51.57±2.52	79.04±4.31	67.10±6.91	64.31±3.70	57.84±1.87	79.93±2.91	65.71±4.47	65.23±2.50	58.78±2.10	70.25±2.99	**61.65±2.46**	81.38±1.27	79.16±1.65	70.97±1.14	62.00±1.52
	99:66	**65.62±3.81**	53.67±1.30	81.27±5.01	69.60±4.91	64.57±2.62	57.21±2.28	80.38±1.98	65.74±5.37	**66.57±3.12**	58.82±4.11	69.42±1.53	58.14±1.37	80.71±1.63	76.44±3.84	71.86±1.12	61.82±1.06
	99:77	64.57±4.14	53.79±2.36	80.93±3.62	68.50±5.49	64.04±2.14	58.05±1.98	81.37±3.13	69.79±4.65	65.50±2.96	57.78±2.48	70.58±4.07	58.90±2.82	**82.58±1.26**	**79.37±1.92**	71.84±1.06	62.59±1.18
	99:88	64.74±2.27	**54.66±2.48**	79.94±4.62	66.33±4.47	65.16±6.04	56.89±4.28	81.86±4.39	70.52±5.97	63.95±1.84	56.00±2.98	70.67±2.54	58.51±2.33	79.60±1.97	76.39±2.25	71.39±1.50	61.70±1.70
	99:99	63.42±3.22	52.12±2.79	82.33±4.69	66.88±5.17	64.22±4.84	**59.34±3.02**	**83.18±3.46**	**71.71±5.11**	66.44±3.51	**59.88±3.08**	69.93±2.23	60.05±1.51	80.86±1.50	78.02±1.63	**71.86±1.08**	**62.88±1.55**

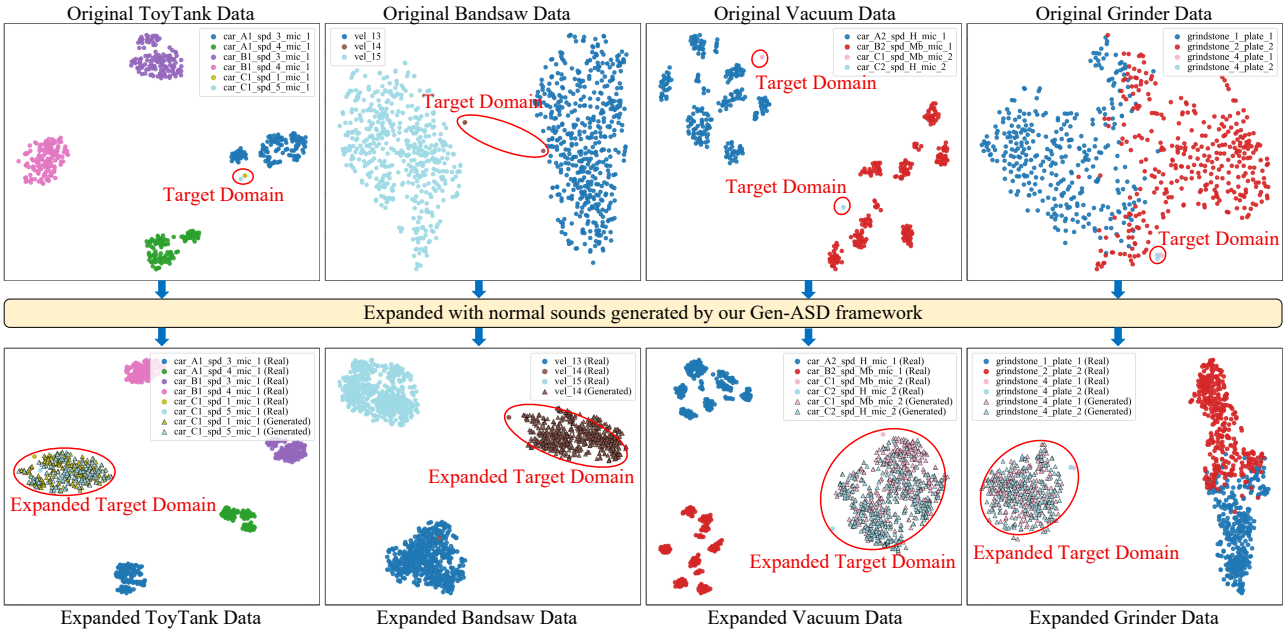


Fig. 6. The t-SNE illustrations of the audio embeddings from the pooling layer before KNN detector of original training data and expanded training data with our Gen-ASD. Each column represents a specific machine type, e.g., ToyTank, Bandsaw, Vacuum, and Grinder, while colors indicate different operating conditions recorded in the metadata. The circular dots represent the real machine sound (i.e., Real), and the triangular dots represent the generated machine sound (i.e., Generated).

