MIMII-Gen: Generative Modeling Approach for Simulated Evaluation of Anomalous Sound Detection System

Harsh Purohit, Tomoya Nishida, Kota Dohi, Takashi Endo, and Yohei Kawaguchi
Research and Development Group, Hitachi, Ltd.

Abstract*—Performance evaluation is essential for developing anomalous sound detection systems. However, in many real-world settings, it is challenging to record actual anomalous sounds, making it difficult to effectively validate the system performance. To address this issue, this study proposes a novel latent diffusion model designed to generate realistic anomalous machine sounds under real-world conditions. The proposed method integrates an encoder-decoder framework with the Flan-T5 model, encoding captions derived from audio metadata to facilitate conditional audio generation using a U-Net architecture. Moreover, operating within the latent space of EnCodec, the method enables the generation of high-quality audio signals that are contextually appropriate. Evaluations of the generated audio using Fréchet Audio Distance (FAD) and other metrics demonstrate that the proposed method outperforms existing approaches, producing audio closely resembling real-world anomalies. Furthermore, when the anomalous sound detection system was evaluated using the anomalous data generated by the proposed method, the AUC score showed only a 4.8% difference compared to using real anomalous data. This confirms the effectiveness of performance validation based on the anomalous data generated by the proposed method. The audio samples can be found at https://hpworkhub.github.io/MIMII-Gen.github.io/.

***Index Terms*—Unsupervised anomalous sound detection, audio generation, latent diffusion model,**

I. Introduction

Anomalous sound detection (ASD) is vital in industrial applications, as subtle machine sound deviations can signal critical faults [1]. However, evaluating ASD models is difficult due to the scarcity of real anomalous recordings, limiting reliable performance validation. This limitation hinders the development of robust and accurate detection systems. To address this, we propose a latent diffusion model to generate realistic anomalous machine sounds. Existing generative models [2]–[4] excel in speech and music, however fail to accurately replicate the fine-grained audio differences and complexities of machine sounds due to different operational environments. Our research named as MIMII-Gen seeks to advance machine sound generation similar to recorded MIMII-DG data [5], enabling practical applications, particularly in evaluating anomaly detection systems.

To list our contributions, (i) We validate the robustness of ASD systems using generated audio anomalous data and demonstrate its effectiveness though AUC score comparison with respect to real anomalous data, (ii) To obtain various operational and environmental conditions, we enhance descriptive quality of weak metadata of sound clips by converting them into rich, human-like captions for audio generation, (iii) We carefully design the U-Net of latent diffusion model for improved guidance through conditional embeddings of captions from Flan-T5 [6]. As shown in IV-C our method outperforms current baseline generation models by producing reliable machine audio samples on Fréchet Audio Distance (FAD) and other metrics.

II. Related work

In this section, we review the recent advancements in audio generation using diffusion-based models and unsupervised anomaly detection in machine sounds.

A. Text-to-Audio Generation Using Diffusion Models

Text-to-audio (TTA) generation has garnered considerable attention, with approaches like AudioGen [7] focusing on learning audio representations by leveraging paired audio-text data to overcome the challenges of data scarcity and quality variability. AudioGen employs a Transformer-decoder to generate discrete tokens in an autoregressive manner. By implementing data augmentation techniques, such as mixing audio samples and distilling language descriptions into simplified labels, AudioGen increases the diversity of training data. However, this comes at the cost of losing intricate spatial and temporal relationships in the text descriptions, which can impact the fidelity and contextual richness of the generated audio. On the other hand, diffusion models have become a dominant framework for generative tasks, including text-to-audio conversion. Diffsound [8], uses a non-autoregressive decoder based on discrete diffusion model to generate audio using text by refining mel-spectrogram tokens through iterative steps rather than sequential predictions typical of autoregressive decoders.

Generation approaches like AudioLDM [3], Make-An-Audio [9], and Tango [4] typically employ pre-trained text encoders (e.g., CLAP [10], T5) and VAEs to extract text embeddings and latent audio features. Using a latent diffusion model (LDM) architecture, these systems generate audio latent features conditioned on text inputs, which are subsequently transformed into mel-spectrograms and waveforms using VAEs and neural vocoders.

