APPLIED COMPUTER SCIENCE

Applied Computer Science 2025, vol. 21, no. 4, pp. 13–20
https://doi.org/10.35784/acs_7549

Submitted: 2025-04-08 | Revised: 2025-05-18 | Accepted: 2025-06-01

Keywords: text-to-sound generation, diffusion model, cumulative learning, mel-spectrogram tokens

Haitham ALHAJI [1*], *Alaa Yaseen TAQA* [2]
[1*] Computer Science Department, College of Computer Science and Mathematics, University of Mosul, Nineveh, Iraq, haithamtalhaji@yahoo.com
[2*] Computer Science Department, College of Education for Pure Science, University of Mosul, Nineveh, Iraq, alaa.taqa@uomosul.edu.iq
[*] Corresponding author: haithamtalhaji@yahoo.com

SoundCrafter: Bridging text and sound with a diffusion model

Abstract

Text-to-sound systems have recently attracted interest for their ability to synthesize common sounds from textual descriptions. However, previous research on sound generation has shown limited generation quality and increased computational complexity. We present SoundCrafter, a text-to-sound generation framework that utilizes diffusion models. Unlike previous methods, SoundCrafter operates within a compressed domain of mel spectrograms and is driven by semantic embeddings derived from the CLAP model, which stands for contrastive language audio pretraining. SoundCrafter improves generation quality and computational efficiency by learning the sound signals without modeling the cross-modal interaction. In addition, we employ a curricular learning technique by progressively increasing spectrogram resolution to stabilize training and improve output fidelity. SoundCrafter distinguishes itself by integrating CLAP-conditional semantic embeddings with a diffusion model that operates in the compressed domain of mel-spectrograms. Using the AudioCaps dataset, it achieves superior text-to-sound synthesis with a Fréchet Distance (FD) of 23.45 and an Inception Score (IS) of 7.57 - exceeding the performance of previous models while requiring significantly less computational resources and training on a single GPU.

1. INTRODUCTION

TTS, which stands for text-to-sound, synthesis, is an ongoing task in generative artificial intelligence that aims to synthesize audio waveforms from natural language descriptions, similar to other contemporary AI applications in various domains, such as (Talal & Anas, 2025). Although TTS and text-to-music systems have advanced due to specialized data sets and organized patterns, sound effects (SFX) synthesis presents a broader and more ambiguous challenge. It presents a novel set of problems where techniques and datasets for TTS or music often prove inadequate. This challenge involves the synthesis of diverse non-speech, non-music audio, including environmental sounds, animal vocalizations, mechanical operations, and human activities, derived from open-ended textual prompts (e.g., "a dog barking in the rain," "footsteps in a hallway") (Barahona-Ríos & Collins, 2021).

Unlike speech or music, sound effect audio presents unique challenges: it often exhibits weak temporal regularity, may involve multiple overlapping sources, and generally requires the modeling of highly variable acoustic properties. The limited availability of large, high-quality paired datasets (text-SFX audio) hampers large-scale supervised training (J. Huang et al., 2023). These issues require the creation of robust, generalizable models capable of understanding diverse stimuli and producing coherent, lifelike audio. In addition, there are several techniques to improve the audio, such as (Issa & Al-Irhaym, 2021).

Recent advances in generative modeling-specifically, autoregressive transformers, diffusion models, and latent representation learning-have led to significant improvements in the fidelity and controllability of audio generation. Early methods, such as AudioGen, demonstrated the feasibility of autoregressive token generation for text-based audio synthesis. Later, models such as DiffSound implemented discrete and latent diffusion mechanisms that significantly improved the quality and efficiency of audio generation. In addition, all of the aforementioned methods rely primarily on neural networks (Hasoon & Al-Hashimi, 2022). Recent initiatives, such as TANGO and AudioLCM, are exploring methods to reduce data requirements and inference time while maintaining superior generation quality, thereby increasing the accessibility and practicality of TTA models.

