MIDI-AudioLDM: MIDI-Conditional Text-to-Audio Synthesis Using ControlNet on AudioLDM

Autora: Laura Ibáñez Martínez
Director: José Manuel Cuadra Troncoso
Co-Director: Fabian-Robert Stöter

Trabajo Fin de Máster - Master's Thesis

Máster Universitario en Investigación en Inteligencia Artificial
Escuela Técnica Superior de Ingeniería Informática
Universidad Nacional de Educación a Distancia (UNED)

Septiembre 2023

Contents

Abstract

Text-to-audio systems have gained attention in recent months, achieving impressive results in general audio synthesis. However, they often lack fine-grained control over the musical output, as note-level adjustments cannot be determined by text. In this work, we present MIDI-AudioLDM, which implements MIDI conditioning into AudioLDM with the use of ControlNet. This enables MIDI-conditional text-to-audio synthesis, which adds up to AudioLDM's previous capacities, including direct text-to-audio synthesis as well as audio style transfer and inpainting. Like AudioLDM, the model uses contrastive language-audio pretraining (CLAP) latents and is trained on audio embeddings, while using text embeddings for inference. In contrast to unconditional audio synthesis, MIDI-AudioLDM offers detailed control over various musical aspects such as notes, genre, mood, and timbre, which makes it a more valuable tool for the music production process. A demo is available at https://huggingface.co/spaces/lauraibnz/midi-audioldm.

***Keywords*:** audio synthesis, MIDI conditioning, text-to-audio systems, AudioLDM, ControlNet

1 Introduction

Over the past years, text-to-image models like DALL·E (Ramesh et al., 2021), Midjourney (Holz et al., 2022) and Stable Diffusion (Rombach et al., 2022) have achieved remarkable success, and AI-generated images have made their way into popular culture. Similarly, the first text-to-audio models such as MusicLM (Agostinelli et al., 2023), AudioLDM (H. Liu, Chen, et al., 2023) or MusicGen Copet et al., 2023 have been introduced and remain a significant focus of research.

More particularly, AudioLDM (H. Liu, Chen, et al., 2023) is a latent diffusion model that employs contrastive language-audio pretraining (CLAP) latents to perform text-to-audio generation in the spectrogram domain. This model has demonstrated impressive results in general audio synthesis, as well as text-conditional audio-to-audio tasks such as style transfer and audio inpainting. However, this and similar works lack detailed control over the musical output, as only broader aspects like mood or timbre can be deter-

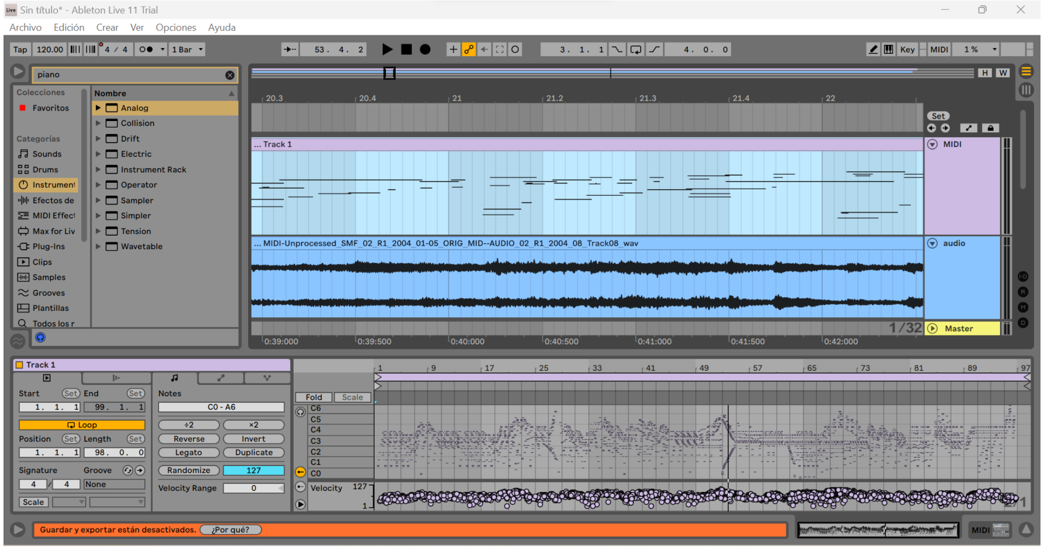


Figure 1: MIDI and audio from the same musical piece in Ableton Live.

mined by text. This makes them less practical for use in the music production process, where note-level control is often necessary.

Musical Instrument Digital Interface (MIDI) (MIDI Manufacturers Association, 1996), on the other hand, is a symbolic representation that "describes music using a notation containing the musical notes and timing, but not the sound or timbre of the actual sound" (DuBreuil, 2020). As in sheet music, MIDI contains information such as note pitch, loudness, onset and offset, but has no sound by itself and needs to be played by an instrument. This is often carried out with the use of Virtual Studio Technology (VST) (Steinberg, 1996) instruments, often integrated into Digital Audio Workstations (DAWs) like Ableton Live. An example of a MIDI track and an audio signal from the same musical piece is shown in Figure 1.

Thus, the hypothesis driving this study is that *the addition of MIDI conditioning to AudioLDM can provide further musical control over the generated output, including note-level pitch and loudness adjustment, which can serve as a valuable tool during the music production process.*

To achieve this goal, we incorporate MIDI conditioning into AudioLDM with the use of ControlNet (Zhang & Agrawala, 2023), a neural network architecture that enables the addition of conditional control to latent diffusion models. This model, primarily implemented into text-to-image models like Stable Diffusion (Rombach et al., 2022), locks a production-ready large diffusion model and reuses its layers as a backbone to learn a set of additional controls. In our study, we adapt ControlNet to the AudioLDM architecture, allowing it to accept a MIDI file as conditioning by converting it previously to a spectrogram-like image. The resulting model, named MIDI-AudioLDM, is implemented in Hugging Face's Diffusers library (von Platen et al., 2022). A demo[1] hosted in Hugging Face's Spaces is provided for experiments and further research.

The rest of this document is structured as follows: section 2 presents a review of relevant works in audio synthesis and text-to-audio systems; section 3 describes the proposed method, including the original AudioLDM architecture and the adaptation of the ControlNet structure; section 4 outlines the experimental setup, including the dataset selection, model implementation, training and evaluation process; section 5 presents results and a comparison with state-of-the-art models; section 6 discusses ethical and social implications of an application of this type; and section 7 draws conclusions and presents potential avenues for future research.

2 Related Work

This section provides an overview of the existing approaches to neural audio synthesis, including recent advances in conditional audio generation and state-of-the-art text-to-audio systems.

2.1 Neural Audio Synthesis

Audio synthesis is the process of generating sound using electronic hardware or software. This task has been traditionally carried out with the use of synthesizers, through methods such as additive (Weidenaar, 1995), subtractive (Moog, 1964), or FM (Chowning, 1977) synthesis. However, as highlighted by (Y. Wu, Manilow, et al., 2022), these often provide detailed expressive controls at the expense of realism. In recent years, an increasing number of studies have focused on audio synthesis using deep learning techniques, commonly referred to as "neural audio synthesis". While the terms "audio

[1] https://huggingface.co/spaces/lauraibnz/midi-audioldm

synthesis" and "audio generation" are often used interchangeably, "synthesis" typically denotes the process of transformation from another representation into audio, whereas "generation" tends to involve the creation of a musical piece. Similarly, the term "music generation" often implies the use of symbolic music representations such as MIDI.

