HARMONYLM: ADVANCING UNIFIED LARGE-SCALE LANGUAGE MODELING FOR AUDIO AND MUSIC GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The fields of sound generation and music generation have seen notable advancements with the development of specialized models tailored to each domain. However, these domains share commonalities, and the use of specialized models can lead to increased hardware resource requirements. On the other hand, recent breakthroughs in large language models, particularly in natural language processing, have showcased their ability to capture complex patterns and generate coherent and contextually relevant outputs in various tasks. Leveraging the success of these language models, we present HarmonyLM, a unified framework designed to synthesize sound and music from discrete representations. HarmonyLM adopts a unified perspective in modeling sound and music, discrete tokens are modeled from text descriptions using a decoder-only model, which are converted back to harmonious and consistent audio outputs. HarmonyLM offers significant advantages as a unified sound and music generation framework. (1) Model Scalability: the model we use in acoustic modeling a decoder-only transformer, which is free to scale up model size. (2) Data Scalability: the acoustic modeling and reconstructing audio models do not require any annotations, which accommodate different scales of data. Experimental results demonstrate the effectiveness of HarmonyLM, as it achieves superior audio quality compared to competitive baseline models.[1]

1 INTRODUCTION

Text-to-sound/music (Huang et al., 2023c; Copet et al., 2023; Huang et al., 2023a) aims to generate high-quality audios given text descriptions, which makes significant progress with the development of deep generative models like diffusion models (Ho & Salimans, 2021; Yang et al.). They are of importance for live dubbing, game sound effects, film and television soundtracks, and virtual reality(VR) and augmented reality(AR) applications. Despite significant success being made in developing specialized models for sound and music generation, separate models limit their ability to cope with more complex auditory scenarios. For example, battle scenes in games or movies require sound effects such as fighting sounds and explosions to simulate real scenes, and appropriate music adds to the audience's emotional experience and helps establish the atmosphere of the battle scene.

Large language models(LLM) (Touvron et al., 2023a;b; Zeng et al., 2022) have achieved great success in natural language processing(NLP), which proves that large language models can learn more complex paradigms. Recent advancements in self-supervised audio representation learning (Zeghidour et al., 2021; Défossez et al., 2022), sequential modeling (Yu et al., 2023; Brown et al., 2020), and audio synthesis (Borsos et al., 2022; Agostinelli et al., 2023) provide the conditions to develop a large-scale language model framework for music and sound generation. To make audio modeling more tractable, recent studies like SoundStream (Zeghidour et al., 2021) and Encodec (Défossez et al., 2022) proposed representing audio signals as multiple streams of discrete tokens representing the same signal. This allows both high-quality audio generation and effective audio modeling. Current large-scale audio generation systems (Kharitonov et al., 2023; Shen et al., 2023) leverage the codec models to generate discrete tokens and then predict them using language models, which have addressed intelligibility challenges in generated samples through large-scale training.

[1] Audio samples are available at `https://HarmonyLM.github.io`

While previous sound or music generation models have proven their effectiveness in their respective fields, most of them have been developed independently despite generating “audio” as a common objective. Thus, the methodologies developed for each application remain scattered in research fields, which is inefficient since we still need to optimize separated models for sound or music generation tasks.

In this work, we introduce HarmonyLM, a unified sound and music framework for synthesizing high-quality audio from discrete representations. HarmonyLM takes a unified perspective to model sound and music, which model discrete tokens from text description with a decoder-only language model, and then map them back to high-fidelity waveforms from discrete tokens. HarmonyLM demonstrates notable advantages as a unified audio synthesis framework: 1) Template universality: we define a novel prompt template to combine any texts and discrete units. 2) Model scalability: the acoustic backbone adopts a variant of transformer decoders, and thus the model capacity could be scaled up. Experimental results demonstrate that HarmonyLM achieves new state-of-the-art results. Both subjective and objective evaluation metrics show that HarmonyLM exhibits superior audio quality and style similarity compared with baseline models.

Our contributions can be summarized as follows:

- We propose a unified sound and music generation model called HarmonyLM, which effectively generates consistent and harmonious audio with a language model based on text prompts.
- We investigate the model and data scalability if large language models for both sound and music generation.
- Experimental results on two tasks demonstrate that HarmonyLM achieves state-of-the-art results in terms of subjective and objective metrics. We demonstrate our scalability by conducting experiments with different capacity language models and text encoders.