Despite this progress, there is still a lack of targeted evaluation and analysis of models developed explicitly for text-to-sound generation, not for speech or musical output, but for intricately layered ambient and contextual soundscapes. SoundCrafter addresses the research gap created by the lack of high-quality text-to-sound generation models that operate effectively without relying on large paired text-audio datasets or discrete token representations. This study presents SoundCrafter, an innovative framework for text-to-sound generation.

Our key contributions are as follows:

- CLAP-conditioned semantic embeddings: We use contrastive language-audio pretraining (CLAP) to provide robust and semantically enriched conditioning for sound synthesis.
- SoundCrafter uses compressed mel-spectrogram diffusion for denoising, unlike previous models that operate on raw audio or high-dimensional spaces, improving both speed and stability.
- Curriculum learning for resolution scaling: We use a curriculum learning approach that incrementally increases spectrogram resolution, thereby stabilizing training and increasing output quality.

2. RELATED WORK

In recent years, there has been a surge of research in text-to-speech (TTS) generation, driven by developments in generative modeling and the availability of linked text-audio datasets (Cherep et al., 2024). Initially, research focused primarily on speech and music synthesis; however, there is a growing interest in the creation of sound effects (SFX), a more complex and understudied field that encompasses a variety of often overlapping non-speech sounds, including ambient noise, animal calls, and mechanical activity (Yuan et al., 2023).

Kreuk et al. (2022) is one of the pioneering and influential efforts in this area, using an autoregressive transformer to generate discrete audio tokens based on textual input. AudioGen demonstrated the feasibility of modeling general sound effects with a GPT-style architecture, and introduced multi-stream decoding to more effectively describe overlapping audio events. Despite its remarkable results, the autoregressive nature of the model resulted in long inference times and the possibility of error accumulation over long sequences.

To alleviate these limitations, DiffSound, Yang et al. (2022) proposed a discrete diffusion model operating in the tokenized mel-spectrogram domain using a VQ-VAE encoder. DiffSound achieved significant improvements in generation speed and audio quality by replacing the autoregressive decoder with a non-autoregressive diffusion process. It emphasized the ability of diffusion models to parallelize sequence generation for complex auditory environments.

Tab. 1. Comparative examination of text-to-image models in relevant studies

Year	Model	Arch.	Dataset	Metrics	Limitations
2022	DiffSound	VQ-VAE, AR, non-AR	AudioCaps, AudioSet	MOS, FD	elevated computational expense and minor degradation in quality with accelerated inference
2023	AudioGen	AR Transformer	AudioSet, BBC, AudioCaps, Clotho, FSD50K	FAD, KL	Prolonged inference duration, diminished quality in multi-stream modelling, inadequate speech synthesis, restricted temporal comprehension, and dataset bias stemming from YouTube-derived data.
2023	Make-An-Audio	latent diffusion, CLAP, HiFi-GAN	AudioSet, BBC Sound Effects, Audiostock, AudioCaps	FD, KL	Demands gradual, resource-intensive diffusion processes; deteriorates with insufficient data.
2023	TANGO	LDM, T5, VAE, HiFi-GAN	AudioCaps, AudioSet, Freesound, WavCaps	FD, KL, FAD	Limited compositional control due to small training datasets, leading to similar outputs for distinct prompts and requiring larger datasets for better differentiation.
2024	AudioLCM	LCM, Transformer	AudioCaps, WavCaps, WavText5K	FD, KL	Quality loss in one-step sampling, limited gains beyond 10 steps

Subsequent work investigated instruction-tuned language models as text encoders for TTA. TANGO Ghosal et al. (2023) used a static Flan-T5 model to encode prompts, allowing for improved comprehension of complex and subtle instructions. Although trained on slightly smaller datasets, TANGO outperformed several baselines, demonstrating the benefits of using large-scale NLP models for audio production.

To improve scalability and data diversity, (R. Huang et al., 2023) augmented training by generating pseudo-captions for unlabeled audio, facilitating the use of over one million audio recordings using Make-An-Audio. The model used a latent diffusion pipeline with multimodal conditioning, achieving superior performance in both objective and subjective evaluations, especially in complex SFX situations.

