DID YOU HEAR THAT? INTRODUCING AADG: A FRAMEWORK FOR GENERATING BENCHMARK DATA IN AUDIO ANOMALY DETECTION

Ksheeraja Raghavan[*]
Carnegie Mellon University

Samiran Gode[*]
University of Technology Nuremberg

Ankit Shah[*]
Carnegie Mellon University

Surabhi Raghavan
University of Pittsburgh

Wolfram Burgard
University of Technology Nuremberg

Bhiksha Raj & Rita Singh
Carnegie Mellon University

ABSTRACT

We introduce a novel, general purpose audio generation framework specifically designed for Audio Anomaly Detection (AAD) and Localization. Unlike existing datasets that predominantly focus on industrial and machine-related sounds, our framework focuses a broader range of environments, particularly useful in real-world scenarios where only audio data are available, such as in video-derived or telephonic audio. To generate such data, we propose a new method, Audio Anomaly Data Generation (AADG), inspired by the LLM-Modulo Kambhampati et al. (2024) framework, which leverages Large Language Models (LLM) as world models to simulate such real-world scenarios. This tool is modular, allowing for a plug-and-play approach. It works by first using LLMs to predict plausible real-world scenarios. An LLM further extracts the constituent sounds, the order and the way in which these should be merged to create coherent wholes. We include a rigorous verification of each output stage, ensuring the reliability of the generated data. The data produced using the framework serves as a benchmark for anomaly detection applications, potentially enhancing the performance of models trained on audio data, particularly in handling out-of-distribution cases. Our contributions thus fill a critical void in audio anomaly detection resources and provide a scalable tool for generating diverse, realistic audio data.

1 INTRODUCTION

Detecting anomalies is crucial for various reasons, such as preventing harm Saligrama et al. (2010), early detection of unexpected events Patcha & Park (2007) Akoglu et al. (2015), and ensuring safety in critical systems Chandola et al. (2009) as well as data integrity. Audio is often the only usable modality in scenarios like extortion cases via phone calls Bidgoli & Grossklags (2017), where actionable information can help solve crimes. Similarly, protest videos with unclear visuals but clear audio and audio from surveillance/CCTV cameras capturing events outside the camera's field of view highlight the need for audio-specific anomaly detection. Detecting out-of-distribution cases in such contexts is essential. In this paper, we take inspiration from the computer vision community to define anomalies Ramachandra et al. (2020) Saligrama et al. (2010), this paper considers only single-scene anomalies.

Definition 1 Audio Anomalies are audio events that stand out within a scene due to their unusual nature relative to the surrounding sounds. This anomaly can arise from the event's position within the audio timeline, its incongruity with the expected auditory context, or the inherent rarity of the sound itself.

There are several benchmark datasets for video-based anomaly detection, such as Street Scene Ramachandra & Jones (2020), CUHK Avenue Lu et al. (2013), ShanghaiTech Liu et al. (2018), and UCSD Ped1 and Ped2 Mahadevan et al. (2010) Li et al. (2013). However, there is a lack of dedicated datasets for audio-only anomaly detection, despite their importance. Existing audio anomaly

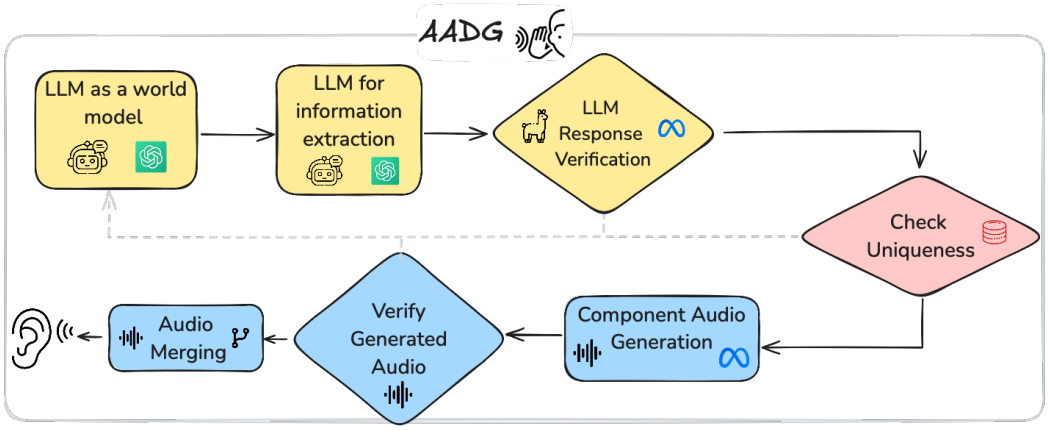

Figure 1: Audio Anomaly Data Generation (AADG), a framework that synthetically generates real life Audio Data with Anomalies by leveraging LLMs as a world model

