https://doi.org/10.31891/2307-5732-2025-351-3
УДК 004.8:004.93

BABIAK YURII
Lviv Polytechnic National University
https://orcid.org/0009-0009-2771-3389
e-mail: yurii.babiak.mnsam.2023@lpnu.ua
MEDIAKOV OLEKSANDR
Lviv Polytechnic National University
https://orcid.org/0000-0002-2580-3155
e-mail: oleksandr.mediakov.mnsam.2023@lpnu.ua
BASYUK TARAS
Lviv Polytechnic National University
https://orcid.org/0000-0003-0813-0785
e-mail: taras.m.basyuk@lpnu.ua

EVALUATION AND COMPARISON OF TEXT-TO-AUDIO GENERATION MODELS FOR MEDIA APPLICATIONS

In this research paper we aim to evaluate and compare the performance of several state-of-the-art text-to-audio generation models in producing audio effects for media applications. To achieve this, we created a new evaluation framework, including curated dataset of text-audio pairs that can be used in media products, and a comprehensive set of metrics, namely: Kullback–Leibler divergence between classification labels of true and generated audio, the Contrastive Language–Audio Pretraining (CLAP) embedding similarity, text-caption cosine similarity, and Fréchet Audio Distance (FAD) between expected and generated audios. Our results demonstrate that Stable Audio Open exhibited the highest performance across most metrics, indicating superior audio quality and semantic alignment. This comprehensive study not only quantifies the performance of these models but also provides a detailed analysis of their strengths and weaknesses in a real-world media production context. The findings reveal the intricate relationship between model architecture, training strategies, and the resulting audio quality. We also found that increasing inference steps generally improved semantic alignment but with diminishing returns beyond 100 steps. Our results also include investigation into the trade-off between models' sizes, training strategies and performance. Scientifically, this study provides a new solid benchmark for evaluating text-to-audio generation models and contributes to a deeper understanding of diffusion-based audio synthesis. Practically, our findings offer clear guidance for media creators and developers in selecting appropriate models for specific applications, facilitating the integration of advanced audio generation into media production. Furthermore, the curated dataset and defined metrics serve as valuable resources for future research in this field.

Keywords: diffusion models, audio generation, reverse diffusion, text-to-audio generation, generative AI evaluation.

БАБ'ЯК ЮРІЙ
МЕДЯКОВ ОЛЕКСАНДР
БАСЮК ТАРАС
Національний університет "Львівська політехніка"

ОЦІНКА ТА ПОРІВНЯННЯ ТЕКСТ-ДО-АУДІО ГЕНЕРАТИВНИХ МОДЕЛЕЙ ДЛЯ ЗАСТОСУВАННЯ У МЕДІА

У цій роботі запропоновано оцінення та порівняння якість кількох найсучасніших моделей генерації аудіо з тексту у створенні звукових ефектів для застосування у медіа продуктах. Для цього у роботі пропонується нова система оцінювання, що включає відібраний набір даних з пар текст-аудіо, які можна використовувати в медіа-продуктах, та комплексний набір з чотирьох метрик. Це комплексне дослідження не лише кількісно оцінює продуктивність цих моделей, але й надає детальний аналіз їхніх сильних та слабких сторін у контексті реального застосування у медіа продуктах. Висновки розкривають комплексний зв'язок між архітектурою моделі, стратегіями навчання та отриманою якістю аудіо. Результати також включають дослідження компромісу між розмірами моделей, стратегіями навчання та якістю генерованого аудіо. З наукової точки зору, це дослідження надає новий надійний бенчмарк для оцінювання моделей генерації аудіо з тексту та сприяє глибшому розумінню синтезу аудіо на основі дифузії. З практичної точки зору, висновки у роботі пропонують визначені рекомендації для медіа-творців та розробників у виборі відповідних моделей для конкретних застосунків, сприяючи інтеграції передової генерації аудіо в медіа продукти. Крім того, відібраний набір даних та визначені метрики слугуватимуть цінними ресурсами для майбутніх досліджень у цій галузі..