Recent initiatives have focused on optimizing efficiency and facilitating real-time generation. LAFMA (Guan et al., 2024) presented a latent flow matching method that significantly reduces the number of inference steps while maintaining quality, and AudioLCM (Liu et al., 2024) has been enhanced by the use of consistency models to facilitate one- or two-step generation, making real-time applications feasible.

Together, these models represent the pinnacle of text-to-SFX generation technology. While diffusion-based approaches dominate, innovations in data efficiency, semantic alignment, and rapid inference continue to shape the evolution of the field. (R. Huang et al., 2023).

3. PROPOSED METHODOLOGY

SoundCrafter is a text-to-sound generation system that uses a diffusion model conditioned on semantic embeddings derived from a pre-trained CLAP model (Karchkhadze et al., 2024). The technology enables high-quality sound synthesis from natural language prompts with coupled text-audio data during training. The system consists of three main parts: a CLAP-based semantic coder, a latent diffusion model, and a VAE-based spectrogram reconstruction module. The entire generation process takes place in the latent space of mel spectrograms, allowing for more efficient training and inference compared to raw waveform generation.

We apply a curriculum learning technique by initially training the model on low-resolution spectrograms and steadily increasing the resolution as training progresses. This method helps stabilize early learning and improves final audio fidelity. At each stage of the curriculum, the length and resolution of the spectrogram is incrementally adjusted, ensuring that the model learns from coarse to fine representations.

The SoundCrafter system is capable of generating an audio sample based on a textual description. In probabilistic generative modeling, a diffusion model is used to approximate the true conditional distribution of the data, as shown in equation (1).

$$data\ distribution = q(z_0|E^{text}) \tag{1}$$

Text embedding E^{Text} is derived from the pre-trained text coder in CLAP, while z_0 represents the previous position of an audio sample within the space created by the compressed representation of the mel spectrogram.

Diffusion Models (Zhang et al., 2023) include two procedures: a forward procedure and a regressive procedure. The former was used to transform the data distribution into a standard Gaussian distribution with a specified noise schedule, while the latter was used to incrementally generate data samples from the noise according to an inference noise schedule. The transition probability in the forward process is expressed as in Eq. (2).

$$q(z_n \mid z_{n-1}) = \mathcal{N}\left(z_n; \sqrt{1-\beta_n}\, z_{n-1}, \beta_n I\right) \tag{2}$$

While the reverse process commences with a Gaussian noise distribution and the text embedding E^{text}. A denoising process, conditioned on E^{text}, progressively builds the audio prior z_0 by the Eq. (3):

$$p_\theta(z_{0:N} \mid E^{text}) = p(z_N)\prod_{t=n}^{N} p_\theta(z_{t-1} \mid z_t, E^{text}) \tag{3}$$

We use the Unet shown in Fig. 1The UNet model, which is the basis for stable diffusion, is the underlying structure for the diffusion model in SoundCrafter. The UNet model depends on both the time step and the CLAP embedding. We convey the time step in a one-dimensional embedding and then concatenate it with the embedding as conditioning data. Since our conditioning vector is unidimensional, we do not use the cross-attention technique in StableDiffusion for conditioning. We use the feature-wise linear modulation layer (Perez et al., 2017) to merge the conditioning information with the feature map of the UNet convolution block. The UNet architecture used consists of four encoder blocks, one center block, and four decoder blocks.