the original TWFR-GMM [19], tuned with real anomalies, with Gen-TWFR-GMM, tuned with generated anomalies. Gen-TWFR-GMM achieves comparable performance on several machine types and outperforms TWFR-GMM in some cases. Although the average performance is not uniformly improved across all metrics, the results further confirm that generated anomalies can provide informative surrogate validation cues for parameter tuning.

Overall, these analyses support the use of metadata-assisted generated anomalous sounds for anomaly-aware parameter tuning when real anomalous validation data are unavailable.

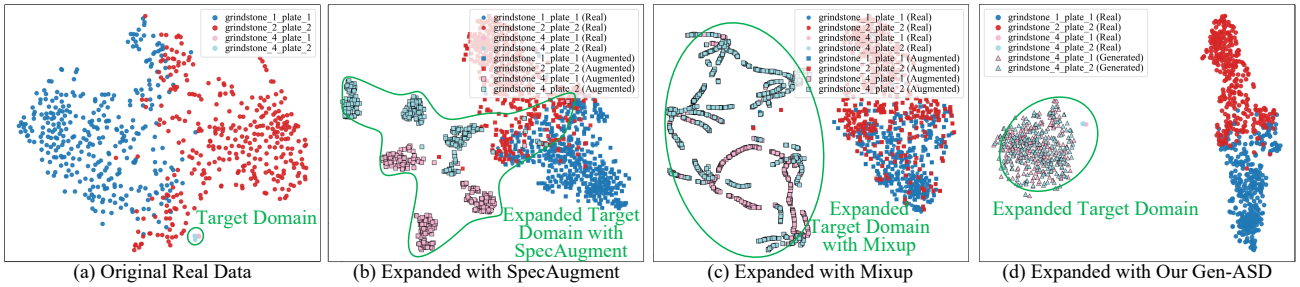
2) Different Amount of Generated Sound for First-shot ASD: To investigate the influence of generated target-domain normal sounds, we evaluate Gen-AnoPatch under different source-to-target ratios, ranging from the original imbalanced setting (99:1) to the balanced setting (99:99). The 99:1 setting contains no generated target-domain data and therefore corresponds to the original AnoPatch. To reduce the influence of random variation, we repeat the experiments with multiple random seeds and report the mean performance and standard deviation in Table XI.

As shown in Table XI, the performance does not vary monotonically with the amount of generated data, and moderate fluctuations occur between adjacent ratios. Nevertheless, an overall improvement is observed after complementing the under-represented target domain. Compared with the original 99:1 setting, most generated-data settings achieve higher overall AUC and pAUC, while larger target-domain proportions generally maintain stronger overall performance. In particular, the balanced 99:99 setting improves the overall AUC from $70.40\% \pm 1.73\%$ to $71.86\% \pm 1.08\%$ and pAUC from $61.52\% \pm 1.17\%$ to $62.88\% \pm 1.55\%$.

These results indicate that generated normal sounds can improve the representation of under-represented target-domain conditions and contribute to ASD performance, although the exact performance varies with the amount of generated data rather than following a strictly monotonic trend.

3) Visualization Analysis: To better understand why our framework can address the first-shot ASD by generating machine sounds based on the physical information within machine metadata, Figure 6 provides t-SNE visualizations of audio features from both the original training set and the dataset augmented with generated normal sounds from our framework, including machine types like ToyTank, Bandsaw, Vacuum, and Grinder.

In the upper row, which visualizes the original training data, target domain samples are notably scarce and form distinct clusters, separated from the rest of audio features. This sepa-



(a) Original Real Data (b) Expanded with SpecAugment (c) Expanded with Mixup (d) Expanded with Our Gen-ASD

Fig. 7. A representative t-SNE visualization of the audio feature distribution for the machine type Grinder. (a) shows the original real data, where the target-domain samples are under-represented as indicated in the green circle. (b) and (c) show data expanded using conventional data augmentation methods, i.e., SpecAugment and Mixup, where the augmented samples form scattered and inconsistent distributions that deviate from the real target-domain structure. (d) shows data expanded with the proposed Gen-ASD, where the generated samples are better aligned with the real target-domain distribution and form a coherent cluster, indicating improved coverage of under-represented target-domain conditions.