2 Related Works

2.1 Text-guided Sound/Music Generation

Text-to-sound and text-to-music generation exhibit commonalities. However, previous works (Huang et al., 2023c; Liu et al.; Dhariwal et al., 2020) often design task-specific inductive biases that limit their generalizability. One popular approach is the use of diffusion models, which naturally operate on continuous representations. Diffsound (Yang et al., 2022), for instance, leverages a pre-trained VQ-VAE (van den Oord et al., 2018) on mel-spectrograms to convert sound into discrete codes. These codes are then utilized by a diffusion model to generate audio outputs. Schneider et al. (2023); Huang et al. (2023b); Maina (2023) proposes the use of latent diffusion models for text-to-music while (Huang et al., 2023a; Liu et al.; Ghosal et al., 2023) for text-to-sound. Alternatively, music or sound samples can be represented as discrete codes using a hierarchical VQ-VAE, allowing the construction of language models on top of it. For example, AudioGen (Kreuk et al.) and MusicGen (Copet et al., 2023) encode raw waveform data into discrete codes and employ auto-regressive models to predict audio tokens based on text features. To advance towards a unified perspective of sound and music generation, this work presents a novel unified framework that capitalizes on a language model for both music and sound. By leveraging the power of language models, this framework seeks to facilitate the generation of coherent and expressive audio outputs in various applications.

2.2 Audio Representation

In recent years, there has been a significant increase in research dedicated to the compression of audio signals into continuous discrete representations. This approach aims to achieve efficient speech processing and high-fidelity audio coding. Notable advancements in this field include Wav2Vec (Baevski et al., 2020) and Hubert (Hsu et al., 2021), which have proposed quantization techniques using k-means to compress speech representations effectively. Drawing inspiration from vector quantization (VQ), SoundStream (Zeghidour et al., 2021) and Encodec (Défossez et al., 2022) have explored the utilization of hierarchical architectures to represent acoustic information. These models offer promising solutions for capturing and reconstructing audio signals with improved quality and fidelity. In a recent study by (Yang et al., 2023), they introduced a novel technique known as group-residual vector quantization (GRVQ), which demonstrates enhanced performance in audio

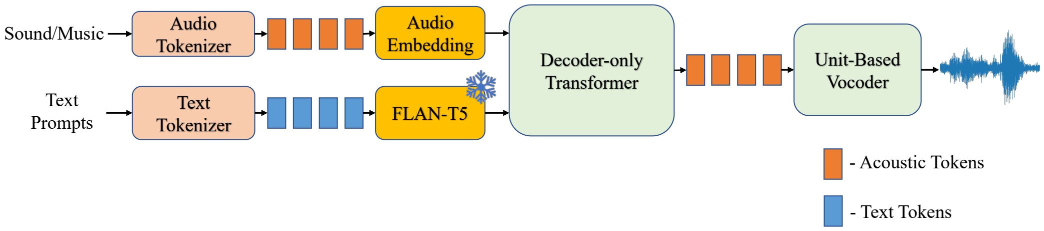

Figure 1: A high-level overview of HarmonyLM. We freeze the FLAN-T5 as our text encoder and use SoundStream as the audio tokenizer. The generated acoustic tokens by the autoregressive transformer are converted back to the raw sound/music with the unit-based vocoder(see Section 3.5).

coding. Based on this breakthrough, our work builds upon the techniques of SoundStream to extract discrete representations as acoustic tokens from sound and music. This enables us to achieve effective audio synthesis and processing, further enhancing the capabilities of our proposed framework.

2.3 Language Models

In the realm of audio synthesis, there has been a growing interest in modeling audio signals within a compact and discrete space. This approach allows for efficient and effective representation of audio using autoregressive transformers. Pioneering works such as AudioLM (Borsos et al., 2022) and MusicLM (Agostinelli et al., 2023) view audio synthesis as a language modeling task and employ a hierarchical structure of coarse-to-fine units. By leveraging this hierarchical approach, these models can generate high-quality audio outputs with fine-grained control. SpeechDLM (Nguyen et al., 2023) takes a similar approach but focuses specifically on spoken language modeling for dialogue. By utilizing the HuBERT representation, SpeechDLM introduces an end-to-end framework for generating realistic and context-aware speech. Furthermore, recent advancements in the field, such as MusicGen (Copet et al., 2023), propose operating over multiple streams of compressed discrete music representations. This allows for the synthesis of complex and expressive musical compositions. In this study, we propose a unified framework for sound and music generation. This framework adopts an autoregressive sequence-to-sequence (seq2seq) approach and utilizes discrete representations.