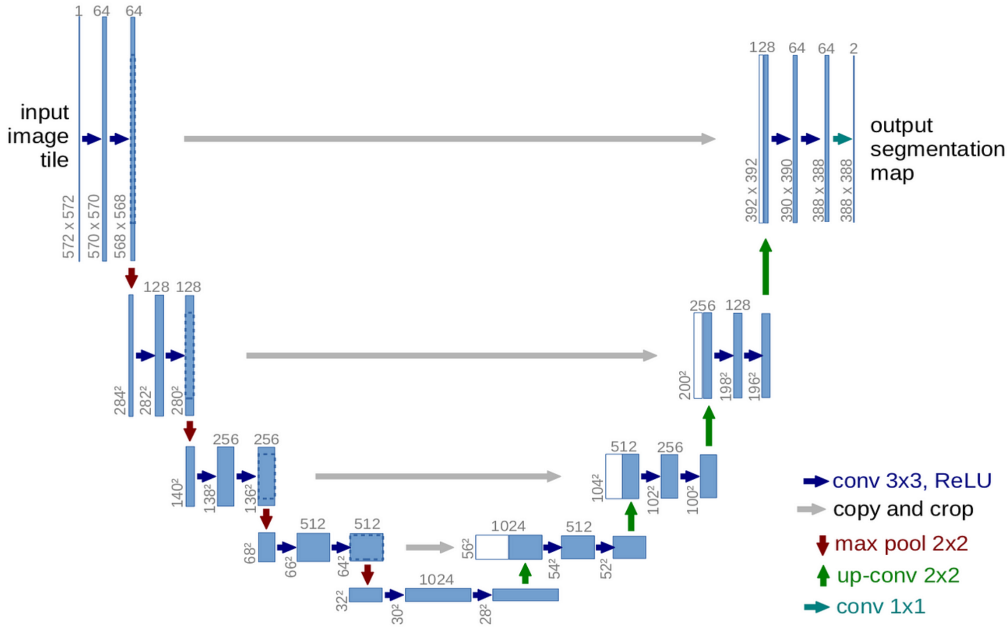


Fig. 1. UNet architecture (Ronneberger et al., 2015)

Text-to-image generation models have demonstrated exceptional sample quality using CLIP (Radford et al., 2021) to produce the image beforehand. Inspired by this, we use CLAP (Wu et al., 2022) to improve sound generation. A text encoder and an audio encoder are used to embed text. $E^{Text}$and an audio embedding E^{Audio}within the dimensions of the CLAP embedding. We have developed an audio encoder based on (Chen et al., 2022) and a CLIP-derived text coder (Radford et al. 2021). Furthermore, we choose a symmetric cross-entropy loss as the training objective.

We use a Variational Autoencoder (VAE) (Berahmand et al., 2024) to compress the mel-spectrogram into a compact latent space. Our Variational Autoencoder (VAE) consists of two elements: an encoder and a decoder, using stacked convolutional modules. Thus, the VAE encoder can preserve the spatial correlation between the mel spectrogram and the latent space. Each module is composed of ResNet blocks (Koonce, 2021) which include convolutional layers and residual connections. We use a reconstruction loss, an adversarial loss, and a Gaussian constraint loss in the training objective. During the sampling procedure, the decoder reconstructs the mel-spectrogram from the sound output produced by the LDMs.

The vocoder aims to transform the produced mel-spectrogram into a waveform. This type of vocoder is a major focus of research. The Griffin-Lim algorithm is a conventional signal processing method that is remarkably fast and easy to implement. However, Griffin-Lim produces low fidelity results when applied to mel spectrograms. WaveNet produces high quality results, but has a somewhat slow generation speed. In this work, MelGAN, a non-autoregressive technique, was used for waveform reconstruction because of its efficiency and high quality output. MelGAN has been widely used in the field of speech synthesis. However, many pre-trained MelGAN models have been constructed using speech or music datasets, making them inappropriate for ambient sound generation. Fig. 2illustrates the architectural framework of the SoundCrafter sound generation system.