TABLE XII
PERFORMANCE COMPARISON BETWEEN MIXUP, SMOTE AUGMENTATION, AND OUR GEN-ASD WITH THE ANOPATCH BACKBONE ON THE DCASE 2023 CHALLENGE TASK 2 EVALUATION SET, WITH THE SAME NUMBER OF ADDED TARGET DOMAIN NORMAL SAMPLES.

Backbone	Strategy	ToyDrone		ToyNscale		ToyTank		Vacuum		Bandsaw		Grinder		Shaker		Overall	
		AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
AnoPatch [29]	Mixup	56.13±3.01	50.48±0.51	**87.60±2.16**	**71.88±2.68**	62.49±1.92	52.71±1.30	**89.45±1.23**	**78.22±1.84**	58.84±4.92	53.76±2.45	68.05±2.86	54.64±3.60	72.80±8.62	64.05±6.60	68.55±1.78	59.25±2.04
	SMOTE	54.70±3.52	**54.42±2.69**	74.09±2.47	59.79±2.80	60.44±2.06	**59.80±1.61**	63.58±1.81	64.32±2.47	64.25±3.31	52.63±1.46	68.34±3.65	59.12±1.99	70.07±5.04	66.12±7.70	64.38±0.81	58.99±0.68
	Gen-ASD (Ours)	**63.81±3.80**	52.65±2.88	82.09±4.57	65.43±5.68	**63.77±4.74**	58.66±2.99	84.09±3.19	71.83±5.25	**66.07±3.34**	**59.17±2.30**	**70.34±2.86**	**59.97±1.37**	**80.72±1.47**	**78.25±1.69**	**71.90±1.16**	**62.59±1.24**

ration is primarily due to variations in operating conditions and inherent physical differences captured in the machine metadata (e.g., different operating speeds and environmental conditions). As a result, there is a significant imbalance in the distribution of training samples, with the target domains being severely underrepresented, resulting in the domain shift.

In contrast, the lower row shows the training data after expansion with generated normal sounds from our Gen-ASD framework. Here, we generate normal machine sounds with the same metadata information of the real sounds in the target domain. As a result, the target domain clusters are significantly denser and more populated, indicating a successful enhancement for audio feature representations. Notably, the generated samples align closely with the real samples from the target domain in feature space, forming coherent, overlapping clusters. This demonstrates that our Gen-ASD effectively captures the relationship between the physical operating status within machine metadata and the underlying acoustic characteristics. These results confirm that Gen-ASD not only alleviates data imbalance by generating realistic, domain-specific audio features but also enhances the representation of rare operational conditions. By bridging metadata and physical acoustic properties, our framework helps stabilize the feature distribution under domain shift, contributing to improved robustness of ASD models in first-shot scenarios.

4) Difference between Gen-ASD and Data Augmentation: To further clarify the distinction between the proposed Gen-ASD and conventional data augmentation techniques, we compare the feature distributions produced by different data expansion strategies in Figure 7. As shown, the target-domain features in Figure 7(a) form a coherent cluster, but are scattered as several clusters, which reflects their inconsistency after applying data augmentation methods (e.g., SpecAugment and Mixup) in Figure 7(b) and (c). This suggests that such augmentations disrupt the original feature distribution of target-domain. In contrast, our Gen-ASD can generate normal sounds with the same operating conditions as in the target domain, which can be clustered with the sounds from the original target domain as a coherent entirety in Figure 7(d), rather than several clusters with inconsistent distribution. This verifies that our Gen-ASD can effectively strengthen feature consistency and mitigate domain shift, by leveraging the relationship between acoustic properties and metadata.

Moreover, we compare Gen-ASD with Mixup and SMOTE using the same AnoPatch backbone and the same number of added target-domain normal samples, as shown in Table XII. Gen-ASD selects the checkpoint with the best performance on its generated validation set, whereas Mixup and SMOTE use the final-epoch checkpoints as a predefined selection rule, without generated-anomaly-based validation. Gen-ASD achieves higher overall AUC and pAUC than both alternatives, although Mixup and SMOTE perform better on some individual machine types or metrics. These results demonstrate that the use of the Gen-ASD framework performs better than simply expanding the target-domain training set with Mixup or SMOTE, which is consistent with the observations of feature-distributions in Figure 7. This comparison evaluates the joint benefits of improved representations of target domain conditions and anomaly-aware checkpoint selection, rather than isolating normal-sound generation, whose contribution is separately assessed in the ablation results shown in Table IV.