datasets are primarily focused on industrial and machine data Dohi et al. (2022b) Dohi et al. (2022a) Harada et al. (2021). Collecting anomalous data is inherently challenging, as it is out-of-distribution by nature and occurs significantly less frequently than general data.

Current datasets used to train Audio Language Models (ALMs) and text-to-audio models lack diversity and do not include complex scenarios, resulting in poor performance for these models when handling complex or anomalous audio Ghosh et al. (2024) Agre & Chapman (1987). State-of-the-art (SOTA) text-to-audio models Vyas et al. (2023) Evans et al. (2024) also struggle with descriptive prompts, unlike text-to-image Podell et al. (2023) Yu et al. (2022) and video models Brooks et al. (2024), which are trained on broader and more diverse datasets. Collecting real-life audio samples with anomalies is challenging and time consuming. Given the importance of audio anomaly detection, training and benchmark data are crucial. To address this, synthetic data generation can augment existing datasets, as real-world training data often lacks such scenarios, necessitating alternative approaches. This paper addresses the lack of diverse audio anomaly datasets utilizing Large Language Models (LLMs) Guan et al. (2023), such as GPT-4 Achiam et al. (2023) and LLama Touvron et al. (2023), which are trained on vast datasets and capable of generating plausible anomalous and out-of-distribution scenarios. These scenarios can be converted into audio using SOTA text-to-audio models Evans et al. (2024) Vyas et al. (2023), which, while effective for simple cases, struggle with more complex scenarios. Recent advances in LLMs and text-to-audio models provide an opportunity to synthesize realistic and diverse audio data for training and benchmarking anomaly detection models. Building on Kambhampati et al. (2024), which shows that LLMs excel at generating and verifying plans, we leverage their ability to create plausible anomalous scenarios. These scenarios are processed to generate component sounds and their order, which are merged using predefined methods. The final audio outperforms SOTA text-to-audio models in handling complex and out-of-distribution (OOD) cases. Our modular plug-and-play framework, independent of specific language or text-to-audio models, generates synthetic data (with text descriptions, component audios, and timestamps) for training and benchmarking audio anomaly detection models. This creates the first general-purpose audio anomaly dataset for improved audio perception and localization. Our key contributions are as follows.

Novel Framework for Audio Anomaly Data Generation: AADG (Audio Anomaly Data Generation) uses LLMs as world models to synthetically generate realistic audio data with anomalies, addressing the scarcity of diverse datasets beyond industrial settings.

Modular and Extensible Approach: The framework's modular design enables a plug-and-play approach, independent of the specific language or text-to-audio model, and can adapt to future advancements in both LLMs and audio generation.

Creation of the First General-Purpose Audio Anomaly Dataset: The synthetic data will form the first general-purpose audio anomaly dataset—with text descriptions, component audios, and timestamps—crucial for training and benchmarking anomaly detection models.