Despite these advances, current TTA models often fall short when applied to machine sound generation, where the complexity of acoustic environments and subtle variations

ISBN: 978-9-46-459362-4

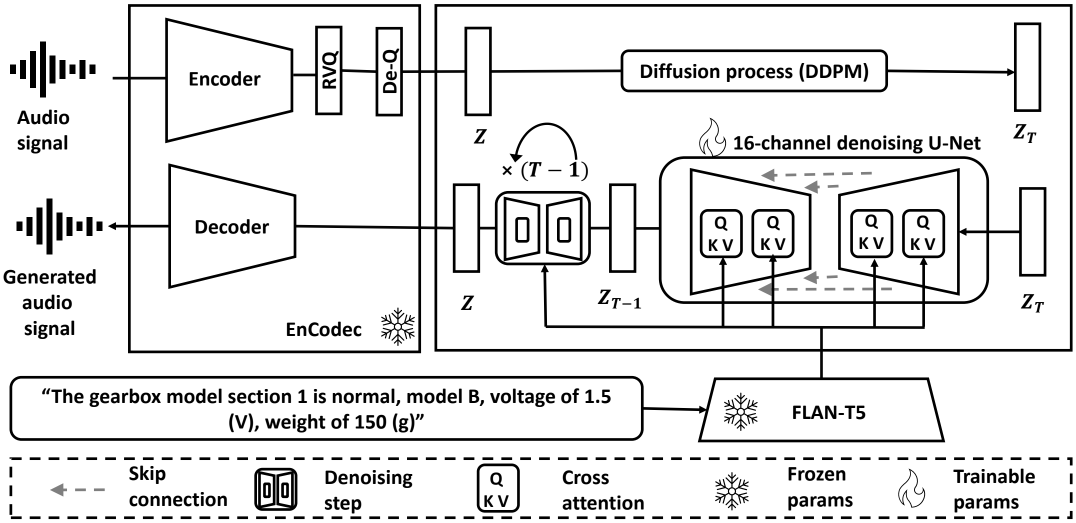


Fig. 1: Block diagram of proposed approach for machine sound generation.

in sound are critical. Current TTA methods primarily focus on speech and music, with limited exploration of industrial machine sounds. This gap underscores the need for specialized generative models tailored to the unique challenges of machine audio.

B. *Unsupervised Anomaly Detection*

Anomalous sound detection [11]–[14] aims to identify deviations from normal sounds, a task complicated by the rarity and variability of anomalous events. Traditional ASD approaches often rely on labeled anomalous data, which is scarce in real-world applications, limiting their ability to generalize to new or unseen conditions. Consequently, unsupervised ASD, which trains only on normal sounds, has emerged as a viable yet challenging alternative.

The DCASE-2023 Challenge Task-2 introduced a first-shot (FS) approach to unsupervised ASD, targeting the detection of anomalies in machine types not seen during training [15]. However, unsupervised ASD methods struggle to adapt to first-shot scenarios due to a lack of diverse training data encompassing unseen machine types and operational conditions. Limited anomalous data for evaluation also hinders adaptability and reliability in real-world scenarios where anomalies vary widely.

Recent work by Zhang et al. [16] uses generation of anomalies for training to improve anomaly detection systems, whereas we focus on generating anomalies to evaluate the robustness of existing anomaly detection system. Our approach combines generative modeling with EnCodec and Flan-T5 embeddings to produce machine audio that captures subtle variations crucial for anomaly detection. It generates diverse samples for anomaly detection evaluation when real-world industrial acoustic data is scarce.

III. Proposed approach

We propose to thoroughly evaluate existing anomaly detection systems across a wide range of operational and environmental conditions by using generated realistic and diverse anomalous machine sounds difficult to obtain in real-world settings. This synthetic data enables us to assess the robustness and effectiveness of these systems, determining how well they generalize and reliably detect anomalies even in scenarios not present in their original training data. To generate these machine sounds under various conditions, we develop a condition-based latent diffusion model. The crucial parts of the generation model as well as diffusion process are explained below.

A. *Overview of Diffusion Models*

Diffusion models are probabilistic generative models [17] that learn the data distribution $p(x)$ by progressively denoising Gaussian noise. They consist of a forward process that adds noise step-by-step in a fixed Markov chain of length T, and a reverse process that iteratively removes noise. The reverse process can be viewed as a sequence of denoising autoencoders, where $\epsilon_\theta(x_t, t)$ predicts the noise added to the noisy input x_t at time step t.