Spectrograms, as described in (DuBreuil, 2020), have been a popular way of handling audio in machine learning, as they are compact and enable easier feature extraction in comparison to waveforms. A spectrogram is the result of applying the Fourier transform (Fourier, 1822) to overlapping frames from an audio signal and therefore decomposing it into its constituent frequencies. This visual representation contains time in the horizontal axis and frequency in the vertical axis, with color intensity corresponding to the audio amplitude. A mel-spectrogram, in turn, is a spectrogram where the frequencies have been converted to the mel scale. This scale is a perceptual scale of pitch in which equal distances sound equally distant to listeners. A common way of converting a spectrogram back into audio is with the use of neural vocoders.

As outlined by (DuBreuil, 2020), audio synthesis is a complex task, as it requires models to handle large numbers of samples per second while simultaneously keeping track of the broader structure. This is due to the sequential nature of the data, and the existence of long-range temporal dependencies such as musical patterns or phrases in speech. In the case of music generation, these systems are also expected to be expressive and realistic, and provide certain degrees of control in order to guide the audio synthesis process. To address these challenges, various strategies have been developed. These are described next.

Autoregressive waveform modeling. An initial approach to this problem is autoregressive waveform modeling . Some of the earliest and most popular models employ this technique, including WaveNet (Oord et al., 2016), SampleRNN (Mehri et al., 2017) and WaveRNN (Kalchbrenner et al., 2018). These models, which are based on recurrent neural networks (RNNs) or variations, predict waveforms one sample at a time, conditioning each sample on the previously generated ones. While they have demonstrated their ability to consistently capture the long-term structure of audio, both the training and inference process can be slow and computationally expensive, particularly for long and high-quality audio signals. Moreover, as shown in Figure 2, a waveform's shape does not perfectly correspond to how sound is perceived.

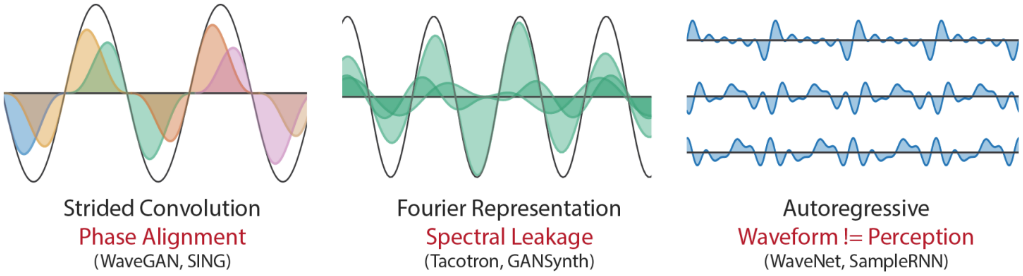


Figure 2: Challenges of neural audio synthesis, from the DDSP paper (Engel et al., 2019).

Adversarial training. A second approach involves the use of generative adversarial networks (GANs). This is the case of other early methods such as GANSynth (Engel et al., 2019) and WaveGAN (Donahue et al., 2019), which aimed to replicate the latest advances in the image generation field at the time. Some of these models, like WaveGAN, work by modeling waveforms directly in the time domain, while others such as GANSynth generate waveforms from their corresponding Fourier coefficients in the frequency domain. Both of these representations are general and can describe any waveform, but they can suffer from high bias (Engel et al., 2020). As audio signals are comprised of multiple frequency components with varying periods, the model is required to precisely learn to align waveforms and apply filters to cover all possible phase variations. In addition, Fourier-based models can suffer from spectral leakage, as multiple neighboring frequencies and phases are often combined to represent a single sinusoid when Fourier basis frequencies do not perfectly match the audio. These challenges are depicted in Figure 2.

Oscillators. A third approach consists of generating audio using oscillator models, such as vocoders and synthesizers, as seen in Differentiable Digital Signal Processing (DDSP) (Engel et al., 2020). These models combine classic signal processing elements with deep learning methods in a modular way, utilizing prior knowledge of how sound is naturally generated and perceived. As discussed in (Hayes et al., 2023), DDSP models are often praised for their interpretability and efficiency, as they are very fast to train. However, they currently rely on accurate estimates of the fundamental frequency, which complicates the rendering of polyphonic music and of non-harmonic sounds such as drums.

Variational autoencoders. More recently, models like Jukebox (Dhariwal et al., 2020) or RAVE (Caillon & Esling, 2021) have incorporated variational autoencoders (VAEs) to address the audio generation task. While Jukebox implements three separate VQ-VAE (van den Oord et al., 2018) models with different temporal resolutions for raw waveform modeling, Realtime Audio Variational autoEncoder (RAVE) adopts a two-stage training procedure, consisting of *representation learning* and *adversarial fine-tuning*. The model achieves real-time audio synthesis at a high sample rate by leveraging a multi-band decomposition of the raw waveform.

Diffusion models. Finally, models such as DiffWave (Z. Kong et al., 2021) and WaveGrad (N. Chen et al., 2020) have applied diffusion models to audio synthesis tasks. Diffusion probabilistic models, initially introduced in (Sohl-Dickstein et al., 2015) and popularized by (Ho et al., 2020), involve a forward diffusion process, which gradually adds noise to the training data through a Markov chain, followed by a denoising reverse process. After their success in the image generation field, diffusion models have proven promising results in waveform generation tasks, including speech synthesis and, more recently, text-to-audio generation.

2.2 Conditional Audio Generation

With regard to conditional audio generation, some models like Jukebox (Dhariwal et al., 2020) have enabled artist, genre or lyrics conditioning with the use of Transformers. Transformer models, first introduced in (Vaswani et al., 2017), learn context and meaning by establishing relationships in sequential data such as words. Therefore, they are often used for encoding speech and lyrics, as well as for predicting MIDI note events in symbolic music generation (C.-Z. A. Huang et al., 2018; Shih et al., 2022). In the case of Jukebox, this information is used to condition a series of Transformer models in order to generate the VQ-VAE codes which are then decoded into audio. This enables users to have more control over the style and content of the generated music.

On the other hand, and more related to the current research, some studies have enabled MIDI-to-audio synthesis (Cooper et al., 2022; Hawthorne et al., 2019, 2022; J. W. Kim et al., 2018; Manzelli et al., 2018; Y. Wu, Manilow, et al., 2022). The earliest of these models (Hawthorne et al., 2019; J. W. Kim et al., 2018; Manzelli et al., 2018) work by conditioning a WaveNet

(Oord et al., 2016) model, and therefore generate audio in an autoregressive manner. Alternatively, MIDI-DDSP (Y. Wu, Manilow, et al., 2022) has integrated MIDI conditioning into DDSP (Engel et al., 2020) models. While DDSP works by generating audio from frequency and loudness control parameters, MIDI-DDSP extracts these directly from MIDI notes by accurately predicting performance attributes. More recently, (Hawthorne et al., 2022) presents multi-instrument MIDI-to-audio synthesis using diffusion models. This model uses an encoder-decoder Transformer, where the decoder is trained as a diffusion model, to perform spectrogram to audio synthesis, followed by a GAN spectrogram inverter to acquire the resulting waveform. This work obtains promising results by prioritizing interactivity and generality in contrast to instrument-specific and less controllable audio synthesis methods.