Ключові слова дифузійні моделі, генерація аудіо, зворотна дифузія, генерація тексту-в-аудіо, оцінка генеративного ШІ.

Стаття надійшла до редакції / Received 11.04.2025
Прийнята до друку / Accepted 26.04.2025

Introduction

The creation of audio effects through advanced generative neural models is a game-changer for improving media products in various fields, such as music production, virtual assistants, and immersive environments. These generative AI models are able to produce dynamic and realistic audio, going beyond the limitations of traditional methods by learning intricate patterns in audio data. This capability enables the creation of high-quality sounds and music, generating novel soundscapes and simulating acoustic environments.

Our paper focuses on the evaluation and comparison of several state-of-the-art text-3-audio generative models for the audio effect generation task. To achieve this, we have created an evaluation dataset of text-audio pairs relevant to media applications. To do this, we've put together an evaluation dataset filled with text-audio pairs that are relevant to media applications. We then defined four key performance metrics to assess the quality of the generated audio, including CLAP embedding similarity for semantic alignment, text-caption similarity for semantic consistency, FAD (based on VGGish embeddings) for audio quality and perceptual similarity and KL divergence (based on PaSST model) for semantic content similarity.

This research helps deepen the understanding of current audio generation techniques, paving the way for improvements in sound quality, efficiency, and control, ultimately enhancing how AI integrates into creative media production.

Analysis of related works on conditional text-to-audio generative models

The use of diffusion models has led to notable progress in the creation of audio from textual descriptions [1]. These models are types of neural networks [2] where the stochastic process of progressively turning data into noise is reversed. Denoising Diffusion Probabilistic Models (DDPMs) [3], which iteratively eliminate noise to restore the original data distribution, are essential to this strategy. Efficiency of classical diffusion models can be improved by Latent Diffusion Models (LDMs) by leveraging reverse diffusion process in the smaller latent space instead of the original one [6]. Other similar techniques include Noise Conditional Score Networks (NCSNs)[4], which improve denoising by conditioning on noise levels and estimate the gradient of the data distribution, while Score-based Generative Models (SGMs)[5] are continuous-time framework for simulating noise addition and reversal is offered by stochastic differential equations (SDEs). Additionally, Variational Diffusion Models (VDMs)[6] incorporate variational inference for optimization. The comparison of the strong and weak points of the listed types of generative models' architectures that can be used for audio generation are put together in Table 1.

Table 1

Generative models' architectures comparison

Model type	Strong points or improvements	Weak points
Noise Conditional Score Networks [4]	A clear and well-defined training objective based on score matching, which allows for principled model comparisons.	Sampling using Langevin dynamics can be computationally expensive, and model's performance is heavily reliant on the accuracy of the score estimation.
Denoising Diffusion Probabilistic Models [3]	Iteratively refine audio by removing noise, leading to high quality, plus method has a strong theoretical foundation.	Slow sampling speed, high computational cost and can be memory intensive for long audios or big images.
Score-based Generative Models [5]	Can handle complex data distributions, plus provide a unified framework for various diffusion processes.	Requires accurate score estimation. Can be unstable during training and can be computationally demanding.
Latent Diffusion Models [7]	Significantly improve sampling speed and reduce computational cost, as a result enables high-resolution generation with limited resources.	Potential loss of fine-grained details, plus require careful selection of the latent space.
Variational Diffusion Models [6]	Combine diffusion models with variational inference. Provide a probabilistic framework for optimization.	Increase complexity due to variational inference, also due to approximation may affect sample quality.

These diffusion-based techniques have found practical application in conditional text2audio generation, exemplified by models like Stable Audio Open [8], an open-source model optimized for generating short audio samples, sound effects, and production elements. It employs an autoencoder for waveform compression, T5-based text embeddings for conditioning, and a transformer-based diffusion model (DiT) [10] in the latent space. Trained on datasets like Freesound[8] and the Free Music Archive, it produces high-quality stereo audio at 44.1kHz.