5) Qualitative Analysis: To further analyze the generated samples, we provide spectrogram-level analysis in Figure 8. As shown in Figure 8, real normal and anomalous spectrograms exhibit distinguishable differences in time-frequency pattern. In addition, the generated normal and anomalous samples present similar pattern differences (highlighted in blue boxes). Furthermore, this phenomenon is consistently observed across different machine types (e.g., Vacuum and Shaker), indicating transferable anomaly-related acoustic patterns in the evaluated machine types. This consistency suggests that Gen-ASD captures acoustic variations associated with normal and anomalous states.

Moreover, the generated sounds exhibit feature distributions

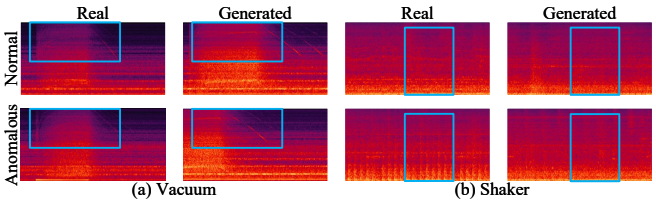


Fig. 8. Signal-level comparison of spectrograms between real and generated machine sounds under normal and anomalous conditions. The generated samples show distinguishable temporal-spectral differences between normal and anomalous states, providing qualitative evidence that Gen-ASD captures structured acoustic deviations rather than simple random perturbations.

consistent with the real target-domain sounds, which is also supported by the coherent clustering observed in Figure 7. Together with the spectrogram-level differences between generated normal and anomalous samples, these observations support the modeling of target-domain acoustic characteristics and anomaly-related variations for constructing surrogate validation samples.

V. Conclusion

We presented Gen-ASD, a metadata-assisted framework for first-shot unsupervised ASD under domain shift. By learning the relationships between machine metadata (e.g., machine states, operating conditions, and environmental factors) and acoustic characteristics, Gen-ASD generates condition-aware normal sounds to improve target-domain representations and anomalous sounds for parameter tuning, enabling tuning-dependent detectors to operate on unseen machine types without real anomalies. Experiments demonstrate its effectiveness across the evaluated deep-learning and non-deep-learning detectors and datasets, with additional evidence supporting generated anomalies as surrogate validation data. These learned relationships may also enable condition-controllable generation for rare-condition augmentation, proxy validation, and robustness assessment, although their physical validity remains unverified. Gen-ASD requires informative metadata and relevant reference acoustic data, and substantial differences in machine mechanisms, fault patterns, or operating and recording conditions may limit transfer. Future work will address incomplete metadata and investigate how reference-machine selection and acoustic conditions affect transferability.

References

[1] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida *et al.*, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, 2023, pp. 31–35.

[2] Z. Mnasri, S. Rovetta, and F. Masulli, "Anomalous sound event detection: A survey of machine learning based methods and applications," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5537–5586, 2022.

[3] J. Wu, F. Yang, and W. Hu, "Unsupervised anomalous sound detection for industrial monitoring based on ArcFace classifier and Gaussian mixture model," *Appl. Acoust.*, vol. 203, p. 109188, 2023.

[4] Y. Wang, Y. Zheng, Y. Zhang, Y. Xie, S. Xu, Y. Hu *et al.*, "Unsupervised anomalous sound detection for machine condition monitoring using classification-based methods," *Appl. Sci.*, vol. 11, no. 23, p. 11128, 2021.

[5] K. Wilkinghoff and F. Kurth, "Why do angular margin losses work well for semi-supervised anomalous sound detection?" *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2023.

[6] E. Di Fiore, A. Ferraro, A. Galli, V. Moscato, and G. Sperlì, "An anomalous sound detection methodology for predictive maintenance," *Expert Syst. Appl.*, vol. 209, p. 118324, 2022.

[7] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 816–820.