3 HarmonyLM

3.1 Overview

HarmonyLM stands as a unified framework that seamlessly combines language modeling and discrete representations to facilitate text-to-sound and text-to-music synthesis. As depicted in Figure 1. HarmonyLM employs autoregressive generation to transform audio samples into discrete tokens, which enables the model to capture the intricate acoustic characteristics and nuances embedded within the provided text prompts. By effectively mapping text prompts to corresponding acoustic tokens, HarmonyLM establishes a strong foundation for subsequent audio reconstruction. In the following, the model maps the acoustic token back to audio using a unit-based vocoder.

Once the backbone network of HarmonyLM has been trained, the model can be applied to two distinct domains: text-to-sound and text-to-music generation. Both tasks can efficiently utilize a shared language model with discrete representations. Flan-T5, a powerful text encoder, is employed in both domains to effectively encode the textual prompts. The proposed unified prompt template is then leveraged to seamlessly combine the text prompts with the acoustic tokens, facilitating a cohesive integration of textual and acoustic information. Following the combination of text prompts and acoustic tokens, acoustic tokens are generated and reconstructed into audio.

3.2 Discrete Audio Representation

Audio codec models such as SoundStream (Zeghidour et al., 2021) and Encodec (Défossez et al., 2022) have recently shown that encoder-decoder architecture excels at learning acoustic information in a self-supervised manner, where the representation can be used in a variety of generative tasks.

The acoustic codec model typically consists of an audio encoder, a residual vector-quantizer (RVQ), and an audio decoder: 1) The audio encoder E consists of several convolutional blocks with a total downsampling rate of 320 and generates continuous representations at every 20-ms frame in 16kHz. 2) The residual vector-quantizer Q produces discrete representations a_q with a codebook size of K_2, using a vector quantization layer (Vasuki & Vanathi, 2006). 3) The audio decoder G reconstructs the signal $\hat{y}$, from the compressed latent representation a_q.

3.3 Textual Representation

Text-guided synthesis models require powerful semantic text encoders to effectively capture the meaning of arbitrary natural language inputs. These encoders can be categorized into two major groups: 1) Contrastive pretraining. Similar to CLIP (Radford et al., 2021) pre-trained on image-text data, recent progress on contrastive language-audio pretraining (CLAP) (Elizalde et al.) brings audio and text descriptions into a joint space and demonstrates the outperformed zero-shot generalization to multiple downstream domains. 2) Large-scale language modeling (LLM). Saharia et al. and Kreuk et al. (2023) utilize language models (e.g., BERT (Devlin et al., 2018), T5 (Raffel et al.), FLAN-T5 (Chung et al., 2022)) for text-guided generation. Language models are trained on text-only corpus significantly larger than paired multi-modal data, thus being exposed to a rich distribution of text.

The FLAN-T5 models have undergone pre-training on a large-scale chain-of-thought (CoT) and instruction-based dataset. This extensive pre-training enables the models to learn new tasks effectively by leveraging in-context information and mimicking gradient descent through attention weights. This property is not present in older large models or contrastive models like T5 and CLAP. Considering the advantages mentioned above, we opt to utilize the pre-trained FLAN-T5-LARGE model as our text encoder, which is frozen following the common practice (Ghosal et al., 2023; Ramesh et al., 2022).

3.4 Acoustic Modeling

3.4.1 Unified Prompt Template

To facilitate a wide range of text-guided tasks, our model has been designed to handle various combinations of text prompts and discrete units, denoted as <Text, Acoustic>. Drawing inspiration from the patch embedding technique used in vision transformers (Dosovitskiy et al., 2021), we chunk the input sequences into P patches, where P corresponds to the number of quantization levels.

To represent the text prompts, we introduce a special token ([Continuous Token]) and replicate it P times. These replicated tokens are then concatenated with the corresponding acoustic tokens. Subsequently, the continuous tokens are replaced with text features after undergoing a patch embedding process. To differentiate between the two types of sequences (text and acoustic), we utilize two special tokens ([TYPE_START] and [TYPE_END]), where TYPE refers to the type of sequence being processed. The unified prompt template can be defined as follows:

$$x_p = [\mathrm{T5\_Start}]\ \mathrm{C}\ [\mathrm{T5\_End}]\ [\mathrm{Acoustic\_Start}]\ \mathrm{A}\ [\mathrm{Acoustic\_End}] \tag{1}$$

where C denotes continuous tokens and A is acoustic tokens, and all special tokens repeat P times.