2.3 Text-to-Audio Systems

Text-to-audio systems present another form of conditional audio generation. These models have emerged following the latest advances in the image generation field and, more specifically, in text-to-image synthesis (Holz et al., 2022; Ramesh et al., 2021; Rombach et al., 2022). The success of these models has been made possible by the existence of multi-modal models such as CLIP (Radford et al., 2021), which, combined with diffusion models, enable the generation of images with high adherence to specific text descriptions. CLIP models establish relationships between the visual and language domains by learning an embedding space common to both. During the last years, some works have extended CLIP to the audio domain, as is the case of AudioCLIP (Guzhov et al., 2021) or Wav2CLIP (H.-H. Wu et al., 2022). Similarly, some studies have presented Contrastive Language-Audio Pretraining (CLAP) models (Elizalde et al., 2022; Q. Huang et al., 2022; Y. Wu, Chen, et al., 2022), which learn audio concepts directly from natural language. A third approach involves the use of large-scale language models (LLM) (Saharia et al., 2022), which are trained on larger text-only corpus and are therefore exposed to a richer distribution of text.

One of the first text-to-audio works to appear in the context of diffusion models is Riffusion (Forsgren & Martiros, 2022). This model directly applies Stable Diffusion (Rombach et al., 2022) to spectrogram generation by fine-tuning it on a set of spectrogram images paired with text. This, combined with a vocoder model that converts spectrograms back into waveforms, enables audio generation directly from text prompts using a text-to-image diffusion model. Another early study is Diffsound (Yang et al., 2023), a discrete diffusion model for text-to-audio generation, also in the frequency domain. Diffsound extracts tokens from the input text using a CLIP model, and then feeds these to a spectrogram decoder. It is trained on AudioCaps (C. D. Kim et al., 2019), a large-scale dataset of audio and human-written text captions. AudioGen (Kreuk et al., 2023), on the other hand, learns an audio representation directly from the raw waveform, and uses an auto-regressive setting to allow the generation of high-quality samples.

Slightly later in the timeline, (Agostinelli et al., 2023) presents MusicLM, a hierarchical sequence-to-sequence model that can be conditioned on both text and melody, using joint music-text embeddings from MuLan (Q. Huang et al., 2022). The paper also introduces the MusicCaps[2] dataset of music-text pairs, with rich descriptions provided by human experts. The results provided demonstrate that the model outperforms previous works, and the generated audio shows great adherence to the input text descriptions. Unfortunately, neither MuLan nor MusicLM are available for public use. Also at this time, the models Make-An-Audio (R. Huang et al., 2023) and Moûsai (Schneider et al., 2023) are released, both of which follow a latent diffusion approach. While Make-An-Audio uses CLAP embeddings (Elizalde et al., 2022) and focuses on the problem of data scarcity in these tasks, Moûsai uses a large-scale language model (Saharia et al., 2022) and achieves successful results in real-time audio generation at a high sample rate.

Following these works, AudioLDM (H. Liu, Chen, et al., 2023) is introduced. This model uses CLAP embeddings (Y. Wu, Chen, et al., 2022) to train a latent diffusion model (LDM). Its structure is detailed in the following sections, and it has been chosen as an appropriate starting point for this study due to its successful results in text-to-audio synthesis, along with providing open-source code and pre-trained checkpoints. The base model has been trained on AudioSet (Gemmeke et al., 2017), AudioCaps (C. D. Kim et al., 2019), Freesound (Font Corbera et al., 2013) and BBC Sound Effect library (SFX) (British Broadcasting Corporation, 1997). Further fine-tuning using the MusicCaps dataset (Agostinelli et al., 2023) has been carried out

[2] https://www.kaggle.com/datasets/googleai/musiccaps

only for some of the checkpoints. In addition to text-to-audio synthesis, AudioLDM is capable of performing text-driven audio-to-audio tasks such as audio inpainting and style transfer.

In parallel to the current research, a few other relevant works in the field of text-to-audio synthesis have been presented. A notable example is MusicGen (Copet et al., 2023), a language model comprised of a single-stage Transformer that utilizes token interleaving patterns, eliminating the need for cascading multiple models. It can be conditioned on both text and melody, and has shown great efficiency in generating high-quality samples. Another case is AudioLDM 2 (H. Liu, Tian, et al., 2023), which introduces a novel approach to speech, music or sound effects generation with means of a general audio representation named “language of audio” (LOA).

On the other hand, some works like ControlNet (Zhang & Agrawala, 2023) have introduced conditional control in text-to-image diffusion models, enabling the use of additional inputs in order to guide the image generation process. Some of its applications, as demonstrated with Stable Diffusion (Rombach et al., 2022), are canny edge maps or human pose control. Due to its achievements and the architectural similarities between AudioLDM and Stable Diffusion, ControlNet has been chosen as an appropriate way to apply additional conditioning to this latent diffusion text-to-audio model.

3 Methods

In order to implement MIDI conditioning into AudioLDM with the use of ControlNet, a thorough study of how these models work is conducted. In the current section, a description of both of these models is provided, including their architecture and any relevant loss functions.

3.1 AudioLDM

AudioLDM (H. Liu, Chen, et al., 2023) is a latent diffusion model, very similar in structure to previous text-to-image models like Stable Diffusion (Rombach et al., 2022). Its architecture is depicted in Figure 3. While this diagram includes both text-to-audio (left) and audio-to-audio (right) tasks, the current work focuses on the first. An overview of the different blocks that conform AudioLDM is given next.

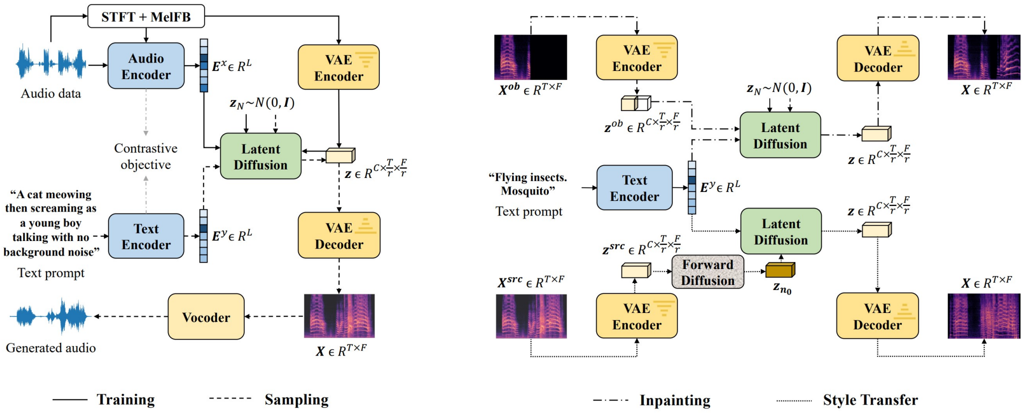


Figure 3: AudioLDM architecture for text-to-audio and audio-to-audio tasks, provided in the original paper (H. Liu, Chen, et al., 2023).

