Metadata-Free Text-to-Audio Normal Synthesis and Latent Gradient Perturbation for Unsupervised Anomalous Sound Detection

Technical Report

JeongSik Kim[*] JongWoo Sung[*] HyoenJun Bae[*] SukHwan Lee[*]

[*]Dong-A University Computer Engineering Dept,
Busan, Republic of Korea
{2343783, planthehuman, gull}@donga.ac.kr
skylee@dau.ac.kr

ABSTRACT

This technical report shows a fully metadata-free framework for unsupervised anomalous sound detection that synthesizes both normal and anomalous training examples. First, we generate diverse normal audio clips by training and adapting a pretrained Tango text-to-audio model: we apply LoRA and fine-tune Text Encoder and VAE in Tango, and full tuning UNet using three automated prompt strategies (fixed templates, spectrogram-statistic descriptions, and CLAP-filtered captions). Next, we create realistic anomalous spectrograms by perturbing encoded normal representations with gradient ascent and enforcing their magnitude via truncated projection. These synthetic normal and anomalous samples are then used to train a downstream spectrogram-based detector, yielding marked improvements in detection accuracy. In future work, we will close the gap between synthetic and real distributions and extend our approach to direct anomalous audio generation.

Index Terms— Unsupervised, Anomaly Detection, Audio Synthesis, Gradient Ascent

1. INTRODUCTION

Unsupervised anomalous sound detection is critical for predictive maintenance in industrial settings, where genuine fault recordings are scarce. Conventional reconstruction-based autoencoders often fail to distinguish subtle anomalies, while feature-based one-class classifiers may lack sensitivity to near-normal faults.

To address these limitations, we introduce a dual-synthesis framework that generates diverse synthetic anomalies at both global and local scales. First, an encoder–decoder is trained to reconstruct normal spectrograms. Next, we apply gradient-guided perturbations in latent space followed by controlled projection to create "synthetic" anomalies that sit near the reconstruction boundary. Simultaneously, we produce local spectral masks using thresholding and Perlin-noise overlays to simulate fine-grained faults.

Training a lightweight convolutional discriminator on these enriched normal and synthetic samples yields more reliable anomaly scores, improving detection accuracy and interpretability without reliance on fault metadata.

2. DATASET

In the development dataset [1] and additional dataset [5] of DCASE 2025 Task2 [1], First and last few seconds of audio data does not impact nor improve result of the performance. This lead us to crop out extra audio in "discriminator stage". However our text-to-audio flow does not crop out audio.

Every system is using STFT band size to 128, 50% hop size and 16k sampling rate as an input data.

3. MODEL ARCHITECTURE

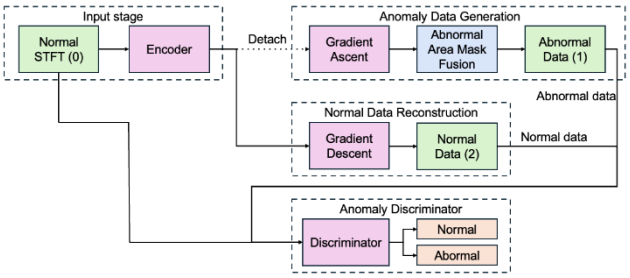

Figure 1: Discriminator flow

Discriminator used in both first and second system is same. Encoder is treated as backbone of the system and we've mixed and tested different type of convolution based models like resnet and efficientnet. In the reconstruction stage of the model we've used convolutional upscaling as gradient ascent model and for pre-training encoder for audio data we've used simple 2D transpose for gradient descent which reconstructs normal audio.

Our discriminator has four different stages for optimally generating anomaly data and training discriminator.

Normal Data Reconstruction serves two prepose of generating STFT of normal audio and pre-training encoder which also frozen in Anomaly Discriminator stage.

Anomaly Data Generation stage first truncated projects latent vector from encoded latent vector just near out side of normal latent vector's boundary and makes it robust in first show context. From this latent vector our gradient ascent model reconstructs "Anomaly Patch" or "Anomaly Region" which we overly on top

of the normal STFT data. This makes generating anomaly data more controllable.