3.4.2 Decoder-Only Transformer

After constructing a unified input template, a decoder-only transformer is adopted to map text prompts into acoustic tokens. To achieve a unified large-scale language model for audio and music generation and leverage the capabilities of language models for reconstructing acoustic tokens, we employ the following strategies:

- Model Scalability: The transformer model, renowned for its parallel computation and self-attention mechanisms, enables the handling of large volumes of input and output data. This design choice allows our model to efficiently scale up to larger capacities, accommodating the demands of more complex sound and music generation tasks.
- Data Scalability: To enhance data scalability, we utilized discrete representations for modeling sound and music. In the acoustic modeling stage, we employ a decoder-only model

to generate discrete units from textual descriptions. This discrete representation not only reduces storage and transmission costs but also improves training and inference efficiency. By adopting discrete representations, our framework eliminates the need for annotated data and achieves better handling of large-scale sound and music data, facilitating sophisticated modeling and generation of sound and music.

These approaches enable us to realize a unified framework for sound and music generation, leveraging the capabilities of language models for acoustic unit reconstruction. The model scalability empowers us to handle larger and more complex tasks, while the data scalability optimizes the handling of large-scale sound and music datasets, enhancing efficiency and accuracy in modeling and generation.

3.5 Reconstructing Audio

Upon completion of the training process, we can utilize language models to generate acoustic tokens based on the given text prompts. Subsequently, a unit-based vocoder is employed to synthesize the corresponding audio waveforms. It is worth noting that the acoustic codec used, such as SoundStream, leverages multiple quantization levels, typically 12, to enhance the quality of audio reconstruction. Thus, reducing the number of codebooks during the inference stage might result in a noticeable drop in perceptual quality.

To ensure that the quality of the generated audio waveforms remains uncompromised, we adopt a unit-based neural vocoder that is trained from scratch for waveform generation from acoustic units. This vocoder achieves high-quality audio reconstruction using only three quantization levels. Taking inspiration from the BigVGAN model (Lee et al., 2022), our synthesizer consists of a generator and a multi-resolution discriminator (MRD). The generator incorporates a set of look-up tables (LUT) that embed the discrete representations, along with a series of blocks. Each block comprises transposed convolutions and a residual block with dilated layers. The transposed convolutions upsample the encoded representation to match the input sample rate, while the dilated layers increase the receptive field.

3.6 Training and Inference Procedures

3.6.1 Training

During the training of language models, we compute the cross-entropy (CE) loss between the generated and target units. In the phase of audio reconstruction, we train the enhanced vocoder using a combination of different loss functions. These include the least-square adversarial loss, the feature matching loss, and the spectral regression loss on mel-spectrograms. These loss functions are carefully weighted and summed together, following the formulations and hyperparameters established by previous works such as Kong et al. (2020a); Lee et al. (2022)

The training data can be scaled up to a large-scale dataset, facilitating the modeling and generation of sound and music in a sophisticated manner. Moreover, HarmonyLM leverages the inherent scalability of transformer models. This scalability enables us to adapt the framework to different model sizes, accommodating the demands of more complex sound and music generation tasks.

3.6.2 Inference

HarmonyLM exhibits efficient advantages as a unified audio framework with discrete tokens and language modeling. Text-to-sound and text-to-music can be tackled by generating acoustic representations with text prompts: sound or music sample is tokenized into acoustic tokens, and unified prompt template are applied to combine text prompts and acoustic tokens.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 DATA

For training text-to-sound models, we use a combination of several datasets: AudioSet, BBC sound effects, Audiostock, AudioCaps-train, ESC-50, FSD50K, Free To Use Sounds, Sonniss Game Effects, WeSoundEffects, MACS, Epidemic Sound, UrbanSound8K, WavText5Ks, LibriSpeech, and Medley-solos-DB. For audios without natural language annotation, we apply the pseudo prompt enhancement to construct captions aligned well with the audio. Overall we have $\sim$3k hours with 1M audio-text pairs for training data. For evaluating text-to-sound models (Yang et al., 2022; Kreuk et al., 2023), the AudioCaption validation set is adopted as the standard benchmark, which contains 900 samples with five human-annotated captions in each audio clip. For a more challenging zero-shot scenario, we also provide results in the Clotho (Drossos et al.) validation set which contains multiple audio events.

For training text-to-music models, we use the LP-MusicCaps-MSD, which includes 500,000 pieces of music with 2.2M Caption. We evaluate the text-to-music models on the LP-MusicCaps-MC evaluation set. We convert the sampling rate of all audios to 16kHz and set the maximum length of the text to 77.

4.1.2 MODEL CONFIGURATIONS

For acoustic representation, we train the SoundStream model with 12 quantization levels, each with a codebook of size 1024 and the same downsampling rate of 320. We take 3 quantization levels as the acoustic tokens, representing each frame as a flat sequence of tokens from the first, second, and third quantization layers.