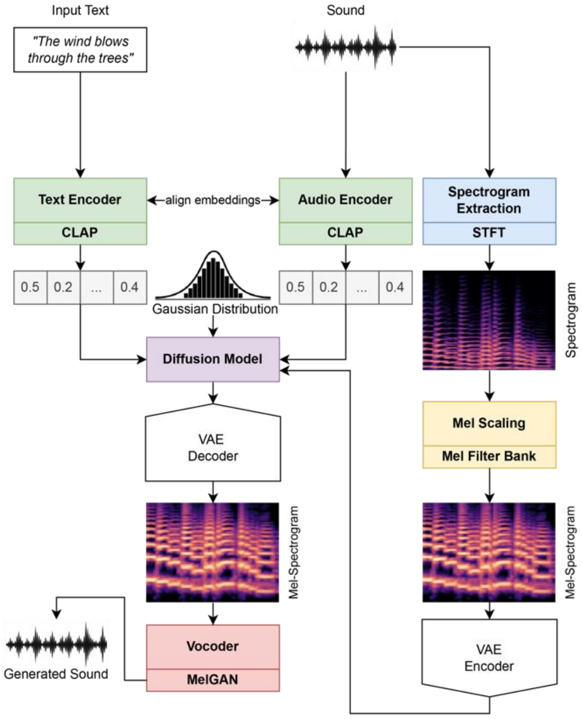


Fig. 2. Architectural overview of the SoundCrafter system for sound generation

4. TRAINING AND INFERENCE

This work uses the AudioCaps dataset (Kim et al., 2019). AudioCaps is the largest existing audio captioning dataset, containing approximately 90,000 audio clips sourced from AudioSet. AudioCaps has three subsets: training, validation, and test sets. The training, validation, and test sets contain 91,256, 494, and 957 audio clips, respectively. The AudioCaps collection contains 136.87 hours of audio samples, including various natural sounds, audio effects, music, and human activities. Each audio clip in the training set contains a single human-annotated caption, while each clip in the validation and test sets provides five captions. We use the AudioCaps training set to train our algorithms. We evaluate our approaches using the AudioCaps validation set. The majority of the data in AudioCaps consists of in-the-wild audio obtained from YouTube, so the audio quality is not guaranteed. The training was conducted using a rented online cloud computing instance equipped with state-of-the-art technologies, as described by (Al-kateeb & Abdullah, 2024a).

We evaluate the model using AudioCap. Each audio clip in AudioCap consists of five textual captions. We generate the evaluation set by randomly selecting one of these as the textual condition. The creators of AudioCap intentionally omit audio categorized as music in order to evaluate the model's effectiveness over a wider range of sounds; therefore, we randomly select 10% of the audio samples from AudioCap. The main metrics used for objective evaluation are the Fréchet distance, which quantifies the similarity between generated and target samples, and the Inception Score (IS), which assesses both the quality and diversity of the samples.

The model uses a latent diffusion technique to remove noise from compression mel-spectrograms trained on the AudioCaps dataset. A pre-trained CLAP audio encoder provides static semantic embeddings for conditioning, which are incorporated into the UNet architecture. Training uses 1000 diffusion steps with a linear beta schedule, culminating in 400,000 steps, and targets a spectrogram length of 1024 frames. SoundCrafter is designed for computational efficiency, enabling real-time or near real-time audio production. The model can produce a 10-second audio clip in approximately 1.2 seconds using a single NVIDIA RTX 3090 GPU. Optimization is performed using AdamW with a learning rate of 1e-4 and cosine annealing. Mixed precision and gradient checkpointing are used to save memory and computation. **Błąd! Nie można odnaleźć**

źródła odwołania. the correlation between sample quality and training progress. **Błąd! Nieprawidłowy odsyłacz do zakładki: wskazuje na nią samą.** the main training hyperparameters used.

Tab. 2. Training hyperparameters of SoundCrafter

Hyperparameters	Value
Dataset	AudioCaps
Diffusion Steps	1000
Noise Schedule	Linear beta schedule
Attention Heads	8
Dropout	0.1
Batch Size	16
Optimizer	AdamW
Learning Rate	1e-4
Training Steps	400,000 V
Spectrogram Length	1024 frames

5. RESULTS

We evaluate the effectiveness of the proposed SoundCrafter system using quantitative indicators as well as an analysis of training time. The SoundCrafter is evaluated on the AudioCaps benchmark using the Fréchet Distance (FD) and Inception Score (IS) metrics. The FD metric is derived from the Fréchet distance used in generative modeling and quantifies the similarity between the distributions of actual and produced audio features computed within the embedding space of a pre-trained audio classification model.