Both Normal Data Reconstruction and Anomaly Data Generation stage were trained on mix of SSIM Loss [6] and L1 Loss.

Finally Anomaly Discriminator stage uses same frozen encoder latent vector from both normal and generated abnormal data. Then encoded vector goes into fully connected 4 layer model for classifying end result. This way of training model might not be optimal in training stage but on the inference stage we only use encoder and Discriminator's four layer FC which makes inference of the actual model near real-time.

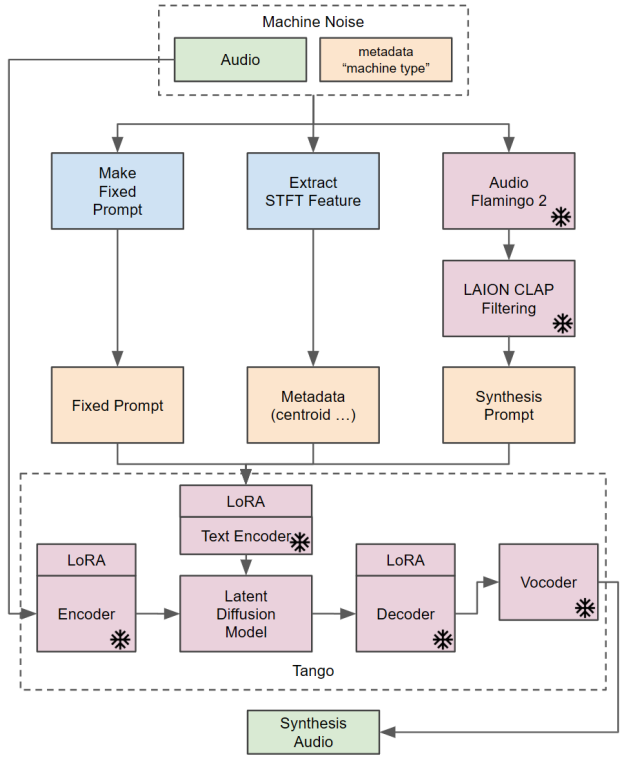

Figure 2: Text to Audio system flow

Our framework is based on Tango [9], a latent diffusion-based text-to-audio generation model. The architecture consists of four components: a frozen text encoder based on FLAN-T5 [11], a variational autoencoder (VAE) [13], a vocoder, and a latent diffusion model (LDM) [12], which is the only trainable component. All input audio clips are 10 seconds long and sampled at 16 kHz. These are transformed into log-mel spectrograms using STFT with the following parameters: n_fft = 1024, hop_length = 128, mel_bins = 64, and a Hann window function.

While this structure provides a stable foundation for general-purpose audio generation, it presents notable limitations in domain-specific applications such as machine sound synthesis. The frozen text encoder struggles to interpret specialized terminology, and the VAE [13] often fails to retain fine-grained, domain-relevant acoustic features in its latent representation.

To address these issues, we apply Low-Rank Adaptation (LoRA) [10] to both the text encoder and the VAE [13], which were originally frozen, in order to give them the capacity to adapt to machine-specific audio characteristics. Each LoRA [10] module is configured with a rank of 32 and a scaling factor of 8. The LDM [12], which operates within the diffusion process, is instead fully fine-tuned.

To enhance controllability and diversity in audio generation, we experimented with three types of text conditions. These conditions were designed under the assumption that metadata is unavailable, which is a common scenario in real-world machine sound datasets. Our goal was to explore how text prompts can be constructed from raw or minimally preprocessed audio.

The first condition uses a fixed prompt template: "The normal condition machine sound of {machine_type}." This approach provides a uniform semantic context across all samples but suffers from rigidity, making it incapable of capturing instance-level variation in the data.

The second condition involves constructing metadata-style prompts based on low-level acoustic features extracted from each spectrogram. Specifically, we compute six attributes: spectral flux, centroid, flatness, kurtosis, RMS energy, and zero-crossing rate. These values are normalized and formatted into structured, human-readable text. This approach enables the encoding of both continuous and categorical properties into the prompt, supporting the grouping of acoustically similar samples and allowing controlled perturbations for data augmentation.