We use FLAN-T5-Large as our text encoder. Autoregressive acoustic modeling global models are 30-layer transformers with an attention dimension of 1920 and an FFN dimension of 7680. As for the unit-based vocoder, we use the modified V1 version of BigVGAN. A comprehensive table of hyperparameters is available in Appendix B.

4.1.3 TRAINING AND EVALUATION

During training, we train acoustic modeling transformers respectively for 50K steps using 8/80 NVIDIA A100 GPUs with a batch size of 10000 tokens for each GPU on the publicly-available *fairseq* framework (Ott et al., 2019). Adam optimizer is used with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$. For the acoustic modeling, we crop the waveform to a random length of up to 10 seconds. Reconstructing audio model is optimized with a segment size of 8192 and a learning rate of 1×10^{-4} until 500K steps using 4 NVIDIA A100 GPUs. During inference, we use batch size 1 of autoregressive decoding in acoustic modeling.

4.1.4 EVALUATION METRICS

To evaluate the performance of HarmonyLM on text-to-sound and text-to-music tasks, we follow the common practice of Huang et al. (2023a) and Copet et al. (2023). Specifically, to evaluate the text-to-sound models, we include both objective metrics Frechet distance (FD), Kullback–Leibler (KL) divergence, Frechet audio distance (FAD), and CLAP score, and subjective metrics including MOS-Q and MOS-F to assess the audio quality and the text-audio alignment faithfulness. The FAD is a reference-free perceptual metric that measures the distance between the generated and ground truth audio. FD measures the similarity between the generated and ground truth audio samples while CLAP score is a reference-free metric to measure audio-text alignment. As for subjective evaluation, we leverage the Amazon Mechanical Turk, a crowd-sourced platform, to perform the subjective evaluation on metrics including MOS-Q and MOS-F. We use similar evaluation metrics except for FD score, following Copet et al. (2023). More information regarding the evaluation process can be found in Appendix C.2

Model	Params	FD↓	KL↓	FAD↓	CLAP↑	MOS-Q↑	MOS-F↑
GroundTruth	-	-	-	-	0.671	86.47	84.31
AudioGen-S	285M	-	2.09	3.13	-	-	-
AudioGen-L	1B	-	1.69	1.82	-	-	-
Make-An-Audio	453M	18.32	1.61	2.66	0.593	69.54	65.45
AudioLDM-S	454M	29.48	1.97	2.43	-	-	-
AudioLDM-L	1.01B	23.31	1.59	1.96	0.605	70.91	67.41
TANGO	1.21B	26.13	1.37	1.87	**0.650**	74.10	72.76
Make-An-Audio 2	937M	15.25	1.32	1.80	0.645	78.31	75.63
HarmonyLM	1.3B	**12.5**	1.82	**1.49**	0.43	**80.01**	**76.22**

Table 1: The comparison between HarmonyLM and baseline models on the AudioCaps dataset. All the diffusion-based models run with 100 DDIM steps for a fair comparison. We used the model released by the authors on Huggingface to test scores.

4.2 Text-to-Sound

We compare the generated audio samples with several popular audio generation systems, including 1) GT, the ground-truth audio; 2) AudioGen (Kreuk et al., 2023); 3) Make-An-Audio (Huang et al., 2023c); 4) AudioLDM (Liu et al.); 5) TANGO (Ghosal et al., 2023); 6) Make-An-Audio 2 (Huang et al., 2023a); For easy comparison, the results are compiled and presented in Table 1, and we have the following observations: 1) For the quality of generated sounds, HarmonyLM has achieved higher FAD scores and FD scores than all baselines, both latent diffusion models(LDM) and language models, indicating that HarmonyLM can generate accessible sound of good quality as most previous LDM families. 2) Regarding subjective evaluation results, HarmonyLM has achieved the highest MOS-Q of 80.01 and MOS-F of 76.22 compared with the baseline models, demonstrating that HarmonyLM can better simulate real-world sounds. 3) For text-to-audio similarity, although HarmonyLM has a relatively lower CLAP score compared to the LDM benchmark, we argue this is mainly due to the difference caused by HarmonyLM not using classifier-free guidance and other data augmentation methods for text.