A reduced FD indicates that the produced audio samples are more similar in distribution to the real samples. The Inception Score (IS) extends this by evaluating the significance and diversity of the generated output, as determined by the entropy of the class predictions over the generated clips. Both metrics provide quantitative standards for model comparison and are widely used in the current text-to-sound generation literature. They are standard metrics for assessing the quality and variety of generated audio.

Our best-performing model achieves an FD of 23.47 and an IS of 7.57. It also uses 181 million parameters, as shown in **Błąd! Nieprawidłowy odsyłacz do zakładki: wskazuje na nią samą.**. These results are comparable to or better than previous methods such as AudioGen and DiffSound, but use fewer resources. Although the majority of AI models are now trained in the cloud, recent research highlights the need for energy-efficient methods (Al-Kateeb & Abdullah, 2024b).

Tab. 3. The comparison between SoundCrafter and other existing models

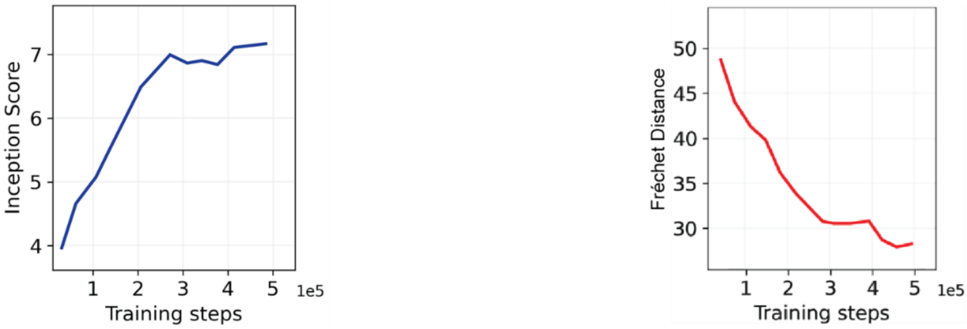
Method	FD ↓	IS ↑	#Params (M)
AudioGen	38	4.7	285
DiffSound	47	4.01	400
TANGO	24.52	7.27	866
Our	23.45	7.57	181

To analyze the training dynamics of the model, we plot the FD and IS scores over training steps. **Błąd! Nie można odnaleźć źródła odwołania.**shows a consistent decrease in the FD during the initial training phase, indicating an improved match between the produced and actual audio distributions.

These results demonstrate that SoundCrafter produces high-quality, semantically coherent audio samples while maintaining efficiency in terms of model size and data requirements, as shown in Fig. 4

Although SoundCrafter generally produces high-quality and semantically coherent audio, specific limitations and error patterns have been identified during the production process. A common error is overlapping or unclear auditory descriptions, as shown in Fig. 5. For example, cues such as "wind and waves together" may provide either a dominant sound or an unnatural blend, suggesting challenges in isolating and integrating multiple sources within a single latent diffusion pathway.

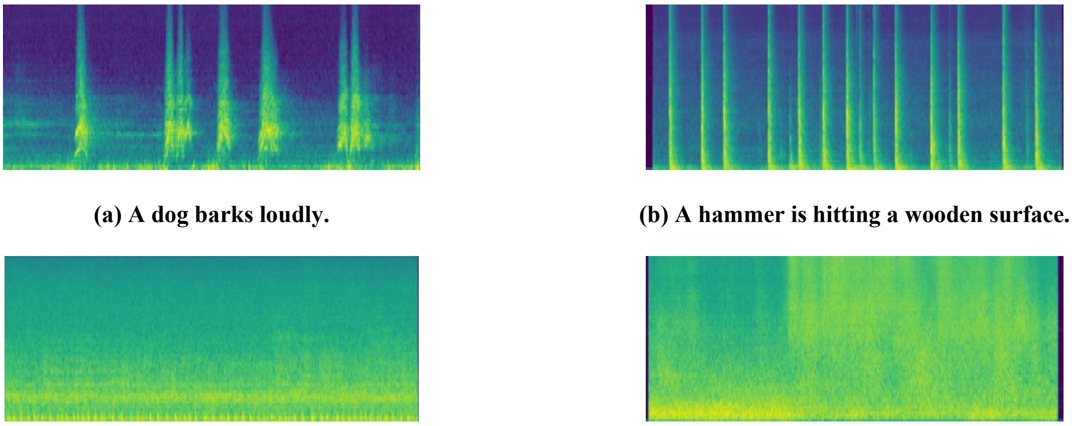
An additional concern arises in textual prompts with abrupt auditory events, such as "a door slams and a dog barks. In such cases, the temporal coherence of the produced audio can deteriorate, resulting in misaligned timing or missing events.