3.1.1 Contrastive Language-Audio Pretraining

One of main elements of AudioLDM is the use of contrastive language-audio pretraining (CLAP) latents, which allow to establish relationships between the audio and language domains. In AudioLDM, the CLAP model is trained first, using a dataset of audio-text pairs. The procedure followed by the author is based on (Y. Wu, Chen, et al., 2022), which suggests HTSAT (K. Chen et al., 2022) as the audio encoder and RoBERTa (Y. Liu et al., 2019) as the text encoder. The following symmetric cross-entropy loss is used as the training objective:

$$L = \frac{1}{2N} \sum_{i=1}^{N} (l_1 + l_2), \tag{1}$$

$$l_1 = log \frac{exp(\mathbf{E}_i^a \cdot \mathbf{E}_i^t / \tau)}{\sum_{j=1}^{N} exp(\mathbf{E}_i^a \cdot \mathbf{E}_j^t / \tau)} \tag{2}$$

$$l_2 = log \frac{exp(\mathbf{E}_i^t \cdot \mathbf{E}_i^a / \tau)}{\sum_{j=1}^{N} exp(\mathbf{E}_i^t \cdot \mathbf{E}_j^a / \tau)}, \tag{3}$$

where N is the batch size, $\mathbf{E}^a$ and $\mathbf{E}^t$ are the audio and text embeddings respectively, and τ is a learnable temperature parameter for scaling the loss. Once trained, the embeddings ($\mathbf{E}^a$, $\mathbf{E}^t$) can be used interchangeably, as a correlation between both domains has been established. Due to the scarcity of paired audio-text data, audio embeddings $\mathbf{E}^a$ are usually preferred for training the latent diffusion model, while using text embeddings $\mathbf{E}^t$ for inference instead. This allows training with a dataset of only-audio, while providing the user with more intuitive control over the generated audio output.

3.1.2 Variational Autoencoder

On the other hand, latent diffusion involves a prior encoding of the input features into a latent space, to reduce their size and therefore facilitate the diffusion process. This is carried out with the use of a variational autoencoder (VAE) (Kingma & Welling, 2022) in the case of AudioLDM, in contrast to the use of discrete representations in Stable Diffusion. The VAE in AudioLDM is conformed by an encoder and a decoder, both made up of stacked convolutional modules. Each module is formed by ResNet blocks (Q. Kong, Cao, et al., 2021), which are made up of convolutional layers and residual connections. A reconstruction loss, an adversarial loss and a gaussian constraint loss are used as training objectives. Once trained, mel-spectrograms that have been extracted from audio can be encoded into a continuous latent vector z, which contains information about the mean and variance of the VAE latent space.

3.1.3 Conditional Latent Diffusion Model

In diffusion models (Ho et al., 2020), the forward diffusion process consists of adding Gaussian noise at each time step t, according to a predefined noise schedule $0 < \beta_1 < ... < \beta_n < ...\beta_N < 1$, such that:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{4}$$

As shown by (Sohl-Dickstein et al., 2015), $\mathbf{x}_t$ can be sampled at any arbitrary noise level conditioned on $\mathbf{x}_0$:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{5}$$

with $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. The reweighted training objective is:

$$L_{DM} = \mathbb{E}_{\mathbf{x}_0, \epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\Big], \tag{6}$$

where:

$$\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2 = \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon, t)\|_2^2. \tag{7}$$

In conditional latent diffusion models (Rombach et al., 2022), the latent representation $\mathbf{z}_t$ is used instead, and the loss function becomes:

$$L_{LDM} = \mathbb{E}_{\mathbf{z}_0,\epsilon\sim\mathcal{N}(0,1),t,\mathbf{E}}\Big[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{E})\|_2^2\Big], \tag{8}$$

where $\mathbf{E}$ are the domain specific extracted embeddings during training, in our case $\mathbf{E}^a$. For inference, after sampling some random Gaussian noise $\mathbf{z}_T$, the following reverse denoising process conditioned on the text embedding $\mathbf{E}^t$ applies:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{E}^t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t, \mathbf{E}^t), \sigma_t^2\mathbf{I})). \tag{9}$$

The mean function and variance are parameterized as follows:

$$\begin{aligned}\mu_\theta(\mathbf{z}_t, t, \mathbf{E}^t) &= \frac{1}{\sqrt{\alpha_t}}(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t, t, \mathbf{E}^t)),\\ \sigma_t^2 &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t,\end{aligned} \tag{10}$$

where $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{E}^t)$ is the predicted noise and $\sigma_1^2 = \beta_1$.

3.1.4 UNet

As in Stable Diffusion (Rombach et al., 2022), a UNet (Ronneberger et al., 2015) backbone is used as the basic architecture of the latent diffusion model. The UNet from (Rombach et al., 2022) is shown in Figure 4. In AudioLDM, however, the cross-attention mechanism is not used, as the conditioning vector is only one-dimensional in this case. Instead, the time step is mapped into a one-dimensional embedding and concatenated with $\mathbf{E}$, and a feature-wise linear modulation layer (Perez et al., 2017) is used to merge this conditioning information with the feature map of the UNet convolution block. The UNet in AudioLDM has four encoder blocks, a middle block, and four decoder blocks. Setting a starting number of channel dimensions c_u, the encoder blocks have [c_u, $2c_u$, $3c_u$, $5c_u$] channel dimensions respectively. The middle block has $5c_u$ channel dimensions and the channel dimensions of the decoder blocks correspond to those of the reversed encoder blocks. An attention block consisting of two multi-head self-attention layers with a fully-connected layer in the middle is added in the last three encoder blocks and the first three decoder blocks.

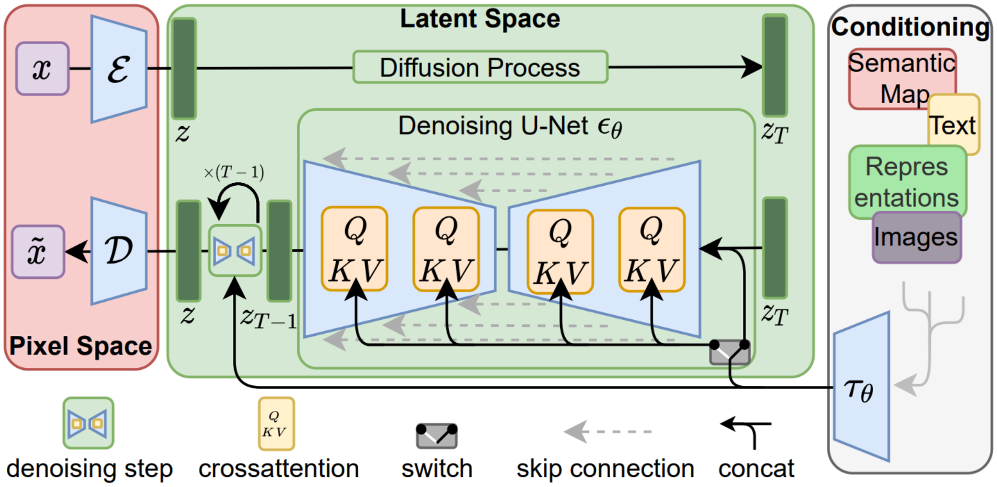


Figure 4: UNet architecture in latent diffusion models (Rombach et al., 2022).