[8] M. Khanjari, A. Azarfar, M. H. Abardeh, and E. Alibeiki, "Anomalous sound detection for machine condition monitoring using 3D tensor representation of sound and 3D deep convolutional neural network," *Multimedia Tools Appl.*, vol. 83, no. 15, pp. 44 101–44 119, 2024.

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

[10] J. Guan, Y. Liu, Q. Kong, F. Xiao, Q. Zhu, J. Tian *et al.*, "Transformer-based autoencoder with ID constraint for unsupervised anomalous sound detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 1, p. 42, 2023.

[11] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proc. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, Tokyo, Japan, November 2020, pp. 46–50.

[12] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine ID based contrastive learning pretraining," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[13] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida *et al.*, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, 2022.

[14] K. Wilkinghoff, "Design choices for learning embeddings from auxiliary tasks for domain generalization in anomalous sound detection," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Local density-based anomaly score normalization for domain generalization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 33, pp. 4642–4652, 2025.

[16] Y. Deng, J. Liu, and W.-Q. Zhang, "AITHU system for unsupervised anomalous detection of machine working status via sounding," DCASE2022 Challenge, Tech. Rep., 2022.

[17] H. Lan, Q. Zhu, J. Guan, Y. Wei, and W. Wang, "Hierarchical metadata information constrained self-supervised learning for anomalous sound detection under domain shift," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 7670–7674.

[18] J. Guan, J. Tian, Q. Zhu, F. Xiao, H. Zhang, and X. Liu, "Disentangling hierarchical features for anomalous sound detection under domain shift," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2025, pp. 1–5.

[19] J. Guan, Y. Liu, Q. Zhu, T. Zheng, J. Han, and W. Wang, "Time-weighted frequency domain audio representation with GMM estimator for anomalous sound detection," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[20] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2023, pp. 191–195.

[21] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini *et al.*, "Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, Tokyo, Japan, October 2024, pp. 111–115.

[22] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 276–280.

[23] N. Jiang, Y. Song, Q. Gu, H. Song, L. Dai, and I. McLoughlin, "An effective anomalous sound detection method based on global and local attribute mining," in *Proc. INTERSPEECH*, 2025, pp. 2620–2624.

[24] T. Fujimura, I. Kuroyanagi, and T. Toda, "Improvements of discriminative feature space training for anomalous sound detection in unlabeled conditions," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2025, pp. 1–5.

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[26] Y. Zhang, J. Liu, Y. Tian, H. Liu, and M. Li, "A dual-path framework with frequency-and-time excited network for anomalous sound detection," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 1266–1270.

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 2613–2617.

[28] H. Zhang, Q. Zhu, J. Guan, H. Liu, F. Xiao, J. Tian *et al.*, "First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 1271–1275.

[29] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen *et al.*, "AnoPatch: Towards better consistency in machine anomalous sound detection," in *Proc. INTERSPEECH*, 2024, pp. 107–111.

[30] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng *et al.*, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 1326–1330.

[31] B. Han, A. Jiang, X. Zheng, W.-Q. Zhang, J. Liu, P. Fan *et al.*, "Exploring self-supervised audio models for generalized anomalous sound detection," *IEEE Trans. Audio, Speech, Lang. Process.*, 2025.

[32] N. Jiang, Y. Song, Q. Gu, H. Song, L. Dai, and I. McLoughlin, "Finetune large pre-trained model based on frequency-wise multi-query attention pooling for anomalous sound detection," in *INTERSPEECH*, vol. 2025, 2025, pp. 3389–3393.

[33] K. Wilkinghoff, T. Fujimura, K. Imoto, and J. Le Roux, "Handling domain shifts for anomalous sound detection: A review," in *Proc. Annu. Meet. Acoust.*, 2025, pp. 101–104.

[34] K. Li, K. Zaman, X. Li, M. Akagi, J. Dang, and M. Unoki, "Machine anomalous sound detection using spectral-temporal modulation representations derived from machine-specific filterbanks," *IEEE Trans. Audio, Speech, Lang. Process.*, 2025.

[35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[36] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic *et al.*, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.

[37] H. Liu, R. Huang, Y. Liu, H. Cao, J. Wang, X. Cheng *et al.*, "AudioLCM: Efficient and high-quality text-to-audio generation with minimal inference steps," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2024, pp. 7008–7017.