Another popular model is AudioLDM [11, 12], inspired by Stable Diffusion. It's a latent diffusion model that leverages Contrastive Language–Audio Pretraining (CLAP) [13, 14] latents to generate realistic audio samples from text prompts. It demonstrates the ability to synthesize a wide range of audio content, including sound effects, speech, and music. We evaluated both small and large variants of AudioLDM versions 1[11] and 2 [12].

MusicLDM [15], built on ideas from both Stable Diffusion and AudioLDM, created specifically for music generation. It utilizes beat-synchronous data augmentation strategies in both the time and latent domains, enhancing the diversity and stylistic fidelity of generated music. The model was trained with 466 hours of music data.

Evaluating the performance of these audio generation models requires usage of robust metrics, covering different aspects of generated audio and its similarity with input prompt. Common approaches include analyzing the semantic alignment between generated audio and input text, as well as assessing audio quality and diversity. To quantify semantic alignment, metrics like the cosine similarity between CLAP embeddings [13] of the input prompt and generated audio are employed. CLAP, based on the similar idea of CLIP [16], provides a common embedding space for texts and audios, enabling their direct comparison via vector metrics. Additionally, the cosine similarity between text prompt embeddings and captions generated for the audio offers another perspective on semantic consistency.

To assess audio quality and perceptual similarity other popular techniques are widely used, starting with the Fréchet Audio Distance (FAD) [17]. FAD, based on VGGish [18] embeddings, measures the distance between the distributions of real and generated audio features. Furthermore, the Kullback-Leibler (KL) divergence between label distributions extracted from pre-trained audio classifiers, such as Patch-based Audio Spectrogram Transformer (PaSST) [19], can evaluate the similarity of semantic content between generated and real audio [20]. These metrics collectively provide a comprehensive evaluation framework for text-to-audio generation models.

Formulation of the purpose of the research

The main purpose of the research is to evaluate and compare the performance of SOTA audio GenAI models, concretely focusing on their ability to produce high-quality audio effects from input textual prompts for media product applications. To achieve this in the work we set up a solid evaluation system that includes a chosen eval-set of task-relevant text-audio pairs and metrics to measure corresponding performance. The created framework is designed to assess six distinct audio generation models, providing a comparative analysis of their strengths and weaknesses, based on the received values. Generally, this research contributes to a deeper understanding of current audio generation models.

Main materials

As it was already mentioned in the previous section, the evaluation framework consists of dataset and a set of metrics. The dataset was constructed using publicly accessible audio files, ensuring compliance with copyright regulations through the selection of open-licensed materials. We gathered over a hundred audio files, each accompanied by descriptive prompts. These prompts served as the ground truth for assessing the semantic accuracy of the generated audio. To give the overall impression of the collected dataset the distributions of the audio length (in seconds) and the distribution of the number unique words in prompt text are visualized on Fig. 1.

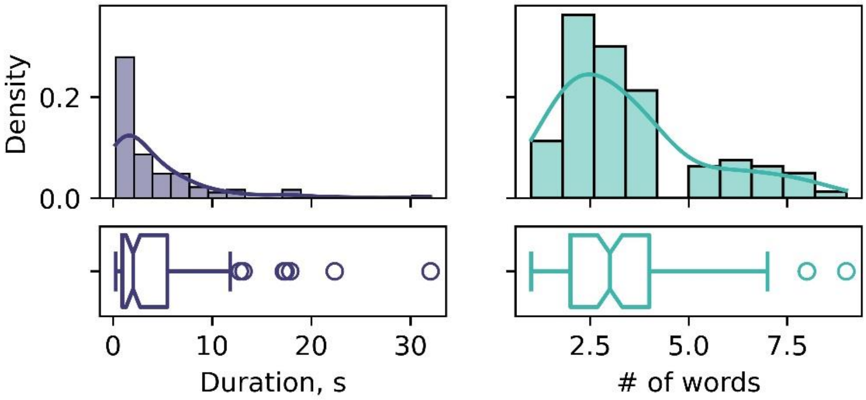

Fig. 1. Distributions of audio duration and prompt text length (number of unique words)

For evaluating the models, we propose using a set of four metrics. Firstly, we used the Kullback-Leibler (KL) divergence between label distributions [20] (for expected and generated audios) extracted from PaSST audio classifier [19]. This metric allowed us to quantify the semantic accuracy of the generated audio by comparing the predicted label distributions against those of real audio samples, revealing how well the generated audio aligned with expected sound categories. The lower the metric value – the better model performs.