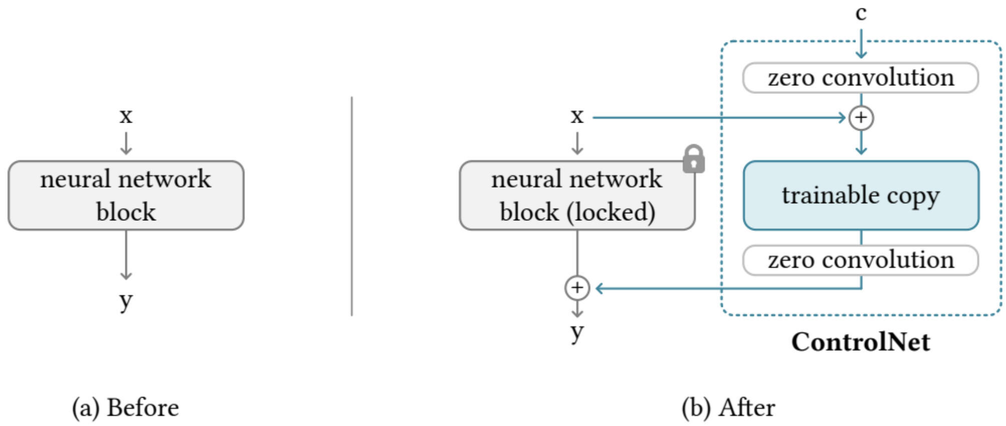
3.1.5 Vocoder

Neural vocoders are commonly employed in audio synthesis for converting mel-spectrograms back into audio. In the case of AudioLDM, the HiFi-Gan (J. Kong et al., 2020) vocoder is used, which has been widely utilized in speech synthesis tasks. Once $\mathbf{z}_0$ is obtained from the latent diffusion model, the VAE decoder transforms it into a mel-spectrogram. This mel-spectrogram is then fed to the vocoder to generate the corresponding waveform, representing the last step of the AudioLDM pipeline.

3.2 ControlNet

ControlNet (Zhang & Agrawala, 2023) is a neural network structure that provides conditional control in text-to-image latent diffusion models like Stable Diffusion (Rombach et al., 2022). It does so by manipulating the input conditions of neural network blocks, which refer to a set of neural layers that are frequently put together creating a unit, as for example a ResNet block (Q. Kong, Cao, et al., 2021). This neural neural network block with a set of parameters θ transforms input features $\mathbf{x}$ to a feature map $\mathbf{y}$, such that:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}; \theta), \tag{11}$$



(a) Before

(b) After

Figure 5: ControlNet applied to a neural network block (Zhang & Agrawala, 2023).

as can be visualized in Figure 5a. In ControlNet, the parameters in θ are locked and cloned into a trainable copy θ_c, to be trained with an external conditioning vector $\mathbf{c}$. Zero convolution layers are added, which consist of a 1×1 convolution layer where both weight and bias are initialized to zero. The resulting structure can be visualized in Figure 5b. This way, the output of the neural network block becomes:

$$\mathbf{y}_c = \mathcal{F}(\mathbf{x}, \theta) + \mathcal{Z}(\mathcal{F}(\mathbf{x} + \mathcal{Z}(\mathbf{c}; \theta_{z_1}); \theta_c); \theta_{z_2}), \tag{12}$$

where $\mathcal{Z}(\cdot;\cdot)$ denotes the zero convolution operation with parameters $\{\theta_{z_1}, \theta_{z_2}\}$. In the first training step, as both the weight and bias of each zero convolution layer are initialized to zeros, $\mathbf{y_c} = \mathbf{y}$. As the training progresses, the zero convolution layers progressively grow from zero to optimized parameters in a learned way.

The implementation of ControlNet into Stable Diffusion is shown in Figure 6. A trainable copy of the 4×3 encoder blocks and 1 middle block from the UNet is created. The encoder blocks are in resolutions (64×64, 32×32, 16×16, 8×8) respectively. The outputs of the trainable copies are added to the 12 skip-connections and 1 middle block of the UNet. This way, the loss function from Equation 8 becomes:

$$L_{LDM} = \mathbb{E}_{\mathbf{z}_0, \epsilon \sim \mathcal{N}(0,1), t, \mathbf{E}, \mathbf{c}} \Big[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{E}, \mathbf{c})\|_2^2 \Big], \tag{13}$$

where $\mathbf{c}$ are the task-specific extra conditions.

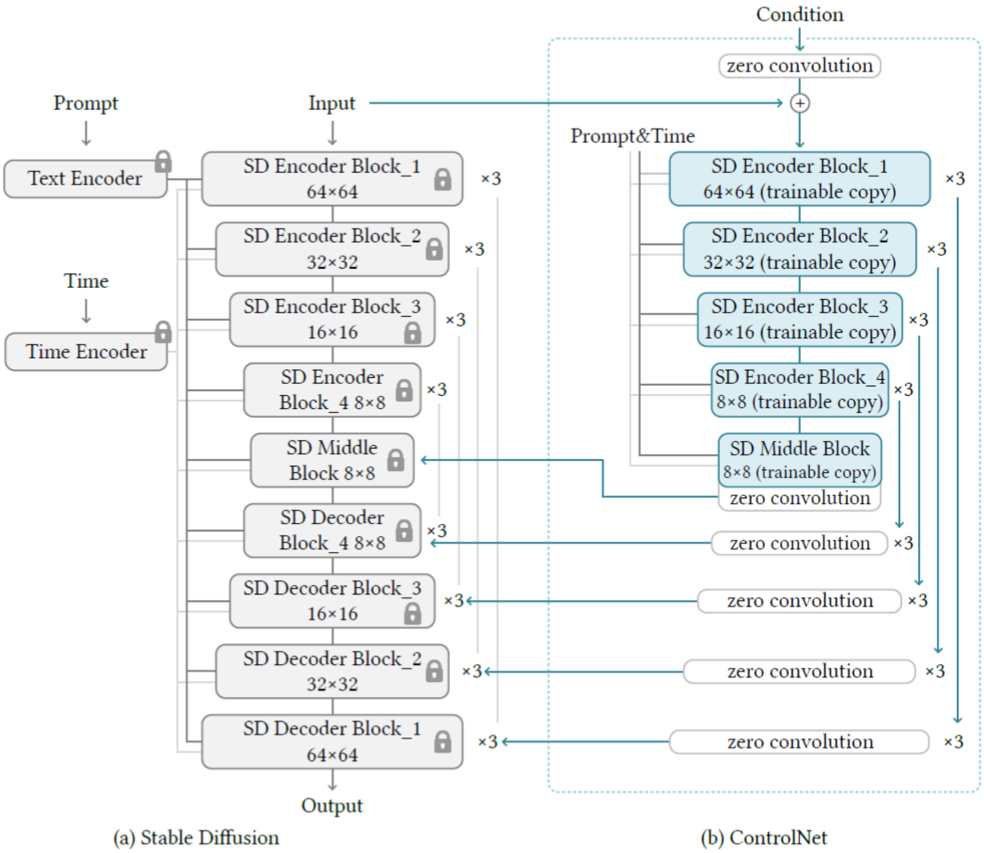


Figure 6: ControlNet (Zhang & Agrawala, 2023) implementation in Stable Diffusion (Rombach et al., 2022).

In the case of AudioLDM, the UNet is conformed by the same number of encoder, middle and decoder blocks as Stable Diffusion, although the channel dimensions vary based on an initial value c_u. While the Stable Diffusion encoder blocks have [320, 640, 1280, 1280] channel dimensions respectively, the UNet encoder blocks in AudioLDM with $c_u = 192$ have [192, 384, 576, 960] channel dimensions respectively. The input latent vectors in Stable Diffusion are of size 64×64, while those of AudioLDM are of size 128×128. Finally, AudioLDM does not use a cross-attention mechanism, but concatenates the conditioning vector to the time embedding instead. Therefore, the implementation of ControlNet into AudioLDM is considered feasible, after aplying the according modifications.

In respect to the input conditioning, a piano roll can be used, which consists of a visual representation of MIDI that resembles an audio spectrogram and which is often used for music generation tasks (Briot et al., 2019). The resulting model is named MIDI-AudioLDM.