[38] H. Chen, Y. Song, Z. Zhuo, Y. Zhou, Y.-H. Li, H. Xue *et al.*, "An effective anomalous sound detection method based on representation learning with simulated anomalies," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[39] H. Purohit, T. Nishida, K. Dohi, T. Endo, and Y. Kawaguchi, "MIMII-Gen: Generative modeling approach for simulated evaluation of anomalous sound detection system," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2025, pp. 571–575.

[40] H. Purohit, T. Nishida, K. Dohi, T. Endo, and Y. Kawaguchi, "MIMII-Agent: Leveraging LLMs with function calling for relative evaluation of anomalous sound detection," in *Proc. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, Barcelona, Spain, October 2025, pp. 160–164.

[41] J. Tian, H. Zhang, Q. Zhu, F. Xiao, H. Liu, X. Mei *et al.*, "First-shot anomalous sound detection with GMM clustering and finetuned attribute classification using audio pretrained model," DCASE2023 Challenge, Tech. Rep., 2023.

[42] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen *et al.*, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 5178–5193.

[43] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (ACM SIGMOD)*, 2000, pp. 427–438.

[44] 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. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, 2023, pp. 41–45.

[45] 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. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, 2021, pp. 1–5.

[46] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[47] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[48] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*. PMLR, 2023, pp. 32 211–32 252.

[49] J. Wang, J. Wang, S. Chen, Y. Sun, and M. Liu, "Anomaly sound detection system based on multi-dimensional attention module," DCASE2023 Challenge, Tech. Rep., 2023.

[50] Z. Lv, B. Han, Z. Chen, Y. Qian, J. Ding, and J. Liu, "Unsupervised anomalous detection based on unsupervised pretrained models," DCASE2023 Challenge, Tech. Rep., 2023.

[51] A. Jiang, Q. Hou, J. Liu, P. Fan, J. Ma, C. Lu *et al.*, "THUEE system for first-shot unsupervised anomalous sound detection for machine condition monitoring," DCASE2023 Challenge, Tech. Rep., 2023.

[52] K. Wilkinghoff, "Fraunhofer FKIE submission for task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," DCASE2023 Challenge, Tech. Rep., 2023.

[53] J. Yafei, B. Jisheng, and H. Siwei, "Unsupervised abnormal sound detection based on machine condition mixup," DCASE2023 Challenge, Tech. Rep., 2023.

[54] Y. Zhou and Y. Long, "Attribute classifier with imbalance compensation for anomalous sound detection," DCASE2023 Challenge, Tech. Rep., 2023.

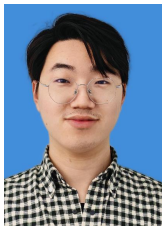
[55] L. Wang, F. Chu, Y. Zhou, S. Wang, Z. Yan, S. Xu *et al.*, "First-shot unsupervised anomalous sound detection using attribute classification and conditional autoencoder," DCASE2023 Challenge, Tech. Rep., 2023.

[56] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe *et al.*, "Description and discussion on DCASE 2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, 2020, pp. 81–85.

[57] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[58] D. Yin, X. Han, B. Li, H. Feng, and J. Bai, "Parameter-efficient is not sufficient: Exploring parameter, memory, and time efficient adapter tuning for dense predictions," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2024, pp. 1398–1406.

[59] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.



Feiyang Xiao received the B.E. and M.S. degrees from Harbin Engineering University, Harbin, China, in 2020 and 2023, respectively. He is currently working toward the Ph.D. degree with Harbin Engineering University. His research interests include intelligent audio signal processing, audio-language understanding, multimodal learning, and anomalous sound detection.



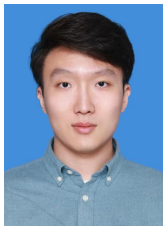
Jian Guan (Senior Member, IEEE) received the B.S. and M.S. degrees from Jilin University, China, in 2005 and 2010, respectively, and the Ph.D. degree from Harbin Institute of Technology, China, in 2018. From 2014 to 2017, he was a visiting Ph.D. researcher with the Centre for Vision, Speech and Signal Processing, University of Surrey, U.K. He is currently an Associate Professor with Harbin Engineering University, China, where he leads the Group of Intelligent Signal Processing (GISP). His research interests include intelligent audio and acoustic signal processing, anomalous sound detection, acoustic scene and event analysis, and multimodal learning. His team has achieved first-place rankings in DCASE 2024 Task 10 and BioDCASE 2026 Task 2, as well as first and second-place finishes in the IEEE ICASSP Signal Processing Grand Challenges in 2025 and 2026.



