Challenge on Sound Scene Synthesis: Evaluating Text-to-Audio Generation

Junwon Lee, Modan Tailleur, Laurie M Heller, Keunwoo Choi, Mathieu Lagrange, Brian McFee, Keisuke Imoto, Yuki Okamoto

Challenge on Sound Scene Synthesis: Evaluating Text-to-Audio Generation

Junwon Lee[*1], **Modan Tailleur**[*2], **Laurie M. Heller**[*3], **Keunwoo Choi**[*4], **Mathieu Lagrange**[*2], **Brian McFee**[5], **Keisuke Imoto**[6], **Yuki Okamoto**[7]
[1]KAIST, [2]Nantes Université, [3]CMU, [4]Gaudio Lab, [5]NYU, [6]Doshisha Univ., [7]UTokyo

Abstract

Despite significant advancements in neural text-to-audio generation, challenges persist in controllability and evaluation. This paper addresses these issues through the *Sound Scene Synthesis* challenge held as part of the Detection and Classification of Acoustic Scenes and Events 2024. We present an evaluation protocol combining objective metric, namely Fréchet Audio Distance, with perceptual assessments, utilizing a structured prompt format to enable diverse captions and effective evaluation. Our analysis reveals varying performance across sound categories and model architectures, with larger models generally excelling but innovative lightweight approaches also showing promise. The strong correlation between objective metrics and human ratings validates our evaluation approach. We discuss outcomes in terms of audio quality, controllability, and architectural considerations for text-to-audio synthesizers, providing direction for future research.

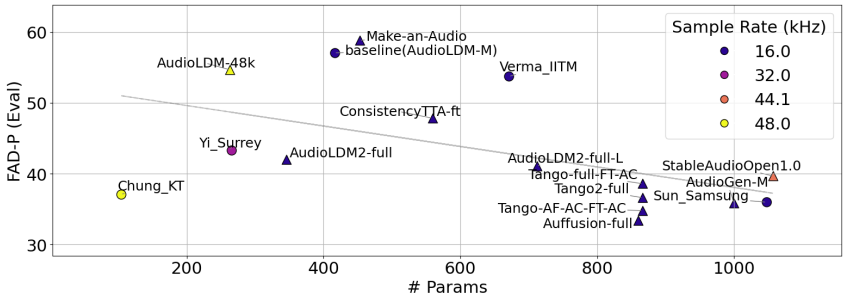


Figure 1: Performance of various Text-to-Audio models (circled markers: challenge submissions, triangular markers: open-source models) on the evaluation set versus their number of parameters. Color depicts the audio sample rate.

1 Introduction

Sound is of paramount importance in the creation of an immersive user experience in multimedia content such as movies and games, not to mention real-time applications such as the metaverse. By generating sound that aligns with a target sound description, audio generation systems would offer creators a greater range of options and streamline the workflow, reducing time and cost.

Recent advances in audio generation models have demonstrated considerable potential for automating and streamlining the process. However, the models face two significant challenges: a lack of audio quality that meets professional standards and a limited control over shaping desired sound sources.

[*]Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

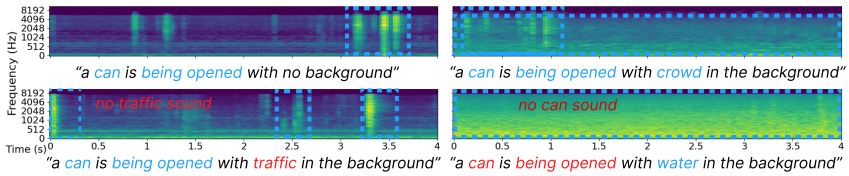


Figure 2: Examples that show limited controllability of a recent text-to-audio model (AudioLDM-M [1]) while controlling sound sources.

To highlight such limitations and facilitate further research, we organized a challenge on sound scene synthesis. The protocol we proposed and followed includes guidelines for dataset construction, objective metrics, and a human evaluation scheme to answer these research questions: 1) How can a model generate high-fidelity(quality) audio 2) How can we improve the diversity of the generated sounds with diverse foreground and background sound sources 3) How can we enhance the controllability of the model to generate audio given a corresponding text caption 4) How can we evaluate the category appropriateness, perceptual quality, and diversity of model-generated sounds.