4 Experimental Setup

In order to prove the hypothesis presented at the beginning of this work, a series of experiments are carried out. The current section provides a detailed description of the experimental setup, including dataset selection, model implementation, training configuration and evaluation process.

4.1 Datasets

For training MIDI-AudioLDM, a dataset of aligned MIDI and audio pairs is needed. A thourough study of the existing datasets of this kind is carried out, and the ones found most appropriate for this work are described next.

MAESTRO. The MIDI and Audio Edited for Synchronous TRacks and Organization (MAESTRO)[3] dataset, presented in (Hawthorne et al., 2019), has been widely used for a variety of tasks involving symbolic representations of music. The dataset contains around 200 hours of recordings and MIDI files from an international piano competition, with fine alignment ($\approx$ 3 ms) between note labels and audio waveforms. Some of its principal uses have been music transcription (Bittner et al., 2022; Gardner et al., 2022; Hawthorne et al., 2019; Q. Kong, Li, et al., 2021), music information retrieval (Zeng et al., 2021), and symbolic music generation (Dong et al., 2020). More in line with the current research, MAESTRO was used in its original paper (Hawthorne et al., 2019) and other works (Cooper et al., 2022; Hawthorne et al., 2022) to perform MIDI-to-audio synthesis.

URMP. The University of Rochester Multi-Modal Music Performance (URMP)[4] dataset, initially introduced in (Li et al., 2019), comprises a number of simple multi-instrument musical pieces, made up of aligned but separately recorded performances of each individual track. The corresponding MIDI files are provided, as well as ground-truth pitch annotations. As

[3] https://magenta.tensorflow.org/datasets/maestro

[4] https://labsites.rochester.edu/air/projects/URMP.html

in the case of MAESTRO, the dataset has been used for music transcription tasks (Gardner et al., 2022), but also for MIDI-conditional audio synthesis (Hawthorne et al., 2022; Y. Wu, Manilow, et al., 2022).

Slakh. The Synthesized Lakh (Slakh)[5] dataset, presented in (Manilow et al., 2019), contains 2100 multi-track audio, which are synthesized from individual MIDI tracks from the Lakh MIDI dataset (Raffel, 2016), using professional and sample-based virtual instruments. Every track in the dataset contains at least piano, bass, guitar, and drum stems, as well as the mixes created from all the stems. The Slakh dataset has been used to perform music transcription (Gardner et al., 2022), as well as MIDI-to-audio synthesis (Hawthorne et al., 2022).

A summary of these datasets, based on the one provided in Gardner et al., 2022, is shown in Table 1. This includes the hours of audio, the number of instruments and the alignment quality of the dataset annotations. In addition, it provides information about whether the audio is synthetic or not, and if it includes drums or mixes. As can be seen, each dataset has a set of advantages and disadvantages. Slakh contains a substantial amount of audio from a variety of instruments, but its audio is synthetic and might not be ideal for audio synthesis tasks. On the other hand, MAESTRO has good aligment quality and contains a fair amount of recordings, but its audio comes from a single instrument, piano. URMP, in contrast, contains a varied amount of instruments, but is a low-resource dataset and its alignment quality is worse. In the following sections, a number of experiments involving different combinations of these datasets are described.

DATASET	Hrs.	Num. Instr.	Alignment	Low-Resource	Synthetic	Drums	Mix
Slakh	969	35	good		✓	✓	✓
MAESTRO	199	1	good				
URMP	1	14	fair	✓			✓

Table 1: Summary of the selected datasets, based on (Gardner et al., 2022).

[5] http://www.slakh.com/

4.2 Implementation

The implementation of ControlNet into AudioLDM is carried out in Python using the Diffusers library (von Platen et al., 2022). The details of such implementation are described next.

Hugging Face's Diffusers library. In the first place, a thorough study of the official code repositories for AudioLDM[6] (H. Liu, Chen, et al., 2023) and ControlNet[7] (Zhang & Agrawala, 2023) is conducted. Since both models are available in Hugging Face's Diffusers[8] library (von Platen et al., 2022), it is found suitable to develop the project within this framework. The Diffusers library is described by its authors as the "go-to library for state-of-the-art pretrained diffusion models for generating images, audio, and even 3D structures of molecules"(von Platen et al., 2022). It includes diffusion pipelines for easy inference, as well as pretrained models for a variety of tasks. Diffusers uses PyTorch>= 1.4 as well as Hugging Face's Transformers[9] (Wolf et al., 2020) and Accelerate[10] (Gugger et al., 2022) libraries.

AudioLDM pretrained checkpoint. From the different AudioLDM checkpoints available in Hugging Face, `audioldm-m-full`[11] is chosen as an appropriate starting point, as it presents the best results in text-to-audio synthesis and is the only checkpoint that has been trained on audio CLAP embeddings instead of text. However, the audio encoder is not available in Diffusers, so it is converted from the original checkpoint first. For this, an existing conversion script from the Diffusers library is utilized and modified accordingly. A few months later, new AudioLDM checkpoints are released[12], including a version of `audioldm-m-full` but fine-tuned on the MusicCaps dataset (Agostinelli et al., 2023). This version is named `audioldm-m-text-ft`, as it uses CLAP text embeddings during fine-tuning. Although showing worse performance than the previous checkpoint for general audio synthesis tasks, the new checkpoint is expected to work best for purely musical output. The checkpoint is converted to the Diffusers format, and used for the training experiments described next.

[6] https://github.com/haoheliu/AudioLDM

[7] https://github.com/lllyasviel/ControlNet

[8] https://github.com/huggingface/diffusers

[9] https://github.com/huggingface/transformers

[10] https://github.com/huggingface/accelerate

[11] https://huggingface.co/cvssp/audioldm-m-full

[12] https://zenodo.org/record/7813012

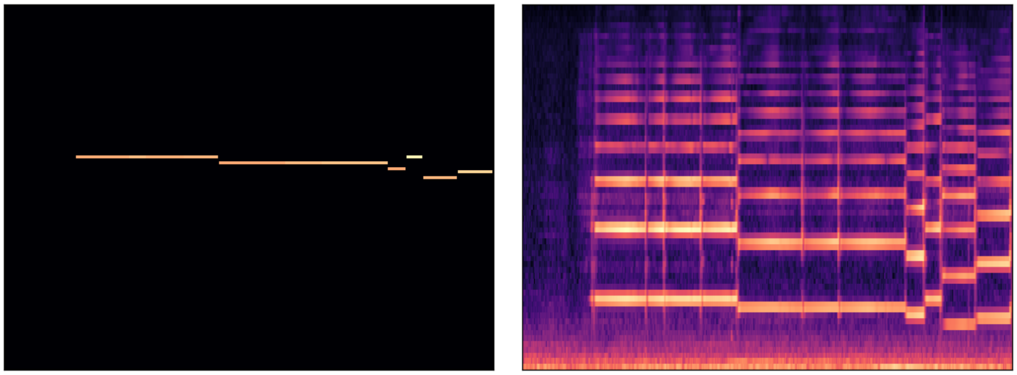
ControlNet architecture adaptation. Once the checkpoint is selected, the ControlNet architecture must be adapted to the architecture of AudioLDM. A ControlNet training script and a Hugging Face blog post[13] (Cuenca & Apolinário, 2023) are provided by its authors, which serves as a starting point for the current training experiments. The ControlNet code in Diffusers is first generalized, making it able to accept other latent diffusion model architectures different to Stable Diffusion. This includes the possibility of concatenating the conditioning vector to the time embedding, in contrast to the cross-attention mechanism used in Stable Diffusion (Rombach et al., 2022). The resulting code is available as a fork of the original Diffusers repository[14].