The objective function for diffusion models is:

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right], \quad (1)$$

where t is uniformly sampled from $\{1, \ldots, T\}$. This mirrors denoising score matching, enabling effective prediction of clean data from noisy observations.

To reduce computational complexity and focus on semantically relevant features, diffusion models can operate in latent space using low-dimensional representations from an encoder. The training objective in the latent diffusion framework becomes:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t,t)\|_2^2\right], \tag{2}$$

where z_t is the noisy latent variable from the encoder $\mathcal{E}$, and the reverse process is modeled using a time-conditional U-Net backbone, as shown in Fig.1.

For conditional generation modeling $p(z|y)$, the conditioning variable y is projected to a representation $\tau_\theta(y)$ via an encoder τ_θ. The loss function for the conditional latent diffusion model is:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t,t,\tau_\theta(y))\|_2^2\right], \tag{3}$$

where both the denoising network ϵ_θ and the conditioning encoder τ_θ can be jointly optimized.

B. Model Architecture

As illustrated in Fig.1, we generate captions from the metadata of audio files, which describe operational settings, environmental conditions, anomaly types, and machine models. These captions are encoded using the Flan-T5 model [6] to obtain 768-dimensional condition embeddings, serving as inputs for the diffusion model.

We employ EnCodec, an off-the-shelf VQ-GAN model, to obtain compressed latent representations of audio signals, capturing essential features efficiently. Unlike traditional methods that rely on VAEs and vocoders—requiring additional training on spectrograms [3] and models like HiFi-GAN [18]—we use EnCodec similar to AudioJourney [19], simplifying the architecture and reducing model complexity.

In EnCodec, the encoder outputs a continuous latent representation, which is converted into a discrete set of codebook indices through residual vector quantization (RVQ). The latent dimension becomes equal to the number of selected codebooks, N_q. Variable bandwidth training in EnCodec randomly selects codebooks in multiples of four, corresponding to bandwidths of 1.5, 3, 6, 12, or 24 kbps at 24 kHz. After experimenting with these options, we selected 24 kbps based on generation quality.

The discrete representation is converted back to a continuous vector by summing the corresponding codebook entries before the decoder, via the dequantization block (De-Q in Fig.1). This continuous latent vector is used as input to the diffusion process during training. During inference, the sampled latent vector is input directly to the decoder to generate the audio clip. The diffusion model is trained using the Denoising Diffusion Probabilistic Model (DDPM) framework with a wide-channel denoising U-Net, where the condition embeddings are combined with audio representations.

U-Net Design and Noise Scheduling: We employ a wide-channel U-Net with 16-channel input to effectively utilize the EnCodec encoder's latent space, differing from the typical 1 or 3 channels in other audio generation methods. To address minimal variance in the 16-dimensional latent encodings, we reshape the latent vectors from a single-channel 128×750 format to a 16-channel 8×750 format. This restructuring allows convolutional blocks to fully encompass the latent representation within their receptive fields, enhancing audio generation fidelity. Each channel is separately normalized to zero mean and unit variance, aiding the U-Net in learning the noise distribution $N(0, I)$. While inspired by AudioJourney [19], our reshaping dimensions, number of channels, and U-Net design differ due to specific data characteristics. These transformations are reversible, enabling decoding back into waveforms using the EnCodec decoder. Additionally, we use cross-attention instead of embedding addition in the U-Net architecture to preserve the original audio embeddings throughout each layer, enhancing conditional guidance.

The generated samples are used to evaluate generation quality and as input into an auto-encoder based anomaly detection system [15].

IV. Experiments

A. Dataset

We utilize the MIMII-DG dataset [5], which contains sounds of various machines recorded under different operating conditions. This dataset is employed in anomaly detection Task 2 of DCASE2023 [15], comprising three parts: development dataset, additional training dataset, and evaluation dataset. Metadata includes attributes related to operational and environmental conditions and machine model types. For our work, we used recordings from five machine types (fan, gearbox, bearing, slide rail, valve) from the development dataset and split it to train and evaluate our generation model. Each audio recording is a single-channel file of approximately 10 seconds duration, sampled at 16 kHz.

B. Experimental setup

1) Audio generation: In order to train the diffusion model, a total of 35,146 training samples from all machines are used and 8,787 samples are used for validation of the generation quality. Table II shows the distribution of samples from each machine type.