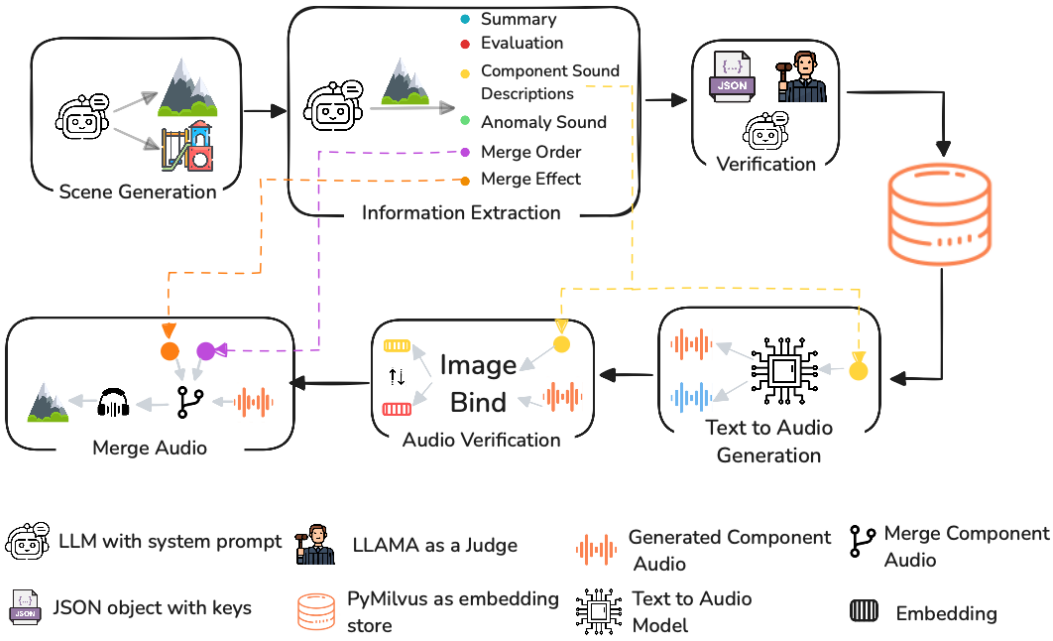

Figure 2: Illustration of the pipeline. The process begins with scene generation, followed by information extraction using a Large Language Model (LLM). Individual audio components are synthesized from text descriptions and meticulously verified for accuracy and merged according to LLM instructions, culminating in a dataset of realistic anomalous audio.

2 METHOD

We use Large Language Models (LLMs) as a world model to generate a scenario. Then another LLM call extracts the summary of the scene, the anomaly, the sound components, and their merging sequence. Next, we employ a text-to-audio model to create each audio component and merge them according to the LLM's instructions. At each stage, we verify the output to ensure that the final audio makes sense. We go through the entire pipeline in detail which is also shown in Fig2.

2.1 SCENARIO GENERATION

LLMs, trained on internet-scale data, excel at generating realistic scenarios and candidate plans Kambhampati et al. (2024). We prompt LLMs to create plausible scenarios with sufficient detail to generate scenes with component sounds and anomalies. By conditioning the prompt, we can adjust the number of anomalies and tailor the scene, enabling adaptation to different audio types. Using GPT-3.5 and GPT-4o in our experiments, we observe that larger models produce more descriptive and creative outputs, aligning scenes with realistic audio characteristics. The prompt ensures scenarios are distinct enough to generate identifiable sounds, typically limited to one anomaly, which suffices for benchmarking current ALMs. Temperature settings significantly impact the results. Higher temperatures enhance creativity but occasionally cause models like GPT-4o to generate nonsensical outputs, likely due to prompt complexity and high temperature amplifying anomalies.

2.2 INFORMATION EXTRACTION

After initial scenario generation, a second call to the LLM extracts and formats useful information. This step summarizes the scenario, identifies the anomaly, and provides instructions for audio generation, including component sounds, their order, and merge types. This structured output, facilitated by the pydantic library Colvin et al. (2024), ensures data consistency as dictionaries for subsequent processing. The LLM's creative capabilities allow it to understand the context, identify suitable sounds for the scene, and determine how to merge them for coherence. It also specifies the anomaly, explains its role, and integrates it with the audio to create a realistic and informative dataset.

2.3 VERIFICATION - LANGUAGE MODEL

Large Language Models though creative suffer from issues such as hallucination Huang et al. (2023) Li et al. (2024) Liu et al. (2024a), Valmeekam et al. (2022) Kambhampati (2024) lack of reasoning, redundant outputs Tirumala et al. (2024) Chiang & Lee (2024) etc. Inspired by LLM Modulo Kambhampati et al. (2024) we use LLM's impressive creative capabilities but also verify its outputs to check for issues which might creep in. We check for logical flaws, alignment with the required output and coherency of the output

2.3.1 LOGICAL VERIFICATION OF OUTPUT

Since the extraction creates outputs that we use in downstream tasks we have to check if they logically make sense. We find that language model fails in number of different ways and thus we try to verify each part. One of the ways it fails is by creating merge types that do not exist in our designated methods, hence we check if the generated merge types lie within what we use. Another way the language model fails and we have to check for is with the number of component sounds, the order and the merge types not being equal. We also find cases where the scenario doesn't make sense and contains nonsensical text. There are also cases where the component audios contain audios that do not make sense with words such as silence, confusion, nervousness etc. These are sounds that need to be checked.