Secondly, we utilized the similarity between CLAP embeddings of the input text prompt and the generated audio. This is the most straight-forward metric in the set. CLAP, designed to learn joint representations of audio and text. We used a pre-trained by Microsoft version [13, 14], providing a cross-modal evaluation of semantic alignment. By default, the CLAP similarity is calculated as a dot-product of the corresponding embeddings. The higher the value – the better the model.

Thirdly, we measured the cosine similarity between the embeddings of the input textual prompt and the caption generated for the audio. This indirect measure of semantic consistency between the intended audio description and the perceived content was calculated using the GIST-Embedding-v0 model [21]. To receive a textual description for the generated audio we used a conditional caption generation processed supported by

CLAP. For each generated audio sample, we have decoded a corresponding description which was compared with an actual input prompt using cosine similarity, given by Equation (1).

$$\text{cos-sim}(\boldsymbol{p}, \boldsymbol{c}) = \frac{\boldsymbol{p} \cdot \boldsymbol{c}}{\|\boldsymbol{p}\|\|\boldsymbol{c}\|}, \quad (1)$$

where vectors p and c are the embedding vectors extracted for an input prompt and generated caption correspondingly. The higher the value – the better the model. For the referencing to the metric the name PCES will be used (from prompt-caption embedding similarity).

Finally, we also propose to use the Fréchet Audio Distance (FAD) [17] with VGGish [18] embeddings to assess the quality and realism of the generated audio. Metric, which measures the distance between the distributions of real and generated audio features, captured both the quality of individual samples and the diversity of the generated set. For the generated audio, we calculated FAD for each batch and then averaged the results over three runs. VGGish embeddings were used as they were preferred by the authors of the original FAD paper [17].

Having two sets of embeddings for a real audios $\mathbf{E}_\text{r}$ and generated $\mathbf{E}_\text{g}$, firstly there is a need to estimate two multivariate Gaussian distribution based on the embedding sets: $N(\boldsymbol{\mu}_\text{r}, \boldsymbol{\Sigma}_\text{r})$, and $N(\boldsymbol{\mu}_\text{g}, \boldsymbol{\Sigma}_\text{g})$, based on which the FAD is defined by Equation (2) [17].

$$\text{FAD}\left(N(\boldsymbol{\mu}_\text{r}, \boldsymbol{\Sigma}_\text{r}), N(\boldsymbol{\mu}_\text{g}, \boldsymbol{\Sigma}_\text{g})\right) = \left\|\boldsymbol{\mu}_\text{r} - \boldsymbol{\mu}_\text{g}\right\|^2 + \text{Tr}\left(\boldsymbol{\Sigma}_\text{r} + \boldsymbol{\Sigma}_\text{g} - 2\sqrt{\boldsymbol{\Sigma}_\text{r}\boldsymbol{\Sigma}_\text{g}}\right), \quad (2)$$

where Tr is a trace of a matrix.

Collectively, these metrics provided a comprehensive framework for evaluating the semantic accuracy, cross-modal alignment, and perceptual quality of the generated audio.

For the evaluation process six models were selected, namely:

- AudioLDM [11, 12] version 1 and 2 (small and large versions), corresponding model checkpoints are:
  - cvssp/audioldm-l-full;
  - cvssp/audioldm-s-full-v2;
  - cvssp/audioldm2;
  - cvssp/audioldm2-large;
- MusicLDM [15], checkpoint: ucsd-reach/musicldm;
- and Stable Audio Open [8] version 1.0, checkpoint: stabilityai/stable-audio-open-1.0

For each of the models we utilized a curated dataset, while also exploring the impact of varying the number of reverse diffusion process steps (50, 100, and 200). Each experiment was run three times (generating three versions of the audio for a given input description) and averaging the resulting metrics. The aggregated metric values are presented in Table 1, with the best values per column highlighted in italics. Visualizations of these metrics are shown in Figure 1, with arrows indicating whether higher or lower values are preferred.