Model	Clotho-eval			AudioCaps-test		
	FD↓	KL↓	FAD↓	FD↓	KL↓	FAD↓
TANGO	32.1	2.59	3.61	31.76	2.04	10.53
AudioLDM-L	28.15	2.6	4.93	31.97	2.39	6.79
Make-An-Audio 2	22.79	**2.52**	2.76	13.78	**1.61**	2.33
HarmonyLM	**20.96**	2.80	**2.64**	**12.5**	1.82	**1.49**

Table 2: Comparison of HarmonyLM, Make-An-Audio 2, AudioLDM-L, and Tango on Clotho-eval and AudioCaps-test datasets.

Zero-shot evaluation. As illustrated in Table 2, when migrating to a more challenging scenario to Clotho in a zero-shot fashion, HarmonyLM still exhibits better FD and FAD scores and comparable KL scores, demonstrating HarmonyLM's effectiveness in constructing diverse object compositions for better generalization.

4.3 Text-to-Music

In this part, we compare the generated audio samples with other systems, including 1) GT, the ground-truth audio; 2) MusicGen (Copet et al., 2023); 3) MusicLM (Agostinelli et al., 2023); 4) Mousai (Schneider et al., 2023); 5) Riffusion (Forsgren & Martiros, 2022). The results are presented in Table 3, and we have the following observations: 1) Compared with diffusion-based models like Mousai and Riffusion, HarmonyLM performs better in both objective and subjective metrics, especially in the fact that the FAD score dropped from 7.5 to 2.95. This demonstrates the potential of language models in the field of music generation to synthesize harmonious and consistent pieces of music.

Model	KL↓	FAD↓	CLAP↑	MOS-Q↑	MOS-F↑
GroundTruth	-	-	0.40	88.42	90.34
Riffusion	2.06	14.8	0.19	79.31	74.20
Mousai	1.59	7.5	0.23	76.11	77.35
MusicLM	-	4.0	-	80.51	82.35
MusicGen	**1.23**	3.4	0.32	80.74	83.70
HarmonyLM	1.50	**2.95**	**0.34**	**82.32**	**85.28**

Table 3: The comparison between HarmonyLM and baseline models on the MusicCaps Evaluation set. We borrow the results from the MusicGen (Copet et al., 2023).

2) Compared with language models like MusicLM and MusicGen, HarmonyLM surpasses MusicLM in FAD score, improving from 4.0 to 2.95, and outperforms MusicGen in FAD and CLAP. The MOS-Q and MOS-F metrics have achieved the highest scores and exceed MusicLM by 1.81 and 2.93 respectively. This shows that HarmonyLM as an efficient unified model is able to generate high-quality music as the separate models do, which saves computing resources and time.

4.4 Analysis and Ablation Studies

To verify the effectiveness of several designs in HarmonyLM, including the scalability of language models and text representations, we conduct ablation studies and discuss the key findings as follows.

Model	Params	FD↓	KL↓	FAD↓	CLAP↑
Model Scalablity					
HarmonyLM-S	400M	14.1	2.36	1.79	0.39
HarmonyLM-M	850M	13.2	2.01	1.6	0.41
HarmonyLM-L	1.3B	12.5	1.82	1.49	0.43
HarmonyLM-XL	2.2B	12.0	1.73	1.43	0.44
Data Scalablity					
HarmonyLM-AC	1.3B	14.0	1.95	1.87	0.36
HarmonyLM-ALL	1.3B	12.0	1.73	1.43	0.44
Text Representions					
T5-Large	400M	18.35	3.15	2.73	0.30
FLAN-T5-Base	400M	14.65	2.42	1.90	0.37
FLAN-T5-Large	400M	14.1	2.36	1.79	0.39
FLAN-T5-XL	400M	13.85	2.13	1.71	0.41

Table 4: We conduct ablation studies on the text-to-sound task with different model sizes and text encoder dimensions. HarmonyLM-AC denotes training HarmonyLM on AudioCaps while Harmony-ALL on all datasets.

4.4.1 Impact of Architecture Scale

We investigate the impact of the model size on the text-to-sound task performance. We train four HarmonyLM models of different sizes(400M, 800M, 1.3B, 2.2B) using discrete representations. As illustrated in Table 4, we find that all metric scores improve substantially with model size, with a 2.1 reduction in FD and a 0.36 drop in FAD moving from 400M to 2.2B model. This proves that larger language models have better understanding and audio reconstruction capabilities while requiring more computing resources and training time.

4.4.2 Scaling the training data.

We investigate the impact of increasing the amount of training data on text-to-sound task performance. We run this analysis on HarmonyLM 1B checkpoint and train this model with an increasing amount of data: (1) The AudioCaption set only. (2) All the datasets described in Section A. We can draw a conclusion from the observation of Table 4. We find that training with increasing amounts of data yields a substantial improvement which demonstrates the advantage of HarmonyLM in terms of data scalablity.