2 Problem and Task Definition

In multimedia sound production, sound artists and engineers typically adhere to a structured process to create a final soundtrack. They first generate Foley sounds or collect samples from databases for each sound source. These samples are then edited to meet specific expectations regarding timbre, nuance, and temporal alignment. Finally, they mix all elements into a cohesive sound scene, often with music. TTA systems are designed to automate this process, but they face a number of challenges.

2.1 Problem in Current Text-To-Audio Systems

First, the quality of the generated audio is usually inadequate to meet commercial standards. Many TTA systems [1–4] generate audio waveforms at a 16 kHz sampling rate for training and inference efficiency, which is significantly lower than the industry standard of 48 kHz or higher. Second, their controllability through the text prompt is limited. Since controllability is crucial to achieve the desired sound characteristics, this limitation is a significant concern. Figure 2 illustrates how the well-known open-source TTA model (AudioLDM [1]) struggles with text-based controls. The generated sound is rarely aligned with both the foreground and background sounds, i.e., their *compositionality* is noticeably limited. This happens particularly when there is a strong positive or negative correlation between the foreground and the background in the training set.

Evaluation is another significant challenge, particularly because captions are incomplete descriptions of audio signals at varying levels of abstraction [5–7]. Fairly evaluating audio generation with a satisfaction score from such varied captioning styles presents considerable difficulties because of the seemingly endless possibilities of factors to consider. When evaluating a generated audio based on a caption *People in a small crowd are speaking and a dog barks* (from AudioCaps), for example, should the number of people speaking be considered? Does "and" imply the sounds to be sequential or simultaneous? How should all these factors be weighted to compute the satisfaction score? Once we answer these questions, how can we aggregate the score of this example with a score of a much simpler prompt? Although it may not be practically possible to answer all these questions, a simplified protocol should be defined to organize a public challenge in a fair manner.

2.2 Task Definition

In general, sound scene synthesis refers to the task of generating environmental sound scenes that can accompany events in multimedia content to enhance the narrative experience, excluding speech and music. This Sound Scene Synthesis task is built on last year's Foley sound synthesis challenge [8, 9], expanding the scope from Foley sounds to general sound scenes by generalizing the conditioning from a single predefined sound category to a natural language prompt. The audio output requirement is a 4-second, 32-bit, 32kHz, mono-channel audio waveform. Each submitted model is required to generate 250 audio files within a 24-hour period using the computing environment of *Colab Pro+*.

The evaluation prompts are limited to the following structure: *'(foreground sound source) with (background sound source) in the background'*, with action-based foreground sounds and ambient background sounds specified within the parenthesis. This format was devised to enable quantified evaluation of diverse text prompts.

3 Official Dataset and Baseline System

Dataset Creation Prompts following the structure described in section 2.2 were crafted manually by the organizing team. We categorized foreground prompts into six categories: "animal," "vehicle," "human," "alarm," "tool," and "entrance." These foreground prompts are paired with five different backgrounds: "crowd," "traffic," "water," "birds," and "no background," except that vehicles are not paired with traffic. The "no background" permits evaluation of clean monophonic foreground audios.

The level of detail in prompts was adjusted depending on the nature of the sound source. For example, the foreground prompt "a jackhammer is pounding" provides a clear and self-sufficient description. Qualifiers such as "small" or "large" would contribute little to the perception of a jackhammer sound, and the action associated with this source is restricted to "pounding." In contrast, other prompts, such as "a dog barking," benefit from more detailed descriptions, where variations in size (e.g., "small dog" vs. "large dog") or action (e.g., "barking" vs. "whining") could yield perceptually different audios. To maintain consistency across the dataset, we empirically balanced the complexity of foreground prompts, acknowledging that certain sounds carry more inherent information and, therefore, do not necessitate additional qualifiers or actions.

A sound engineer from our team created 4-s audio files corresponding to each prompt based on sounds sourced mainly from `Freesound.org` but also from private libraries. In total, our dataset comprises 310 audio-captions, with approximately 50 in each foreground sound category and 60 per background category. The development and evaluation set contain respectively 60 and 250 of these audio-caption pairs. Two background categories, "no background" and "birds," are excluded from the development set. Consequently, the evaluation set contains more samples with "no background" and "birds" compared to the development set.