Table 2

Resulting table

Model	Steps	KL div ↓	CLAP sim. ↑	PCES ↑	FAD ↓
cvssp/audioldm-l-full	50	3.777663	3.255752	0.632183	5.831618
cvssp/audioldm-l-full	100	3.738220	3.296518	0.634056	5.853212
cvssp/audioldm-l-full	200	3.730167	3.302370	0.633000	5.910927
cvssp/audioldm-s-full-v2	50	3.852858	2.946991	0.634231	6.869589
cvssp/audioldm-s-full-v2	100	3.752894	3.022142	0.634018	6.608230
cvssp/audioldm-s-full-v2	200	3.863988	3.029306	0.632912	6.955686
cvssp/audioldm2	50	4.063242	3.025873	0.639320	5.488616
cvssp/audioldm2	100	4.056284	3.065535	0.639993	5.402819
cvssp/audioldm2	200	4.090465	3.099258	0.639077	5.475049
cvssp/audioldm2-large	50	4.360254	2.460987	0.636093	5.434517
cvssp/audioldm2-large	100	4.356609	2.465132	0.636733	5.396087
cvssp/audioldm2-large	200	4.292400	2.641249	0.637267	5.474047
stabilityai/stable-audio-open-1.0	50	*2.458611*	6.364019	*0.675837*	3.697069
stabilityai/stable-audio-open-1.0	100	2.498052	6.383319	0.674974	3.730893
stabilityai/stable-audio-open-1.0	200	2.493545	*6.421551*	0.674869	*3.656341*
ucsd-reach/musicldm	50	5.077684	1.667891	0.623549	9.816323
ucsd-reach/musicldm	100	5.071681	1.697881	0.623988	9.900907
ucsd-reach/musicldm	200	5.079560	1.732906	0.623814	10.001324

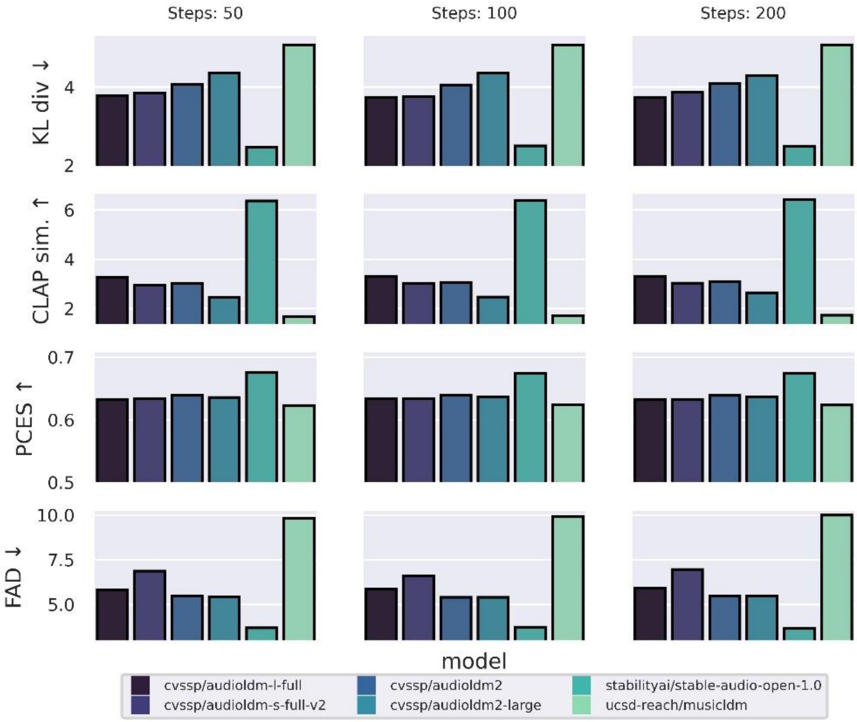

Fig. 2. Comparison matrix by metric per inference steps per model

Our analysis revealed several key observations. Firstly, increasing the number of inference steps generally led to slight improvements in CLAP similarity, suggesting a potential enhancement in semantic alignment with more refined generation. However, the effect on other metrics was less consistent, indicating a complex interplay between inference steps and performance across different evaluation metrics. CLAP similarity also exhibited the widest range of variations, pointing to its sensitivity to the model and parameter changes.