4.4.3 TEXTUAL REPRESENTATION

We explore different text representations including T5-Large, FLAN-T5-Base, FLAN-T5-Large, and FLAN-T5-XL in the text-to-sound task. We freeze the weights of text encoders. For easy comparison, we present the results in Table 4 and have the following observations: 1) A larger text encoder produces higher scores across all metrics, which shows that higher feature dimensions could provide richer information representation and more powerful semantic modeling capabilities. By increasing feature dimensions, text encoders can better capture subtle features and semantic information in text input, thereby improving model performance. 2) FLAN-T5-Large outperforms T5-Large across all metrics especially FD score and FAD score, which demonstrates FLAN-T5 are able to learn a new task better with pre-training of FLAN-T5 models on a large-scale chain-of-thought- (CoT) and instruction-based dataset.

5 CONCLUSION

In this work, we proposed HarmonyLM, a unified model for synthesizing sound and music from discrete representations. HarmonyLM adopted a unified perspective in modeling sound and music: discrete tokens are modeled from text descriptions using a decoder-only model, which are mapped back to audio using codec models for harmonious and consistent audio generation. Experimental results demonstrated that HarmonyLM offered significant advantages as a unified sound and music generation framework: (1) Model Scalability: the model we use in acoustic modeling are decoder-only transformer, which is free to scale up model size. (2) Data Scalability: the acoustic modeling and reconstructing audio models do not require any annotations, which accommodate different scales of data. For future work, we will verify the effectiveness in more general scenarios such as audio generalization. The discussions on limitations and potential negative impacts are included in the Appendix.

6 LIMITATION

HarmonyLM utilizes auto-regressive models for the unified generation of sound and music, which inherently involves an iterative refinement process to achieve better results. The model's ability to produce high-quality outputs is directly influenced by factors such as sequence length and available computational resources. Longer sequence lengths typically require more computational power, and there may be a degradation in performance when training data is limited. One of our future directions is to develop lightweight and parallel models for accelerating sampling.

Furthermore, it is important to acknowledge that HarmonyLM currently lacks certain strategies such as classifier-free guidance and large language model data augmentation, which have been employed in previous latent diffusion models. These strategies play a crucial role in aligning generated audio and text, ensuring coherence and fidelity. In future work, we plan to incorporate these techniques into HarmonyLM to further improve its performance.

7 POTENTIAL NEGATIVE SOCIETAL IMPACTS

HarmonyLM lowers the requirements for harmony and consistent sound and music generation, which may cause unemployment for people with related occupations such as voice actor and composer. In addition, there is the potential for harm from non-consensual sound or music generation of fake media might be over-used than they expect.

REFERENCES

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2023.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. Audio retrieval with wavtext5k and clap training. *arXiv preprint arXiv:2209.14275*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music, 2020.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

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)*, pp. 736–740.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision.

Seth Forsgren and Hayk Martiros. Riffusion-stable diffusion for real-time music generation. *URL https://riffusion. com*, 2022.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned llm and latent diffusion model, 2023.

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 (icassp)*, pp. 131–135. IEEE, 2017.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 workshop on deep generative models and downstream applications*, 2021.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation, 2023a.

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. Noise2music: Text-conditioned music generation with diffusion models, 2023b.

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, 2023c.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*, 2023.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

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)*, pp. 119–132.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. of NeurIPS*, 2020a.

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, 2020b.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually Guided Audio Generation.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation, 2023.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

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.

Kinyugo Maina. Msanii: High fidelity music synthesis on a shoestring budget, 2023.

Irene Martín-Morató and Annamaria Mesaros. What is the ground truth? reliability of multi-annotator data for audio tagging. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 76–80.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11: 250–266, 2023.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

Karol J Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. 21(140):1–67.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding.

J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pp. 1041–1044.

Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Text-to-music generation with long-context latent diffusion, 2023.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.

A Vasuki and PT Vanathi. A review of vector quantization techniques. *IEEE Potentials*, 25(4):39–47, 2006.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete Diffusion Model for Text-to-sound Generation.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*, 2022.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023.

Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. Megabyte: Predicting million-byte sequences with multiscale transformers, 2023.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. Glm-130b: An open bilingual pre-trained model, 2022.