Baseline System We provided AudioLDM [1] as our baseline model. To ensure high quality and controllability, 9k hours of audio from 4 different sources were used for training. In addition, the model leveraged techniques such as the latent diffusion model and pretrained audio-text embedding [10], which made the training efficient. As the baseline model generates 10-second audio, which is longer than our configuration, we 1) chopped audio into 4-second segments with a hop size of 2 seconds and selected the largest energy segment, and 2) resampled it from 16kHz to 32kHz.

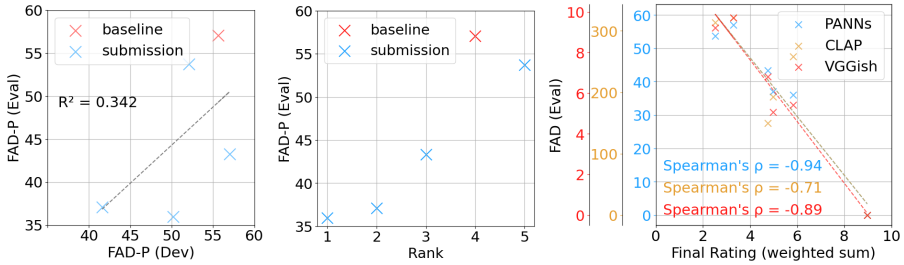
4 Evaluation

Following the previous challenge edition [8] , we conducted a two-stage evaluation scheme including both objective and subjective evaluation.

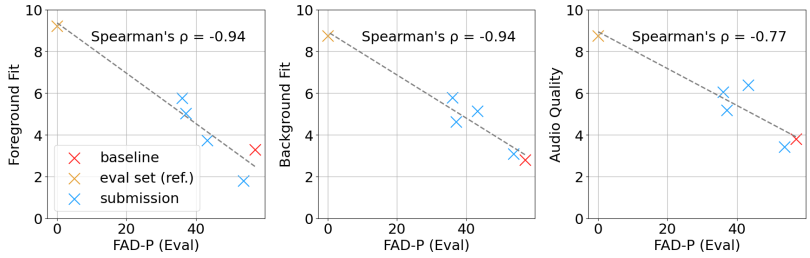
Step 1: Objective metrics To measure audio quality objectively, we adopted Fréchet Audio Distance (FAD) [11]. We chose FAD as it is a widely used metric in audio generation to measure set-wise audio quality and semantics compared to the reference set. For the embedding used in FAD, we used PANNs CNN14 Wavegram-Logmel [12] (denoted as FAD-P) since it showed the highest correlation scores with perceptual rating [13, 14]. We provided an official evaluation software. [2]

Step 2: Subjective metrics Subjective evaluation of audio fit ("how well the audio matches the sound of the prompt") and perceptual quality ("clarity, absence of artifacts and distortion") was performed for the four submitted systems, the provided baseline system, and the Sound-Designer Reference evaluation set. Four prompts from each of the six foreground categories were selected, spanning the five background categories. First, 148 randomly ordered trials were presented online (via the toolkit `Gorilla.sc`) in six sections separated by foreground category. Category orders were varied across raters. Each audio was given a separate rating for its match to the foreground and background portion of the prompt on a scale from 0 (extremely poor) to 10 (extremely good). Subsequently, the same 148 sounds were presented in random order, without a prompt, and were rated for perceptual

[2] `https://github.com/DCASE2024-Task7-Sound-Scene-Synthesis/fadtk`



(a) FAD-P on Evaluation set vs Development set (b) FAD-P on Evaluation set vs Challenge Ranking (c) FAD scores from different embeddings vs Final Rating



(d) FAD-P on Evaluation set vs Foreground Fit (e) FAD-P on Evaluation set vs Background Fit (f) FAD-P on Evaluation set vs Audio Quality

Figure 3: Correlation between FAD scores on evaluation set and other indicators, computed on the 4 submitted systems and the baseline system.