MIDI conditioning input. As mentioned in the ControlNet original paper (Zhang & Agrawala, 2023), the conditioning input must be adapted to fit the dimensions of the input features of the UNet, which are equal to the latent dimensions of the VAE. In the case of AudioLDM, these features are of size 128×128, in contrast to the input features of size 64×64 in Stable Diffusion. The ControlNet code provides its own VAE encoder, to adjust the size of the conditioning input accordingly. In the current work, MIDI is the desired conditioning input. As mentioned earlier, a piano roll can serve as an image-like representation of MIDI. This contains time information in the horizontal axis, and frequency in the vertical axis. An example of a piano roll and the corresponding mel-spectrogram of its audio pair from the MAESTRO dataset is shown in Figure 7. As can be seen, there is a correspondence between both images. Although the vertical axis from the piano roll is not in the mel scale, we can expect the VAE encoder to figure this out at the time of encoding this image into a latent representation of size 128×128. In MIDI-AudioLDM, the piano rolls are extracted from the MIDI files using the `pretty-midi`[15] library (Raffel & Ellis, 2014). The `pretty-midi` method for converting a MIDI file into a piano roll ignores drum tracks by default. A number of training experiments with the different datasets and excluding or including drum annotations are described next.

[13] https://huggingface.co/blog/train-your-controlnet

[14] https://github.com/lauraibnz/diffusers

[15] https://github.com/craffel/pretty-midi



(a) Piano roll

(b) Mel-spectrogram

Figure 7: Piano roll and mel-spectrogram corresponding to a MIDI-audio pair from the MAESTRO dataset.

4.3 Training

In the case of the MAESTRO and Slakh datasets, train, test and validation splits are provided. URMP, on the other hand, consists of a single split, so the available samples are divided randomly into train (90%) and validation (10%) splits. For the following training experiments, train splits are used for training and validation splits are used to monitor the validation loss. The training is run for 10k steps, where the steps per epoch depend on the size of the dataset, and checkpoints are saved every 500 steps. After training, the validation loss is observed, and the checkpoint with the lowest validation loss is selected to avoid overfitting.

The training configurations employed in order to compare the performance of MIDI-AudioLDM with respect to different combinations of datasets are shown in Table 2. For the training runs involving URMP, ground-truth pitch annotations are taken into account to improve the alignment quality of the dataset. In the case of multiple datasets with significant difference in size, weighting is applied to ensure that the amount of audio from each dataset is balanced. In this table, mixes refer only to those from the Slakh dataset, which contain drums and audio effects that are not present in the MIDI files. As mentioned earlier, drum annotations are not present in the piano roll by default. In some cases, these have been forced to appear, hoping that this helps the model learn to apply MIDI conditioning to more percussive sounds.

MODEL	MAESTRO	URMP	Slakh	Drum stems	Mixes	Drum annotations
MIDI-AudioLDM-M	✓					
MIDI-AudioLDM-U		✓				
MIDI-AudioLDM-M-U	✓	✓				
MIDI-AudioLDM-M-U-S-v1	✓	✓	✓			
MIDI-AudioLDM-M-U-S-v2	✓	✓	✓		✓	
MIDI-AudioLDM-M-U-S-v3	✓	✓	✓		✓	✓
MIDI-AudioLDM-M-U-S-v4	✓	✓	✓	✓	✓	✓

Table 2: Summary of the training configurations used.

4.4 Evaluation

Once the training is realized, a series of experiments are conducted in order to evaluate the effectiveness of MIDI-AudioLDM in a number of tasks. For this, the following objective and subjective metrics are used:

Fréchet Audio Distance (FAD). The Fréchet Audio Distance (Kilgour et al., 2019) measures the perceptual similarity between the distribution of the output samples generated by the model and the distribution of the target samples. FAD uses the VGGish (Hershey et al., 2017) model as a classifier. The lower the FAD score, the higher the similarity between the generated and the target audio. This measure is used as an evaluation metric in a number of audio synthesis tasks, including MIDI-to-audio synthesis (Hawthorne et al., 2022), and text-to-audio generation (Agostinelli et al., 2023; Copet et al., 2023; Kreuk et al., 2023; H. Liu, Chen, et al., 2023; H. Liu, Tian, et al., 2023). In our case, a PyTorch implementation[16] of FAD is used.

MT3 Transcription F1. This metric measures how well the model reproduces the notes and instruments from the MIDI data. As described in (Hawthorne et al., 2022), the generated samples are passed through the MT3 transcription model (Gardner et al., 2022), and the F1 score is computed using the "Full" metric from the MT3 paper. A higher F1 implies a higher correspondence between the target annotations and those extracted from the generated audio. In order to calculate this metric, the official MT3 implementation[17] is used.

[16] https://github.com/gudgud96/frechet-audio-distance

[17] https://github.com/magenta/mt3

Multi-resolution STFT. This metric compares ground-truth and generated audio in the frequency domain. Presented in (Yamamoto et al., 2020) as a multi-resolution short-time Fourier transform (STFT) loss function, this measure evaluates the reconstruction capacity of a generative audio model. A similar spectral loss function is employed in DDSP (Engel et al., 2020) and DDSP-based MIDI-to-audio models (Renault et al., 2022; Y. Wu, Manilow, et al., 2022). A PyTorch implementation[18] of this metric is used for the current experiments.

CLAP score. A CLAP score (Y. Wu, Chen, et al., 2022) is computed in order to measure the similarity between the text descriptions and the generated audio. A similar approach is used to evaluate a number of text-to-audio models (Copet et al., 2023; R. Huang et al., 2023; H. Liu, Tian, et al., 2023). As in MusicGen (Copet et al., 2023), the official code implementation[19] is used, along one of the official pretrained CLAP models[20] which is recommended for music.

In respect to subjective evaluation, a number of listening tests are carried out. For this, human participants are asked to listen to a series of 20-second audio recordings and rate them accordingly. The recordings correspond to target audio and audio synthesized by the different models, shown in a random order. In the case of MIDI-to-audio synthesis, a Mean Opinion Score (MOS) is utilized as in (Cooper et al., 2022; J. W. Kim et al., 2018), with a rating scale ranging from 1 (very bad) to 5 (very good). For MIDI-conditional text-to-audio synthesis, two measures are computed, the overall quality (OVL) and the relevance to the text input (REL). This represents a common standard in text-to-audio synthesis (Copet et al., 2023; Kreuk et al., 2023; H. Liu, Chen, et al., 2023; H. Liu, Tian, et al., 2023). In our case, OVL and REL are also rated on a 5-point Likert (Likert, 1932) scale. For all subjective metrics, both the mean and the 95% Confidence Interval are reported.

[18] https://github.com/csteinmetz1/auraloss

[19] https://github.com/LAION-AI/CLAP

[20] https://huggingface.co/lukewys/laion_clap/blob/main/music_audioset_epoch_15_esc_90.14.pt

5 Results

In the current section, the experiments realized in order to evaluate several audio synthesis tasks are described in more detail. The results obtained are presented and discussed, and a comparison between models is carried out.