2.3.2 LLM AS A JUDGE

We utilize Llama Touvron et al. (2023) as an evaluative tool to verify the responses generated by GPT Achiam et al. (2023), by setting up Llama Touvron et al. (2023) as an independent impartial judge. The evaluation uses a Single Answer Grading Framework Zheng et al. (2024) to directly assign a score to the GPT responses. This setup introduces a layer of quality control, ensuring that the responses generated by GPT align with the prompts that AudioCraft Meta AI (2024) can use for component audio generation, while eliminating the need of human intervention. In this study, we leverage Llama as an impartial evaluation tool to assess the quality of the responses generated by GPT-4 . By employing a Single Answer Grading Framework , Llama directly assigns scores to GPT-4's outputs, introducing a layer of quality control. This approach ensures that the generated responses align with prompts suitable for component audio generation in AudioCraft , while eliminating the need for human intervention.

2.4 COMPONENT AUDIO GENERATION

Once we have access to the audio components from the extraction, we pass them to a text-to-audio model which creates the audio components. The advantage of our method is that the text-to-audio models can be replaced as we find better ones, or we could also use multiple text-to-audio models. In practice, we use Audiogen Kreuk et al. (2022). Audiogen is a textually guided model part of AudioCraft Meta AI (2024). We find it to be the best for our case because it is open-source, although there are better models such as Audiobox Vyas et al. (2023) but are not available for use. The text-to-audio model can create good audios as long as the prompt isn't complex, in our case, the extraction helps keep the prompts small. However, we could also make the component sound descriptions more informative by conditioning the prompt of the LLM call for extraction. Based on the model that we are currently utilizing, we have determined that utilizing less detailed texts yields better results in practice.

2.5 VERIFICATION - AUDIO GENERATION

Like LLMs, the text-to-audio model is imperfect and often misaligns with its text prompt. The training dataset may not cover all possible sounds, resulting in OOD cases. For instance, it performs better on "cat meowing" than "lion roaring," likely due to more examples of the former in its trainidata. Prompts involving timing specifications (e.g., "periodic announcement prompts") confuse the model. It also struggles to accurately render conversations unless they are in the background. To ensure that the final audio semantically aligns with the text, we propose using a multimodal model trained to align embeddings representing the same scenario across different modalities. We utilize ImageBind Girdhar et al. (2023), though other models such as Audio CLIP could also be considered.

Since ImageBind has been trained contrastively to align the embeddings of the same label across different modalities, the output embeddings of ImageBind for the generated audio and the text prompt should be close according to a chosen distance metric. We assess the alignment between the text embeddings($\mathbf{E}_{\text{text}}$) and audio embeddings($\mathbf{E}_{\text{audio}}$) using cosine similarity. Their cosine similarity is computed as:

$$\cos\_\text{sim}(\mathbf{E}_{\text{text}}, \mathbf{E}_{\text{audio}}) = \frac{\mathbf{E}_{\text{text}} \cdot \mathbf{E}_{\text{audio}}}{\|\mathbf{E}_{\text{text}}\| \, \|\mathbf{E}_{\text{audio}}\|}.$$

This metric quantifies the alignment between the embeddings, with higher values indicating greater similarity. For each generated audio, if the cosine similarity is above a predefined threshold, we accept the audio as semantically aligned. However, in practice, we found that ImageBind did not perform well with the generated audio even when the audio sounded accurate. Despite this, the similarity score was significantly lower (by an order of magnitude) for semantically dissimilar audios compared to semantically similar ones. To enhance the verification process, we apply a sigmoid regularizer to make the differences more pronounced:

$$\text{RegSim}(\mathbf{E}_{\text{text}}, \mathbf{E}_{\text{audio}}) = \sigma\left(\alpha \cdot \frac{\mathbf{E}_{\text{text}} \cdot \mathbf{E}_{\text{audio}}}{\|\mathbf{E}_{\text{text}}\| \, \|\mathbf{E}_{\text{audio}}\|} - \beta\right).$$

Where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function and α and β are tunable parameters that control the scaling and shift. We only accept audios whose regularized similarity is above a certain threshold, improving the verification.