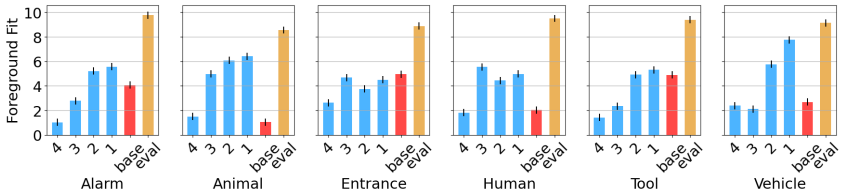


Figure 4: Subjective evaluation results on Foreground Fit. The error bar indicates the standard error.

quality (0-10) regardless of content. Rating one sound per trial was better suited to this purpose than comparing multiple sounds because each sound was unique [15].

Fourteen raters, four from each top team and ten from system-blinded organizers and their lab members, rated sounds from all systems. To avoid bias, for each contestant and each prompt, each self-rating was replaced with a contestant's average responses to that prompt for all other systems; replacement ensured that simple removal of self-ratings would not uniquely raise or lower a system's average. The Final Rating of each system is a weighted sum of its Foreground Fit, Background Fit, and Audio Quality in a 2:1:1 ratio.

5 Results

A total of four systems were received for submission [16–19]. In Figure 3a, the FAD scores of 4 systems and the baseline system are plotted. The (x, y) position represents the FAD score computed on the development set (FAD-P Dev) and the evaluation set (FAD-P Eval), respectively. First, the majority of systems exhibit a tendency to achieve lower FAD-P scores on the evaluation set when they are lower on the development set, with the exception [16, 18]. This is anticipated, as the training is based, at least in part, on the development set. Second, it turns out that FAD-P Dev is a noisy measure to predict FAD-P Eval. The exception may be attributed to the presence of new sound sources in the evaluation set to prevent overfitting, which may result in performance discrepancies between the two sets.

Figure 3b shows the final rankings of the 4 systems and the baseline system, as determined by the weighted summed score from the listening test, in conjunction with the FAD-P Eval and FAD-P Dev. The FAD-P Eval scores align well with the final rankings, while FAD-P Dev does not. The Spearman's correlation coefficient ρ of the ranking by FAD-P Eval and the final ranking is '0.900' ($p = 0.037$), while by FAD-P Dev it is '0.500' ($p = 0.391$). This discrepancy may be due to systems being overfitted to the development set or to the relatively small size of the development set since FAD is a biased metric [14, 20].

To validate the use of PANNs embedding in FAD calculation, we examined the correlation between FAD scores calculated from different embeddings and weighted summed scores, as illustrated in Figure 3c. The PANNs model demonstrated the highest Spearman's correlation coefficient of -0.94, in comparison to CLAP [10] and VGGish [21]. It is noteworthy that only the result of PANNs was statistically significant (i.e., $p < 0.05$). This result corroborates the previous study's findings [13].

Figure 3d to 3f illustrate the correlation between FAD-P and human subjective ratings. FAD-P shows a strong relationship with both foreground and background fit but a weak correlation with overall audio quality. This suggests that FAD-P primarily measures the audio-text correspondence, while it may be less sensitive to factors affecting overall quality, such as noise or generated artifacts.

To apply our dataset for evaluation, we additionally measured the FAD-P on the evaluation set with generated results from other open-source models [2–4, 22–27] (see Figure 1). Our findings revealed a consistent trend whereby scaling up resulted in enhanced performance, which aligns with the prevalent notion in the field of generative models. The number of model parameters was a more dominant factor than the model types (transformers or diffusion models) in general. However, there is one notable exception: Chung_KT [17] demonstrated promising performance in a lightweight GAN-based architecture. Secondly, it was observed that a model generating audio at a higher sample rate did not always achieve a better score in the evaluation set at 32 kHz. Currently, it is relatively under-optimized to train a model with a higher and more production-ready sampling rate.