The metadata associated with all the recorded audio clips are given as input to the T5-large [20] model to obtain descriptive captions that are saved in order to be used later for audio generation. Table I shows examples of the metadata and captions of the dataset. These captions are then encoded into 768-dimensional embedding vectors using another model Flan-T5 as shown in Fig. 1 to give as input condition during training and inference from diffusion model. During training, the parameters of Encodec and the Flan-T5 model are frozen, only the denoising U-Net is trained.

2) Anomaly Detection: We trained the anomaly detection system using 990 normal audio clips from each machine type. For evaluation, we created two datasets for all five machine types: one containing the original 50 normal and 50 anomalous clips, and another comprising 50 original normal clips and 50 generated anomalous clips. The generated anomalous clips were produced by the diffusion model using captions and conditions not encountered during its training. The anomaly

TABLE I: Generated and ground truth audio clips. Listen to them at: https://hpworkhub.github.io/MIMII-Gen.github.io/

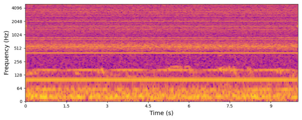
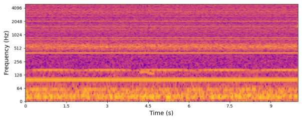
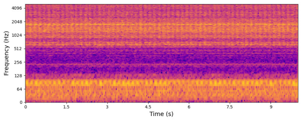
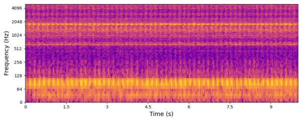
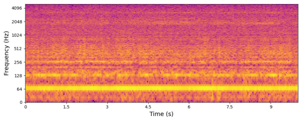
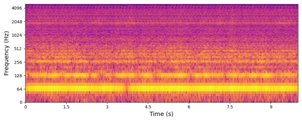
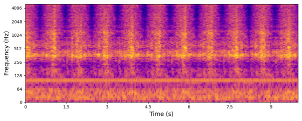
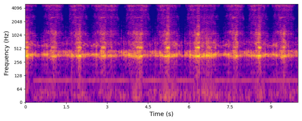
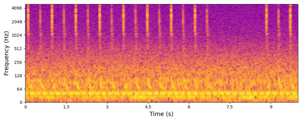
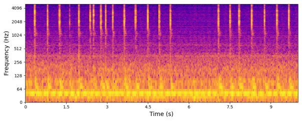
Metadata	Caption	Ground Truth	Generated
Bearing, anomaly, axis damage, velocity of 24 krpm, location "A"	A bearing operating on velocity of 24 krpm with anomaly due to axis damage at location A		
Gearbox, anomaly, damage type 2, model B, voltage of 2.3 (V), weight of 0 (g)	A gearbox model B operating on voltage of 2.3 (V) and weight of 0 (g) with anomaly due to damage type 2		
Fan, model, anomaly, over voltage	A fan model is running on over voltage with anomaly		
Slider, ball-type, anomaly, damage, velocity 1000.0 (mm/s), acceleration 0.3	A ball-type slider operating on velocity of 1000.0 and an acceleration of 0.3 with an anomaly due to damage		
Valve, anomaly, contamination, moving pattern 1, surroundings is open	A valve of moving pattern 1 in open surroundings with anomaly due to contamination		

TABLE II: Counts of audio samples for each machine type in training and validation datasets

Machine Type	Training Samples	Validation Samples
Bearing	Normal: 6381 Anomalous: 637	Normal: 1613 Anomalous: 151
Gearbox	Normal: 5863 Anomalous: 778	Normal: 1485 Anomalous: 192
Fan	Normal: 5961 Anomalous: 850	Normal: 1458 Anomalous: 202
Slide rail	Normal: 5983 Anomalous: 1036	Normal: 1536 Anomalous: 286
Valve	Normal: 7134 Anomalous: 523	Normal: 1707 Anomalous: 157

detection system is same as the autoencoder baseline used in DCASE2023 Task 2 [15].