2.6 Audio Merging

After verifying the component audios, we merge them based on the language model's instructions, using methods like cross-fade, overlay, fade-in, or fade-out. We merge audios in the order specified by the language model, ensuring each is normalized to maintain consistent audio levels. These merge types can be expanded, although we have found these to be sufficient in practice. As we merge audios sequentially, each new audio is appended to the last. For example, with a fade-in, the new audio begins fading in at the previous audio's end. For a fade-out, it is added to the end, fading out as it concludes. In a cross-fade, the previous audio cross-fades into the new one. With an overlay, the new audio's alignment depends on its length compared to the previous merged audio. We store the timestamps of each audio in the final merged audio which can be used for evaluation, training anomaly detection models, or event detection. It should be noted that the merging methods can be altered by adding new methods or utilizing a new learned model.

2.7 Final Data

The final data contains the audio components, the merged audio itself and metadata which contains the scenario, the summary of the scenario, the anomaly, a description of why it is anomalous, the text description of the audio components, the order in which the audios have been merged, the method with which they have been merged and the time stamps of the component audios in the final audio.

3 Evaluation

To demonstrate the usefulness of our model, we illustrate how existing models using audio could be improved by accounting for anomalous scenarios and using synthetic data during training.

3.1 Comparison against state-of-the-art text-to-tudio generation models

We demonstrate that our framework produces better data than SOTA text-to-audio models, particularly with complex or anomalous prompts. We compare against Stable Open Audio Evans et al. (2024), the current SOTA in table 1a. The vocabulary set used by models like Audiobox Vyas et al. (2023) an Stable Open Audio is limited because they are trained on datasets like AudioSet, which only cover a specific range of sounds and scenarios. These models generate audio based on a fixed set of well-represented categories—common environmental noises, simple musical notes, or speech. However, when prompts involve complex or uncommon scenarios, they fail to produce accurate outputs due to a limited sound vocabulary and understanding beyond their training data. Their restricted

Model	PREF
Stable Audio Open	0.12
Ours AADG	**0.88**

(a) Adherence to text prompts in text-to-audio models with complex anomaly-inducing prompts.

Model	MOS
GAMA against simple audios	4.00
GAMA against complex audios	3.21

(b) Mean Opinion Score (MOS) to evaluate SOTA ALM audio understanding. Higher is better.

Table 1: Comparison of text-to-audio models on different evaluation metrics.

Prompt Complexity	FAD
Lower Complexity	5.5015
Higher Complexity	**6.775**

Table 2: Comparing the audio separation capabilities for different levels of prompt complexity. The lower the FAD score the better the match

sound vocabulary produces repetitive or irrelevant outputs when prompts require a broader, specialized range, reducing effectiveness in tasks like anomaly detection that rely on nuanced sounds. This highlights the need for a framework that can simulate a wide variety of audio events.

3.2 Benchmarking Audio Language Models

Current ALMs are trained on anomaly-free datasets, such as clean speech with minimal background noise, which raises concerns about their robustness to real-world audio complexities. To assess the true capabilities of SOTA ALMs, we tested using the current SOTA Ghosh et al. (2024) asking it to predict complex audios. We randomly sample audio and ask participants to choose how accurate the description is for the audio. We ask them to score the model from 1 to 5, 5 being extremely accurate and 1 being inaccurate. We report the MOS of the participants. The results are shown in 1b. We find that the Audio Language Model (ALM) can understand the easier audios but does not fully comprehend the complex audios. This is likely because the model wasn't trained on such data. Models like GAMA, Ghosh et al. (2024) are impressive, but will become much better once trained on datasets augmented with complex data.

3.3 Comparing Against Audio Separation Models

Audio separation models Liu et al. (2022) Liu et al. (2023b) extract a specific component audio, guided by a descriptive prompt, from a test audio containing multiple simultaneous sounds. Audio separator models struggle with complex or anomalous audio because they haven't encountered such components during training, limiting their ability to distinguish individual parts. We show that complex audios and their corresponding prompts break such models. Through our dataset, we have access to the component audio description along with the prompt and the component audios. Audio separator models as described in Liu et al. (2022)Liu et al. (2023b) should, ideally, be able to separate our audio into it's components. The separated audios, extracted using our text descriptions, should match the original audio generated from the same prompt. However, we find that separation performance deteriorates when the audio contains anomalies, is complex, or includes OOD sounds, compared to simpler audio scenarios. To test this, we generate audio samples which are generated using simple and complex prompts, we find that the ones generated with complex prompts are harder to separate 2. We use Frechet Audio Distance (FAD) similar to Vyas et al. (2023)Kreuk et al. (2022) to measure the closeness of separated and original audios, we find that the FAD increases for complex audios.