The mean subjective ratings of Foreground fit, Background fit, and Audio Quality were appropriately low for the baseline system (3.3, 2.8, 3.8), appropriately high for the Reference Set (9.8, 8.8, 9.0), and moderately high for the top-ranked submitted system (5.8, 5.8, 6.0) [3]. Figure 4 shows the mean Foreground Fit ratings (and their standard errors, calculated over the distribution of 14 ratings, showing high interrater agreement, Cronbach's α = 0.959) for each submission within each foreground category. The Background Fit (not shown) correlates highly (r=0.79) with the Foreground Fit, and Audio Quality correlates highly with both Foreground Fit (r=0.85) and Background Fit (r=0.87). Although system rankings vary across categories, and different rankings do not always reflect a large mean difference, the overall winner (submission 1) has the highest rating in most of the foreground categories. The Entrance Category proved the most challenging for generative systems, with no submissions rated a higher fit than the baseline system, while all submissions fared better than the baseline system in the Animal Category.

6 Conclusion

The Sound Scene Synthesis challenge has yielded important insights into text-to-audio generation for environmental sounds. Our evaluation protocol, combining FAD-P metrics and human ratings, revealed both progress and areas for improvement in audio quality, diversity, and controllability. The structured prompt format facilitated diverse captions while enabling effective evaluation. While larger models generally excelled, innovative lightweight approaches also showed promise. Performance varied across sound categories, with some showing substantial improvement over the baseline. The strong correlation between FAD-P and human ratings, particularly for sound source fit, validates its use as a reliable objective metric for future research.

Future work should focus on enhancing the nuance, temporal aspects, and spatial capabilities of generated sounds. Refining evaluation metrics to capture subtle qualitative differences will be crucial. As this task serves as a valuable benchmark for assessing generative audio models, future iterations could incorporate more sophisticated prompts and criteria. Success in this domain could pave the way for more complex audio generation tasks such as video-to-audio synthesis, potentially revolutionizing AI-driven audio production for multimedia content.

[3] `https://dcase.community/challenge2024/task-sound-scene-synthesis-results`

References

[1] 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. In *Proceedings of the 40th International Conference on Machine Learning*, pages 21450–21474, 2023.

[2] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3590–3598, 2023.

[3] 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.

[4] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

[5] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

[6] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[7] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020.

[8] Keunwoo Choi, Jaekwon Im, Laurie Heller, Brian Mcfee, Keisuke Imoto, Yuki Okamoto, Mathieu Lagrange, and Shinnosuke Takamichi. Foley sound synthesis at the dcase 2023 challenge. In *2023 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2023)*, 2023.

[9] Keunwoo Choi, Sangshin Oh, Minsung Kang, and Brian McFee. A proposal for foley sound synthesis challenge, 2022.

[10] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[11] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

[12] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[13] Modan Tailleur, Junwon Lee, Mathieu Lagrange, Keunwoo Choi, Laurie M Heller, Keisuke Imoto, and Yuki Okamoto. Correlation of fréchet audio distance with human perception of environmental audio is embedding dependent. In *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024.

[14] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335. IEEE, 2024.

[15] B Series. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*, 2, 2014.

[16] Xie ZhiDong, Li XinYu, Liu HaiCheng, Zou XiaoYan, and Sun Yu. Sound scene synthesis with audioldm and tango2 for dcase 2024 task7. Technical report, Samsung Research China-Nanjing, Nanjing, China, July 2024.

[17] Hae Chun Chung and Jae Hoon Jung. Sound scene synthesis based on gan using contrastive learning and effective time-frequency swap cross attention mechanism. Technical report, KT Corporation, Seoul, Republic of Korea, July 2024.

[18] Yi Yuan, Haohe Liu, Xubo Liu, Mark D. Plumbley, and Wenwu Wang. Diffusion based sound scene synthesis for dcase challenge 2024 task 7. Technical report, University of Surrey, Guildford, United Kingdom, July 2024.

[19] Sagnik Ghosh, Gaurav Verma, Siddharath Narayan Shakya, Shubham Sharma, and Shivesh Singh. Sound scene synthesis based on fine-tuned latent diffusion model for dcase challenge 2024 task 7. Technical report, Indian Institute of Technology Mandi, Kamand, Mandi, India, July 2024.

[20] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020.

[21] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 131–135. IEEE, 2017.

[22] 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*, 2024.