(a) Inception score

(b) Fréchet distance

Fig. 3. Evaluation of (a) IS and (b) FD during training of SourdCrafter



(a) A dog barks loudly.

(b) A hammer is hitting a wooden surface.

(c) Strong wind blowing through trees

(d) Waves crashing on a rocky shore

Fig. 4. Samples generated via our proposed model

(a) Wind and ocean waves at the same time

(b) Birds chirping while rain falls heavily

Fig. 5. Failure Cases of Our Proposed Model

6. CONCLUSIONS

SoundCrafter provides an effective and adaptable framework for text-to-sound production by combining a diffusion model with CLAP-based semantic conditioning. The text-audio pairs are used by SoundCrafter for training purposes, using the diffusion model with audio embeddings obtained from a pre-trained CLAP encoder. This design significantly reduces data requirements and improves the applicability of the model to large unlabeled audio samples. The architecture utilizes a UNet-based latent diffusion model that operates in the VAE-compressed mel-spectrogram domain, improving memory and computational efficiency. The model achieved a Fréchet distance (FD) of 23.47 and an inception score (IS) of 7.57, using a total of 181 million parameters.

Acknowledgments

The authors wish to express their appreciation and gratitude to the College of Computer Sciences and Mathematics, as well as the College of Education for Pure Science at the University of Mosul, for their invaluable support in the advancement of this study.

Conflicts of interest

The authors declare no relevant conflicts of interest pertaining to the content of this work.

REFERENCES

Al-Kateeb, Z. N., & Abdullah, D. B. (2024a). AdaBoost-powered cloud of things framework for low-latency, energy-efficient chronic kidney disease prediction. *Transactions on Emerging Telecommunications Technologies*, *35*(6). https://doi.org/10.1002/ett.5007

Al-Kateeb, Z. N., & Abdullah, D. B. (2024b). Unlocking the potential: Synergizing IoT, cloud computing, and big data for a bright future. *Iraqi Journal for Computer Science and Mathematics*, *5*(3). https://doi.org/10.52866/ijcsm.2024.05.03.001

Barahona-Ríos, A., & Collins, T. (2022). SpecSinGAN: Sound effect variation synthesis using single-image GANs. *ArXiv, abs/2110.07311*. https://doi.org/10.48550/arXiv.2110.07311

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., & Xu, Y. (2024). Autoencoders and their applications in machine learning: A survey. *Artificial Intelligence Review*, *57*, 28. https://doi.org/10.1007/s10462-023-10662-6

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., & Dubnov, S. (2022). HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. *ArXiv, abs/2202.00874*. https://doi.org/10.48550/arXiv.2202.00874

Cherep, M., Singh, N., & Shand, J. (2024). Creative text-to-audio generation via synthesizer programming. *ArXiv, abs/2405.18698*. https://doi.org/10.48550/arXiv.2406.00294

Ghosal, D., Majumder, N., Mehrish, A., & Poria, S. (2023). Text-to-audio generation using instruction-tuned LLM and latent diffusion model. *ArXiv, abs/2304.13731*. https://doi.org/10.48550/arXiv.2304.13731

Guan, W., Wang, K., Zhou, W., Wang, Y., Deng, F., Wang, H., Li, L., Hong, Q., & Qin, Y. (2024). LAFMA: A latent flow matching model for text-to-audio generation. *ArXiv, abs/2406.08203*. https://doi.org/10.48550/arXiv.2406.08203

Hasoon, S. O., & Al-Hashimi, M. M. (2022). Hybrid deep neural network and long short term memory network for predicting of sunspot time series. *International Journal of Mathematics and Computer Science*, *17*(3), 955–967.