KL divergence showed significant variability w.r.t. models but mostly remained stable within each model across different inference steps. This suggests that the inherent semantic categorization capabilities of each model were largely consistent. Unfortunately, the caption-prompt embeddings cosine similarity remained relatively uniform across all models and inference steps, typically falling within the 0.6-0.7 range. This consistency may indicate limitations in the GIST-Embedding-v0 model's sensitivity to changes in audio descriptions or potentially points to over-simplified or generic generated captions. Further investigation is needed to clarify this observation.

FAD also varied substantially between models, demonstrating consistency within each model across different inference steps. This suggests that the overall audio quality and realism, as captured by FAD, were inherent to each model's architecture and training.

Notably, models with lower KL divergence tended to have higher CLAP similarity, possibly pointing to a correlation between semantic accuracy and cross-modal alignment. For most models, increasing inference steps beyond 100 yielded diminishing returns or even slight degradation in some metrics, particularly FAD, indicating a trade-off between refinement and potential overfitting or artifact generation.

Additionally, to compare the models not only from metrics perspective, but also from size-metrics point of view, we created the visual representation of the dependencies between models' sizes and two selected metrics: CLAP and FAD (given by the fact that they tend to have a correlation), presented on Fig. 3. Number of steps, used to create Fig. 3, was chosen to be a middle value – a 100.

Inspecting results from Fig. 2 and Fig. 3 we can extract insight with respect to the model, rather than generation parameters. As expected, Stable Audio Open demonstrated strong overall performance across most metrics, confirming its efficacy in generating high-quality audio from text prompts. AudioLDM-large exhibited a surprisingly high KL divergence, despite its larger architecture, while showing noticeable improvements in CLAP similarity at 200 inference steps. MusicLDM generally performed poorly across most metrics, likely due to its specialization in music generation rather than general audio effects, validating the suitability of our metrics for evaluating text-to-audio tasks.

Comparing the two versions of AudioLDM, the larger model consistently outperformed the smaller model in CLAP similarity and FAD, indicating enhanced semantic alignment and audio quality. However, KL divergence and caption-prompt cosine similarity remained similar, suggesting that these metrics were less sensitive to model size variations.