[23] Yatong Bai, Trung Dang, Dung Tran, Kazuhito Koishida, and Somayeh Sojoudi. Accelerating diffusion-based text-to-audio generation with consistency distillation. *arXiv preprint arXiv:2309.10740*, 2023.

[24] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generative models through direct preference optimization. In *ACM Multimedia 2024*, 2024.

[25] Zhifeng Kong, Sang-gil Lee, Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, Rafael Valle, Soujanya Poria, and Bryan Catanzaro. Improving text-to-audio models with synthetic captions. *arXiv preprint arXiv:2406.15487*, 2024.

[26] Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *arXiv preprint arXiv:2401.01044*, 2024.

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

A Challenge Task Overview

Figure 5: Overview of *Sound Scene Synthesis* task. A sound synthesis system (i.e., Text-to-Audio model) receives a text prompt as an input, and outputs an audio corresponding to the prompt.

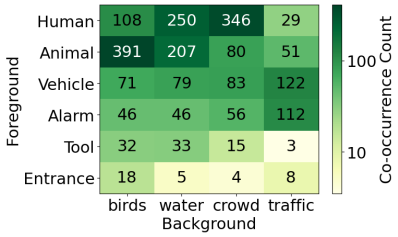


Figure 6: Heatmap of the co-occurrence of foreground-background combinations in AudioCaps[5] trainset. The data imbalance may potentially limit the model's controllability.

B Challenge Results

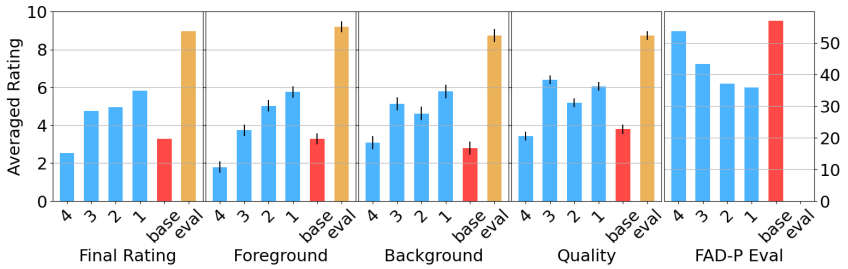


Figure 7: Evaluation score for each system averaged across all sounds: Far left panel, Final Rating, combines subjective ratings of Foreground: Background: Quality with 2:1:1 weighting. Far right panel, Objective evaluation score (FAD-P Eval). The error bar indicates the standard error. The official ranking is as follows: 1^{st} *Sun_Samsung*[16], 2^{nd} *Chung_KT*[17], 3^{rd} *Yi_Surrey*[18], 4^{th} *Verma_IITM*[19].

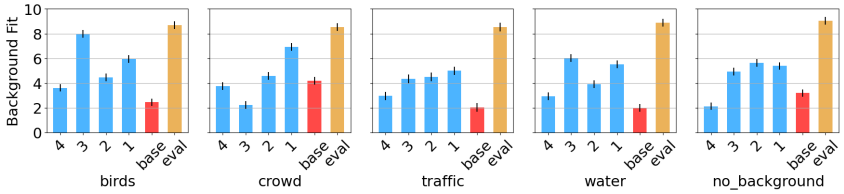


Figure 8: Subjective evaluation results on Background Fit. The error bar indicates the standard error.

These additional figures are provided to display the challenge results on subjective evaluation. The x-axis indicates the official ranking, where "base" refers to the baseline system and "eval" denotes the reference evaluation set created by a sound designer. Figure 7 depicts the Final Rating (i.e., weighted sum as outlined in Section 4), in conjunction with other averaged scores and FAD-P. Note that unlike other metrics, a lower FAD-P score means better performance and the FAD-P for "eval" is zero. Figure 8 illustrates the mean Background Fit within each background category. Figure 9 shows the mean Audio Quality within each foreground category. The resulting inter-rater agreement, among 14