4 Conclusion

The proposed framework, AADG, improves data generation of anomalous audio with greater versatility and scalability compared to current methods. Unlike traditional datasets that focus on industrial or machine sounds, it leverages LLMs to simulate a broader range of real-world scenarios, making it particularly valuable for audio-only applications such as surveillance and telephonic recordings. The modular design enables integration of various LLMs and text-to-audio models, allowing the generation of complex, anomalous scenarios that are hard to capture in real-world data. While current text-to-audio models still face challenges with generating realistic audio for complex prompts and

anomalies, the framework introduces multi-stage verification processes to minimize logical flaws, misalignment, and inconsistent outputs. Additionally, by using multimodal models like ImageBind for verification, the framework improves the reliability of the generated data, although this process still has limitations in handling certain out-of-distribution cases. Overall, this approach fills a critical gap in the field by providing a scalable tool for creating diverse and realistic audio datasets, which are essential for advancing audio anomaly detection technologies as well as complex audio generation

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Philip E Agre and David Chapman. Pengi: An implementation of a theory of activity. In *Proceedings of the sixth National conference on Artificial intelligence-Volume 1*, pp. 268–272, 1987.

Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, 29:626–688, 2015.

Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3526–3535, 2023.

Morvareed Bidgoli and Jens Grossklags. "hello. this is the irs calling.": A case study on scams, extortion, impersonation, and phone spoofing. In *2017 APWG Symposium on Electronic Crime Research (eCrime)*, pp. 57–69. IEEE, 2017.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Cheng-Han Chiang and Hung-yi Lee. Over-reasoning and redundant calculation of large language models. *arXiv preprint arXiv:2401.11467*, 2024.

Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, and Alex Hall. Pydantic, June 2024. URL `https://docs.pydantic.dev/latest/`.

Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, and Yohei Kawaguchi. Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. In *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, pp. 1–5, Nancy, France, November 2022a.

Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido, and Yohei Kawaguchi. MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task. In *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, pp. 1–5, Nancy, France, November 2022b.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. Leveraging pretrained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094, 2023.

Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Masahiro Yasuda, and Shoichiro Saito. ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pp. 1–5, Barcelona, Spain, November 2021. ISBN 978-84-09-36072-7.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Subbarao Kambhampati. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18, 2024.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*, 2024.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Johnny Li, Saksham Consul, Eda Zhou, James Wong, Naila Farooqui, Yuxin Ye, Nithyashree Manohar, Zhuxiaona Wei, Tian Wu, Ben Echols, et al. Banishing llm hallucinations requires rethinking generalization. *arXiv preprint arXiv:2406.17642*, 2024.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, and Li Zhang. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*, 2024a.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023a.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024b.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Separate what you describe: Language-queried audio source separation. In *Proc. Interspeech*, pp. 1801–1805, 2022.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Separate anything you describe. *arXiv preprint arXiv:2308.05037*, 2023b.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pp. 2720–2727, 2013.

Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, 2010. doi: 10.1109/CVPR.2010.5539872.

Meta AI. Audiocraft by meta ai. `https://ai.meta.com/resources/models-and-libraries/audiocraft/`, 2024. Accessed: 2024-08-14.

Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. Nemotron-4 340b technical report, 2024. URL `https://arxiv.org/abs/2406.11704`.

Animesh Patcha and Jung-Min Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer networks*, 51(12):3448–3470, 2007.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. Datadreamer: A tool for synthetic data generation and reproducible llm workflows. *arXiv preprint arXiv:2402.10379*, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL `https://arxiv.org/abs/2307.01952`.

Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2569–2578, 2020.

Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2293–2312, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Venkatesh Saligrama, Janusz Konrad, and Pierre-Marc Jodoin. Video anomaly identification. *IEEE Signal Processing Magazine*, 27(5):18–33, 2010.