C. Results and discussion

The generated audio samples should be of higher quality and diversity in order to validate the anomaly detection systems. So, we employ four objective metrics to evaluate our generated audio: Frechet Audio Distance (FAD) [21], [22], calculated using embeddings extracted by VGGish [23]; Kullback-Leibler divergence (KLpasst) between the outputs of PaSST [24], an audio classification model; Inception Score (ISpasst) [25], which is also based on the outputs of PaSST; and the CLAP score. Lower FAD indicates higher audio quality for the generated samples. The KL divergence assesses the semantic similarity between generated audio and reference ground truth audio. The IS measures the diversity of generated samples, while the CLAP score evaluates how closely the generated audio aligns with the provided textual description. We calculate the CLAP scores for two cases, (i) for original audio and caption (ii) for generated audio and the caption. The CLAP scores for both the cases should be almost same if the generated audio is similar to the original.

Table III presents the performance of the generation models across the evaluated metrics. The CLAP scores for the original (i) and generated (ii) cases are separated by a hyphen in the table. The results indicate that our approach outperforms the Tango baseline, which relies on AudioLDM with a pretrained VAE and vocoder. This performance gap is likely due to the vocoder's limited generalization to non-speech audio. Additionally, our approach utilizes a 16-channel input and a wide-channel U-Net, which effectively captures the latent representations within its receptive field, enhancing denoising capabilities and resulting in improved generation quality.

TABLE III: FAD and other scores for conditional audio generation.

Models	FAD ↓	KLpasst ↓	ISpasst ↑	CLAP score ↑
Tango	6.88	1.74	2.57	0.15-0.10
Our approach	5.43	1.22	3.72	0.15-0.14

Spectrogram provides a visual representation highlighting the unique patterns associated with each machine type's sounds. Table I shows that spectrograms of the generated audio samples for given captions as well as the ground truth audio clips follow similar patterns. We could also successfully generate the audio samples for combinations of different conditions which were not seen during training of the diffusion model.

The anomaly detection system is then evaluated on the anomalous data generated. Table IV shows the AUC scores obtained on both the evaluation datasets, i.e., originally recorded as well as the generated anomalous clips. AUC scores for generated data have an average difference of 4.8% and are correlated to scores of the original anomalous data for all machines.

TABLE IV: AUC scores for all machine types

Machine type	**Original data**	**Generated data**
Bearing	0.5468	0.5916
Gearbox	0.6920	0.7607
Fan	0.9496	0.9713
Slide rail	0.5588	0.6132
Valve	0.5271	0.5517

V. Conclusion

We presented a method for generating high-quality machine audio with fine-grained variations using metadata, demonstrating strong alignment with real recordings. Our approach effectively generates anomalous data across various conditions, enhancing downstream task performance. Leveraging an EnCodec-based approach over VAE and vocoder methods, we achieve superior performance with a simplified pipeline. The anomaly detection system tested on our generated audio shows only a 4.8% AUC deviation from its performance on original data, underscoring our method's practical value in industrial scenarios.

VI. Acknowledgment

We express our sincere gratitude to Prof. Keisuke Imoto of Doshisha University for his invaluable advice throughout this research.

References

[1] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *arXiv preprint arXiv:2106.04492*, 2021.

[2] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "MusicLM: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[3] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[4] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned LLM and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[5] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," *arXiv preprint arXiv:2205.13879*, 2022.

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

[7] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[8] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "DiffSound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023.

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

[10] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[11] Y. Kawaguchi, R. Tanabe, T. Endo, K. Ichige, and K. Hamada, "Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 865–869.

[12] H. Purohit, R. Tanabe, T. Endo, K. Suefusa, Y. Nikaido, and Y. Kawaguchi, "Deep autoencoding GMM-based unsupervised anomaly detection in acoustic signals and its hyper-parameter optimization," *arXiv preprint arXiv:2009.12042*, 2020.

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

[14] H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Hierarchical conditional variational autoencoder based acoustic anomaly detection," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2022, pp. 274–278.

[15] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2305.07828*, 2023.

[16] H. Zhang, Q. Zhu, J. Guan, H. Liu, F. Xiao, J. Tian, X. Mei, X. Liu, and W. Wang, "First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1271–1275.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[18] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[19] J. Michaels, J. B. Li, L. Yao, L. Yu, Z. Wood-Doughty, and F. Metze, "Audio-Journey: Open domain latent diffusion based text-to-audio generation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6960–6964.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[21] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[22] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting Frechet audio distance for generative music evaluation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1331–1335.

[23] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A large-scale audio-visual dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[24] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.