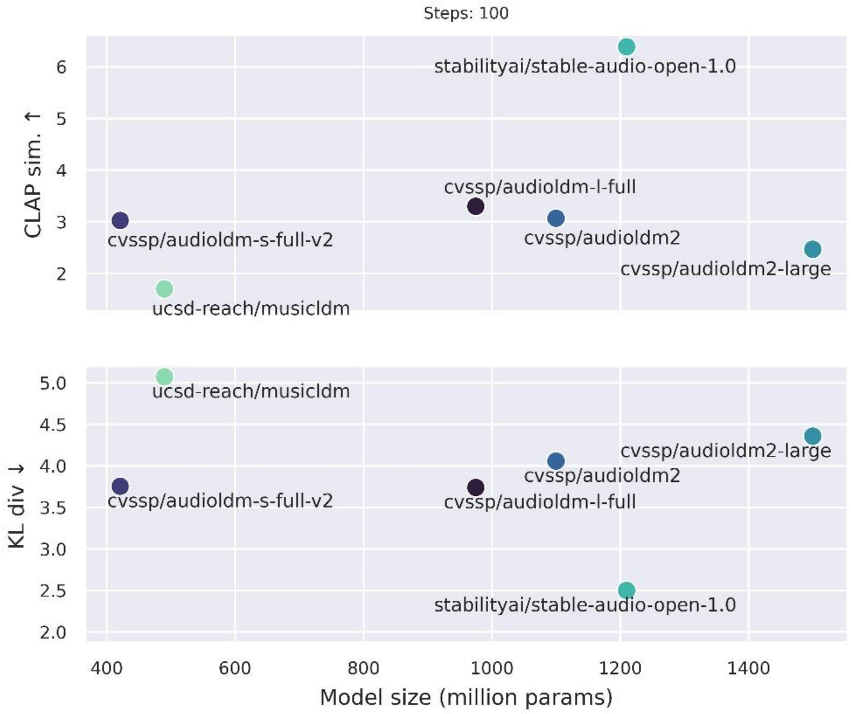

Fig. 3. Models' size by metrics values

Conclusions and further work

Our research aimed to evaluate and compare the performance of six state-of-the-art text-to-audio generation models using a curated dataset and a comprehensive set of metrics.

We used a selected dataset and a wide range of measurements to do this. Our findings showed that Stable Audio Open had the best overall results. It did especially well in Fréchet Audio Distance (FAD) and CLAP similarity. These scores point to better sound quality and a closer match to the intended meaning. We also noticed that more inference steps led to better CLAP similarity, but this improvement slowed down after a while, and other scores, e.g., FAD, could get worse after increasing number of steps. Also, MusicLDM, while designed for music generation, performed poorly across most metrics. Looking at all the models side by side showed how the model structure and training data affect performance.

Future work could aim to create custom audio generation models for specific uses possibly by fine-tuning or coming up with new designs. Additionally, to increase the validity of evaluation framework there is a need to improve how the dataset is constructed, for instance, utilizing data augmentation and self-supervised learning. To make metrics more reliable and suitable for evaluating conditional generation of audio for media production.

References

1. Božić, M., & Horvat, M. (2024). *A survey of deep learning audio generation methods*. https://doi.org/10.48550/arXiv.2406.00146

2. Mediakov, O., & Basyuk, T. (2022). Specifics of designing and construction of the system for deep neural networks generation. In *CEUR Workshop Proceedings: Vol. 3171. Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022)* (pp. 1282–1296).

3. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In H. Larochelle et al. (Eds.), *Advances in Neural Information Processing Systems* (pp. 6840–6851).

4. Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In H. Wallach et al. (Eds.), *Advances in Neural Information Processing Systems*.

5. Song, Y., et al. (2020). *Score-based generative modeling through stochastic differential equations*. https://arxiv.org/abs/2011.13456

6. Kingma, D. P., et al. (2021). *Variational diffusion models*. https://arxiv.org/abs/2107.00630

7. Rombach, R., et al. (2022). High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr52688.2022.01042

8. Evans, Z., et al. (2024). *Stable Audio Open*. https://doi.org/10.48550/arXiv.2407.14358

9. Ramires, A., et al. (2020). *The Freesound loop dataset and annotation tool*. https://arxiv.org/abs/2008.11507

10. Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4172–4182). https://doi.org/10.1109/ICCV51070.2023.00387

11. Liu, H., et al. (2023). AudioLDM: Text-to-audio generation with latent diffusion models. In A. Krause et al. (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 21450–21474).

12. Liu, H., et al. (2024). AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 2871–2883.

13. Elizalde, B., et al. (2023). CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1–5).

14. Elizalde, B., Deshmukh, S., & Wang, H. (2023). *Natural language supervision for general-purpose audio representations*. https://arxiv.org/abs/2309.05767

15. Chen, K., et al. (2024). MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1206–1210).

16. Radford, A., et al. (2021). *Learning transferable visual models from natural language supervision*. https://arxiv.org/abs/2103.00020

17. Kilgour, K., et al. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech 2019* (pp. 2350–2354).

18. Hershey, S., et al. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131–135).

19. Koutini, K., et al. (2022). Efficient training of audio transformers with patchout. In *Interspeech 2022* (pp. 2753–2757).

20. Yuan, Y., et al. (2023). *Leveraging pre-trained AudioLDM for sound generation: A benchmark study*. https://arxiv.org/abs/2303.03857

21. Solatorio, A. V. (2024). *GISTEmbed: Guided in-sample selection of training negatives for text embedding fine-tuning*. https://arxiv.org/abs/2402.16829