Huang, J., Ren, Y., Huang, R., Yang, D., Ye, Z., Zhang, C., Liu, J., Yin, X., Ma, Z., & Zhao, Z. (2023). Make-An-Audio 2: Temporal-enhanced text-to-audio generation. *ArXiv, abs/2305.18474*. https://doi.org/10.48550/arXiv.2305.18474

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., & Zhao, Z. (2023). Make-An-Audio: Text-to-audio generation with prompt-enhanced diffusion models. *ArXiv, abs/2301.12661*. https://doi.org/10.48550/arXiv.2301.12661

Issa, R. J., & Al-Irhaym, Y. F. (2021). Audio source separation using supervised deep neural network. *Journal of Physics: Conference Series*, *1879*, 022077. https://doi.org/10.1088/1742-6596/1879/2/022077

Karchkhadze, T., Kavaki, H. S., Izadi, M. R., Irvin, B., Kegler, M., Hertz, A., Zhang, S., & Stamenovic, M. (2024). Latent CLAP loss for better Foley sound synthesis. *ArXiv, abs/2403.12182*. https://doi.org/10.48550/arXiv.2403.12182

Kim, C. D., Kim, B., Lee, H., & Kim, G. (2019). AudioCaps: Generating captions for audios in the wild. *2019 Conference of the North* (pp. 119–132). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1011

Koonce, B. (2021). *Convolutional neural networks with Swift for TensorFlow*. Apress.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., & Adi, Y. (2022). AudioGen: Textually guided audio generation. *ArXiv, abs/2209.15352*. https://doi.org/10.48550/arXiv.2209.15352

Liu, H., Huang, R., Liu, Y., Cao, H., Wang, J., Cheng, X., Zheng, S., & Zhao, Z. (2024). AudioLCM: Text-to-audio generation with latent consistency models. *ArXiv, abs/2406.00356*. https://doi.org/10.48550/arXiv.2406.00356

Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. (2017). FiLM: Visual reasoning with a general conditioning layer. *ArXiv, abs/1709.07871*. https://doi.org/10.48550/arXiv.1709.07871

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *ArXiv, abs/2103.00020*. https://doi.org/10.48550/arXiv.2103.00020

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *ArXiv, abs/1505.04597*. https://doi.org/10.48550/arXiv.1505.04597

Talal, R., & Anas, H. (2025). Prediction of drug risks consumption by using artificial intelligence techniques. *International Journal of Computing and Digital Systems*, *17*(1), 1–11.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Nezhurina, M., Berg-Kirkpatrick, T., & Dubnov, S. (2022). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ArXiv, abs/2211.06687*. https://doi.org/10.48550/arXiv.2211.06687

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., & Yu, D. (2022). Diffsound: Discrete diffusion model for text-to-sound generation. *ArXiv, abs/2207.09983*. https://doi.org/10.48550/arXiv.2207.09983

Yuan, Y., Liu, H., Liu, X., Kang, X., Wu, P., Plumbley, M. D., & Wang, W. (2023). Text-driven Foley sound generation with latent diffusion model. *ArXiv, abs/2306.10359*. https://doi.org/10.48550/arXiv.2306.10359

Zhang, C., Zhang, C., Zheng, S., Zhang, M., Qamar, M., Bae, S.-H., & Kweon, I. S. (2023). A survey on audio diffusion models: Text to speech synthesis and enhancement in generative AI. *ArXiv, abs/2303.13336*. https://doi.org/10.48550/arXiv.2303.13336