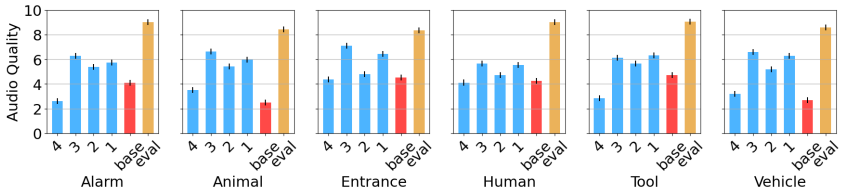


Figure 9: Subjective evaluation results on Audio Quality. The error bar indicates the standard error.

raters over 144 prompts, was high (Cronbach's $\alpha = 0.959$). Please refer to our challenge homepage for further detailed results and numerical data. [4]

C Subjective Evaluation

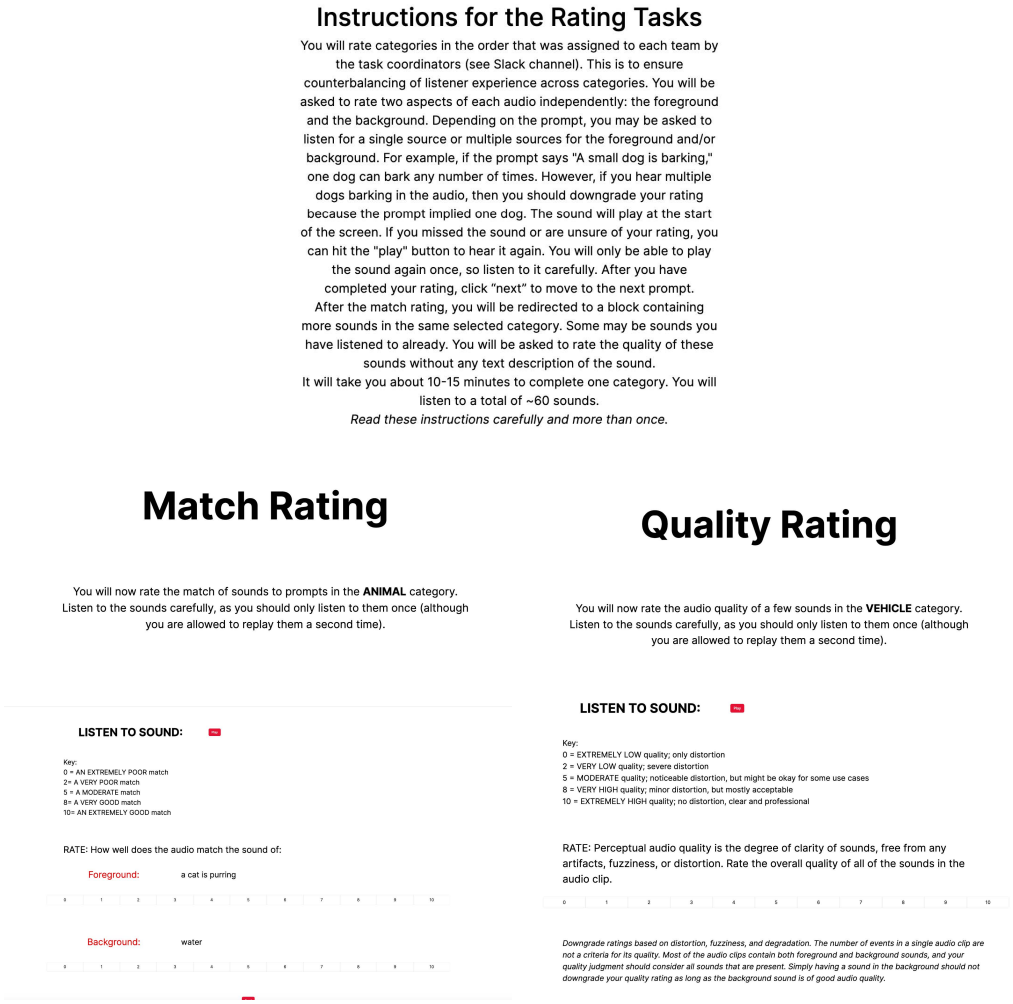


Figure 10: Screenshots of subjective evaluation toolkit

Figure 10 shows the screenshot of the platform used for subjective evaluation. See Section 4 for more details.

[4] https://dcase.community/challenge2024/task-sound-scene-synthesis-results