The third condition leverages AudioFlamingo2 [7], a large pretrained audio-language model, to generate four natural language descriptions per audio clip. Each description is generated using a structured analytical prompt with max_new_tokens = 256. We then calculate CLAP [8] based similarity scores between the audio and each generated description. Using the 630k-best.pt and 630k-audioset-best.pt CLAP models [8], we compute the average similarity score for each description and select the one with the highest average score, following a CLAP-based selection strategy inspired by Tango 2 [14]. This method provides interpretable, semantically rich prompts, and if sufficient audio-text alignment is achieved, it has the potential to support natural language–driven conditional synthesis.

4. EXPEREMENT AND RESULTS DISCUSSION

We have trained our system on 4x Nvidia RTX3090 and 5x V100 at various batch size and learning rate, epoch. Results are all average result of 5 different seeds.

Table 2: System results

Machine	Method	Baseline	***Discriminator***	***Full system***
ToyCar	AUC (Source)	73.17 %	54.54 %	**77.12 %**
	AUC (Target)	53.52 %	58.83 %	**78.03 %**
	pAUC	49.70 %	54.11 %	**66.87 %**
ToyTrain	AUC (Source)	61.76 %	65.65 %	**78.25 %**
	AUC (Target)	56.46 %	61.05 %	**77.41 %**
	pAUC	50.19 %	50.19 %	**68.42 %**
bearing	AUC (Source)	66.53 %	74.47 %	**79.00 %**
	AUC (Target)	59.03 %	64.04 %	**77.99 %**
	pAUC	61.86 %	61.21 %	**69.53 %**
fan	AUC (Source)	77.99 %	72.07 %	**78.14 %**
	AUC (Target)	38.75 %	49.35 %	**76.72 %**
	pAUC	50.82 %	50.47 %	**65.33 %**

gearbox	AUC (Source)	73.79 %	70.14 %	**78.70 %**
	AUC (Target)	51.61 %	52.35 %	**77.89 %**
	pAUC	55.07 %	64.50 %	**69.64 %**
slider	AUC (Source)	73.79 %	75.15 %	**77.94 %**
	AUC (Target)	50.27 %	51.11 %	**76.48 %**
	pAUC	53.61 %	51.34 %	**64.10 %**
valve	AUC (Source)	63.53 %	74.90 %	**79.00 %**
	AUC (Target)	67.18 %	72.72 %	**78.61 %**
	pAUC	57.35 %	56.42 %	**67.27 %**

5. CONCLUSION

This tech report presents a framework that synthesizes diverse normal audio from text without metadata using Tango with LoRA and generates anomalous audio via latent gradient perturbation. The synthetic samples were used to train a spectrogram-based anomaly detector, effectively addressing the data scarcity problem in unsupervised settings and leading to improved detection accuracy.

In future work, we will close the gap between synthetic and real distributions and extend our approach to direct anomalous audio generation from text prompts.

6. REFERENCES

[1] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," In arXiv e-prints: 2506.10097, 2025.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Barcelona, Spain, November 2021, pp. 1–5.

[3] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events Workshop (DCASE2022), Nancy, France, November 2022.

[4] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," Proceedings of the 31st European Signal Processing Conference (EUSIPCO), pp. 191–195, 2023.

[5] Nishida, T., Harada, N., Niizumi, D., Albertini, D., Sannino, R., Pradolini, S., Augusti, F., Imoto, K., Dohi, K., Purohit, H., Endo, T., & Kawaguchi, Y. (2025). DCASE 2025 Challenge Task 2 Additional Training Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15392814

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," In IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[7] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, "Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities," arXiv preprint arXiv:2503.03983, Mar. 2025.

[8] Y. Wu*, K. Chen*, T. Zhang*, Y. Hui*, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2023.

[9] 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, Apr. 2023.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, Jun. 2021.

[11] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling Instruction-Finetuned Language Models," arXiv preprint arXiv:2210.11416, Oct. 2022.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv preprint arXiv:2112.10752, Dec. 2021.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," In arXiv e-prints: 1312.6114, December 2013. (last revised December 2022)

[14] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization. arXiv preprint arXiv:2404.09956, 2024. Accepted at ACM MM 2024.