A Dataset

Dataset	Hours	Type	Source
Clotho	152	Caption	Drossos et al.
AudioCaps	109	Caption	Kim et al.
MACS	100	Caption	Martín-Morató & Mesaros
WavText5Ks	25	Caption	Deshmukh et al. (2022)
BBC sound effects	481	Caption	`https://sound-effects.bbcrewind.co.uk/`
Audiostock	43	Caption	`https://audiostock.net/se`
Filter AudioSet	2084	Label	Gemmeke et al.
ESC-50	3	Label	Piczak
FSD50K	108	Label	`https://annotator.freesound.org/fsd/`
Sonniss Game Effects	20	Label	`https://sonniss.com/gameaudiogdc/`
WeSoundEffects	11	Label	`https://wesoundeffects.com/`
Epidemic Sound	220	Label	`https://www.epidemicsound.com/`
UrbanSound8K	8	Label	Salamon et al.
LibriTTS	300	Language-free	Zen et al. (2019)
LP-MusicCaps-MSD	4283	Caption	Doh et al. (2023)
LP-MusicCaps-MC	15	Caption	Doh et al. (2023)

Table 5: Statistics for the combination of several datasets.

As shown in Table A, we collect a large-scale sound-text dataset consisting of 1M audio samples with a total duration of ∼3k hours. It contains audio of human activities, natural sounds, and audio effects, consisting of several data sources from publicly available websites. For audio with text descriptions, we download the parallel audio-text data. For audios without natural language annotation (or with labels), we discard the corresponding class label (if any) and apply the pseudo prompt enhancement to construct natural language descriptions aligned well with the audio. As for the music dataset, we collect a large-scale music-text dataset consisting of 0.5M music samples with 2.2M captions.

B Model Configurations

We list the model hyper-parameters of HarmonyLM in Table 6.

Hyperparameter		HarmonyLM
Decoder-Only Transformer	Transformer Layer	24
	Transformer Embed Dim	1920
	Transformer Attention Headers	16
	Transformer FFN Embed Dim	7680
	Decoder Dictionary Length	3081
	Number of Parameters	1.3B
Vocoder	Upsample Rates	[5, 4, 2, 2, 2, 2]
	Hop Size	320
	Upsample Kernel Sizes	[9, 8, 4, 4, 4, 4]
	Number of Parameters	121.6M
Total Number of Parameters		1.4B

Table 6: Hyperparameters of HarmonyLM.

We list our architecture scale in Table **??**

C Unit-based Vocoder

The generator of the unit-based vocoder is built from a set of look-up tables (LUT) that embed the discrete representation, and a series of blocks composed of transposed convolution and a residual block with dilated layers.

Params	Layer	Head	Hidden Dim
400M	26	24	1152
850M	30	24	1536
1.3B	30	32	1920
2.2B	30	40	2304

Table 7: Different model scale of HarmonyLM.

D EVALUATION

D.1 SUBJECTIVE EVALUATION

To assess the generation quality, we conduct MOS (Mean Opinion Score) tests regarding audio quality and text-audio faithfulness, respectively scoring MOS-Q and MOS-F.

For audio quality, the raters were explicitly instructed to “focus on examining the audio quality and naturalness.” The testers were presented with audio samples and asked to rate their subjective score (MOS-P) on a 20-100 Likert scale.

For text-audio faithfulness, human raters were shown the audio and its caption and asked to respond to the question, "Does the natural language description align with the audio faithfully?" They had to choose one of the options - "completely," "mostly," or "somewhat" on a 20-100 Likert scale.

Our crowd-sourced subjective evaluation tests were conducted via Amazon Mechanical Turk where participants were paid $8 hourly. A small subset of the generated audio samples used in the test can be found at `https://HarmonyLM.github.io/`.

D.2 OBJECTIVE EVALUATION

Fréchet Audio Distance (FAD) (Kilgour et al., 2018) is adapted from the Fréchet Inception Distance (FID) to the audio domain, it is a reference-free perceptual metric that measures the distance between the generated and ground truth audio distributions. FAD is used to evaluate the quality of generated audio.

KL divergence is measured at a paired sample level between the generated audio and the ground truth audio, it is computed using the label distribution and is averaged as the final result.

Fréchet Distance (FD) evaluates the similarity between the generated and ground truth audio distributions. FD, KL and IS are built upon an audio classifier, PANNs (Kong et al., 2020b), which takes the mel-spectrogram as model input. Differently, FAD uses VGGish (Hershey et al., 2017) as an audio classifier that takes raw audio waveform as model input.

CLAP score: adapted from the CLIP score (Hessel et al., 2021; Radford et al., 2021) to the audio domain and is a reference-free evaluation metric to measure audio-text alignment for this work that closely correlates with human perception.