Ashish Singh, Michael J Jones, and Erik G Learned-Miller. Eval: Explainable video anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18717–18726, 2023.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*, 2022.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix

A.1 Related work

A.1.1 Anomaly detection in video

Detecting anomalies in videos is quite a common problem Ramachandra et al. (2020) Lu et al. (2013) Liu et al. (2018) Mahadevan et al. (2010) Li et al. (2013) with multiple datasets for benchmarking. Ramachandra et al. (2020) talks about single-scene anomaly detection, and the need for benchmarking data for development of new algorithms. There are multiple such datasets for video. Mahadevan et al. (2010), Li et al. (2013) are the most widely usedRamachandra et al. (2020) and contain videos from different static cameras. Lu et al. (2013) Ramachandra & Jones (2020) and Liu et al. (2018) are some other common ones, all focusing on single-scene anomalies. Our work aims to introduce similar progress in audios for anomaly detection. There have been multiple works in detecting audio anomaly data and can roughly be divided into two parts representation learning and detection methods Singh et al. (2023).

A.1.2 TEXT TO AUDIO GENERATION

The field of text-to-audio generation has progressed significantly due to innovations in diffusion-based techniques like AUDIT Wang et al. (2023), AudioGen Kreuk et al. (2022), and AudioLDM Liu et al. (2023a) Liu et al. (2024b). Additionally, there have been improvements with auto-regressive models exemplified by AudioGen Kreuk et al. (2022). AudioGen Kreuk et al. (2022) learns representations from the raw waveform and utilizes a transformer model conditioned on text to produce audio outputs. Moreover, advancements in flow matching have further enhanced text-to-audio generation capabilities, as demonstrated in Vyas et al. (2023). Recently, Evans et al. (2024) introduced a diffusion transformer.

A.1.3 LLMS FOR SYNTHETIC DATA GENERATION

Recent works have shown significant progress in synthetic data generation in the image space in Liang et al. (2024) Ramesh et al. (2021) Sun et al. (2024) Bae et al. (2023) etc. However, similar progress has not been seen in the audio space. Zero-shot text-to-image generation approaches have expanded the scope of synthetic data applications by enabling the generation of novel image data from unseen textual prompts, highlighting the model's ability to generalize from limited examples Ramesh et al. (2021). Digiface-1m Bae et al. (2023) dataset exemplifies the practical applications of these technologies, providing a robust framework for testing and improving face recognition algorithms through access to one million digital face images. Ye et al. (2022) outlines a method to leverage LLMs to create synthetic datasets produced entirely using pre-trained language models (PLMs) without human interference while emphasizing the efficiency and flexibility of using synthetic datasets to train task-specific models. Yu et al. (2024) explores generation of training data that not only focuses on diversity, but also addresses inherent biases within the data generated by LLMs. It highlights the critical role of using diversely attributed prompts that enhance quality and utility of synthetic datasets improving model performance across NLP tasks. Patel et al. (2024) presents a tool designed to streamline synthetic data generation using LLMs providing a platform to generate, train and share datasets and models.

Training on synthetic data can improve the model performance Nvidia et al. (2024) . In Dubey et al. (2024) it has been used to generate training data for text-qualilty classifiers.

A.1.4 LLMS FOR PLANNING

Chain-of-thought (CoT) prompting has emerged as a powerful technique to enhance the reasoning capabilities of LLMs by generating intermediate reasoning steps, thereby improving performance on complex tasks such as arithmetic and commonsense reasoning. Additionally, the LLM Modulo framework has shown promise in iterative planning and reasoning tasks by establishing a robust interaction between generative models and verifiers, leading to significant improvements in domains like travel planning. These methodologies and frameworks underscore the potential of LLMs to revolutionize automated planning and other complex domains, making them a focal point for future research and development.

A.1.5 LLM VERIFICATION

Zheng et al. (2024) investigates the use of GPT-4 as an evaluator of other LLMs, demonstrating the model's capability to assess responses scalably reducing human involvement and enabling faster iterations.

A.2 LIMITATIONS

Anomaly detection in audio presents a major challenge due to the limited vocabulary set used by current SOTA text-to-audio models such as Audiobox and Stable Open Audio. These models are trained on datasets like AudioSet, which only encompass a specific range of sounds and scenarios. Consequently, they struggle to produce accurate audio when presented with complex or uncommon audio scenarios that fall outside their training data. This restricted sound vocabulary results in repetitive or irrelevant outputs for anomalous audio prompts. Our approach, AADG, aims to address this limitation by generating out-of-vocabulary sounds with complex real-life descriptions. However, the generated audio may sometimes sound unnatural, which can be computationally intensive to detect.

The inherent limitation of the current audio generation models start to degrade in audio quality for very long duration audio generation, which limits the audio duration the AADG framework will be able to generate for complex anomaly description in a given acoustic scene.