Kejia Zhang received the B.S. degree from Beijing Normal University, China, in 2004, and the Ph.D. degree from Harbin Institute of Technology, China, in 2012. He is currently an Associate Professor with Harbin Engineering University, Harbin, China. He has published several SCI-indexed papers as the first author and has led a national research project. His research interests include wireless sensor networks, the Internet of Things, privacy protection, deep learning, and computer vision.



Qiaoxi Zhu received the B.Sc. degree in acoustics from Nanjing University, Nanjing, China, in 2012, and the Ph.D. degree in acoustic signal processing from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2018. From 2015 to 2016, she was a visiting Ph.D. student with the Centre for Vision, Speech and Signal Processing, University of Surrey, UK. She is currently a Technologist (Acoustics) with the University of Technology Sydney (UTS), Sydney, Australia. Her research interests include spatial audio, acoustic measurement and metrology, artificial intelligence for audio processing, and architectural and environmental acoustics.



Wenbo Wang received the B.S. and M.S. degrees from Harbin Engineering University, Harbin, China, in 2018 and 2021, respectively, and the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2026. He is currently a Postdoctoral Researcher with Harbin Engineering University. His research interests include recommender systems and multimodal learning.



Haohe Liu (Graduate Student Member, IEEE) received the B.Eng. degree from Northwestern Polytechnical University, Xi'an, China, in 2020, and the Ph.D. degree from the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., in 2025. His first-author research has been published in leading journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE/ACM Transactions on Audio, Speech, and Language Processing, IEEE Journal of Selected Topics in Signal Processing, ICML, AAAI, ICASSP, and INTERSPEECH. His notable projects include AudioLDM, VoiceFixer, AudioSR, and NaturalSpeech. His research interests include audio quality enhancement, audio generation, source separation, and audio recognition. He is known for developing AudioLDM for text-to-audio generation, which has attracted considerable attention from the open-source community.



Wenwu Wang (M'02-SM'11-F'26) was born in Anhui, China. He received the B.Sc., M.E., and the Ph.D. degrees, all in the field of automation, from Harbin Engineering University, China, in 1997, 2000, and 2002, respectively. He then worked with King's College London, Cardiff University, Tao Group Ltd. (now Antix Labs Ltd.), and Creative Labs, before joining University of Surrey, U.K., in May 2007, where he is currently a Professor in Signal Processing and Machine Learning, and an Associate Head in External Engagement, School of Computer Science and Electronic Engineering, University of Surrey, UK. He is also a Principal AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), human-AI collaboration, and statistical anomaly detection. He has (co)-authored over 400 papers in these areas. His works have been recognized with various awards, including the Meta Distinguished Faculty Award (2026), Audio Engineering Society Best Technical Paper Award (2025), IEEE Signal Processing Society Young Author Best Paper Award (2022), DCASE Judge's Award (2020, 2023, and 2024), DCASE Reproducible System Award (2019 and 2020), and LVA/ICA Best Student Paper Award (2018). He has been elected to IEEE Fellow for contributions to audio classification, generation and source separation, since 2026. He is an IEEE Signal Processing Society Distinguished Lecturer (2027-2028). He is a Senior Area Editor (2025-2027) for IEEE Open Journal of Signal Processing and an Associate Editor (2024-2028) for IEEE Transactions on Multimedia. He was a Senior Area Editor (2019-2023) and Associate Editor (2014-2018) for IEEE Transactions on Signal Processing, and an Associate Editor (2020-2025) for IEEE/ACM Transactions on Audio Speech and Language Processing. He was the elected Chair (2023-2024) of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing (MLSP) Technical Committee, and a Board Member (2023-2024) of IEEE SPS Technical Directions Board and a Technical Directions Board Representative (2026-2028) and Executive Sub-Committee Member (2026) of the IEEE SPS Conference Board. He is currently the elected Chair (2025-2027) of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, and an elected Member (2021-2026) of the IEEE SPS Signal Processing Theory and Methods Technical Committee. He was on the organization committee of IEEE ICASSP 2019 and 2024, INTERSPEECH 2022, IEEE MLSP 2013 and 2024, and IEEE SSP 2009. He was a Technical Program Co-Chair of IEEE MLSP 2025. He has been a keynote or plenary speaker at 30+ international conferences and workshops.