5.1 MIDI-to-Audio Synthesis

The first experiment is aimed to test the capacity of MIDI-AudioLDM to synthesize a single-instrument MIDI track as precisely as possible. For this, the MAESTRO test split is utilized, as most existing MIDI-to-audio models are able to synthesize piano. Each MIDI file from the test split is fed to MIDI-AudioLDM, along with the simple caption “piano”. For comparison, a series of baseline models are utilized. This includes FluidSynth (Moebert et al., 2018), which is an open-source real-time software synthesizer available in Python, and the MIDI-to-audio model Spectrogram Diffusion (Hawthorne et al., 2022). For this, the FluidSynth synthesis method from `pretty-midi` is used, as well as the official Spectrogram Diffusion inference code with a pretrained checkpoint[21]. Unfortunately, DDSP-based MIDI-to-audio models (Renault et al., 2022; Y. Wu, Manilow, et al., 2022) are currently not available, as their official code repositories have not been mantained. The AudioLDM checkpoints `AudioLDM-m-full` and `AudioLDM-m-text-ft` are also employed as baselines, using “piano” as text input and without any kind of MIDI conditioning. The results from this experiment are presented in Table 3.

As can be seen from the table, MIDI-AudioLDM acquires the best results according to the FAD metric, which implies that the model is able to synthesize piano sound in a realistic way. This is confirmed by the MOS score, which shows that some of the MIDI-AudioLDM checkpoints, especially `MIDI-AudioLDM-M-U-S-v1`, are preferred over Spectrogram Diffusion and AudioLDM in a listening evaluation. The FAD results also prove that the fine-tuned AudioLDM model works better than the original checkpoint for purely musical output. On the other hand, the MIDI-AudioLDM results for F1 are significantly lower than in the case of previous MIDI-to-audio models, which means that the model is not synthesizing the original notes from the MIDI as accurately as expected. However, the value for the

[21] https://github.com/magenta/music-spectrogram-diffusion

MODEL	FAD↓	STFT↓	F1↑	MOS↑
GroundTruth	-	-	**0.39**	**4.22**$_{\pm 2.23}$
FluidSynth	4.69	2.85	0.37	**4.22**$_{\pm 2.23}$
Spectrogram Diffusion	11.17	**2.49**	0.16	2.67$_{\pm 1.46}$
AudioLDM-m-full	4.99	-	-	2.22$_{\pm 0.84}$
AudioLDM-m-text-ft	**2.61**	-	-	2.56$_{\pm 1.08}$
MIDI-AudioLDM-M	2.38	3.19	0.03	2.78$_{\pm 1.23}$
MIDI-AudioLDM-U	7.11	3.40	0.02	2.78$_{\pm 1.23}$
MIDI-AudioLDM-M-U	2.06	3.18	**0.05**	2.56$_{\pm 1.08}$
MIDI-AudioLDM-M-U-S-v1	2.01	3.24	0.03	**3.11**$_{\pm 1.46}$
MIDI-AudioLDM-M-U-S-v2	**1.95**	**3.04**	0.03	2.44$_{\pm 1.00}$
MIDI-AudioLDM-M-U-S-v3	2.14	3.25	0.02	2.89$_{\pm 1.31}$
MIDI-AudioLDM-M-U-S-v4	2.57	3.81	0.02	2.89$_{\pm 1.31}$

Table 3: MIDI-to-audio evaluation using MAESTRO test split.

STFT metric seems reasonable, so the synthesized audio must share certain musical similarity with the target audio. Therefore, it could be concluded that MIDI-AudioLDM provides a creative interpretation of the MIDI notes, rather than serve as an accurate MIDI-to-audio synthesis method. In respect to the different MIDI-AudioLDM configurations, for this experiment the best objective results are obtained from the model trained on the MAESTRO, URMP and Slakh datasets, excluding drums and including mixes, and without drum annotations. However, the model trained on all datasets, but excluding both drums and mixes, shows the best results in a subjective evaluation. In general terms, we can say that MIDI-AudioLDM outperforms the original AudioLDM model, while enabling MIDI conditioning, which the model interprets in a creative way.

5.2 MIDI-Conditional Text-to-Audio Synthesis

Secondly, an experiment is constructed to assess MIDI-AudioLDM's ability to synthesize a coherent mixture, containing drums and a variety of instruments, from the MIDI file of a single instrument. For this, the Slakh test split is used, as it contains complex mixes as well as individual instrument stems.

A Whisper audio captioning model[22] (Kadlcík et al., 2023) is utilized to acquire descriptive captions from the mix files available in the dataset. Each of these captions is then fed to MIDI-AudioLDM, along with the MIDI file of one of the individual instruments. The individual instrument stem to use in each case is chosen randomly, excluding drums and instruments not present in the MIDI files. For comparison purposes, the selected MIDI file is synthesized with the use of FluidSynth (Moebert et al., 2018), and then fed to the MusicGen (Copet et al., 2023) model with melody conditioning[23], along with the extracted caption from the mixture. In addition, the AudioLDM checkpoints are utilized as baselines, using the extracted captions as text input and with no MIDI or melody conditioning. The results from this second experiment are shown in Table 4.

MODEL	FAD↓	STFT↓	CLAP↑	OVL↑	REL↑
GroundTruth	-	-	**0.26**	**4.22**$_{\pm 2.23}$	3.44$_{\pm 1.69}$
FluidSynth+MusicGen	**2.28**	**3.37**	0.18	3.89$_{\pm 2.01}$	**3.75**$_{\pm 2.04}$
AudioLDM-m-full	7.76	-	0.23	1.86$_{\pm 0.69}$	2.57$_{\pm 1.26}$
AudioLDM-m-text-ft	11.58	-	0.17	1.67$_{\pm 0.58}$	1.57$_{\pm 0.46}$
MIDI-AudioLDM-M	12.05	3.84	0.21	2.00$_{\pm 0.80}$	2.00$_{\pm 0.80}$
MIDI-AudioLDM-U	11.19	3.46	0.19	2.00$_{\pm 0.80}$	2.13$_{\pm 0.84}$
MIDI-AudioLDM-M-U	13.01	3.58	0.20	1.83$_{\pm 0.73}$	1.71$_{\pm 0.57}$
MIDI-AudioLDM-M-U-S-v1	6.49	3.17	**0.26**	**2.29**$_{\pm 1.03}$	**2.57**$_{\pm 1.26}$
MIDI-AudioLDM-M-U-S-v2	6.36	**3.09**	0.24	2.13$_{\pm 0.84}$	2.50$_{\pm 1.11}$
MIDI-AudioLDM-M-U-S-v3	7.77	3.26	0.23	1.71$_{\pm 0.57}$	1.86$_{\pm 0.69}$
MIDI-AudioLDM-M-U-S-v4	**5.94**	3.70	0.25	2.00$_{\pm 0.88}$	2.22$_{\pm 0.84}$

Table 4: MIDI-conditional text-to-audio evaluation using Slakh test split.

The results presented suggest that the current approach is not sufficient for the generation of complex mixtures from a single instrument stem. Even though the CLAP metric reports the best results for MIDI-AudioLDM, the subjective measure REL proves that the adherence to the provided text descriptions is much lower than in the case of MusicGen. Moreover, the FAD measure is notably higher than in the previous experiment, proving lower similarity between the generated and the target audio in terms of quality. In

[22] https://huggingface.co/MU-NLPC/whisper-large-v2-audio-captioning

[23] https://huggingface.co/facebook/musicgen-melody

the same line, the subjective measure MOS demonstrates worse quality than in the case of MusicGen. On the other hand, the STFT metric reports good results, which suggests certain spectral similarity between the synthesized audio and the target mixes. This could mean that some musical aspects are preserved. In respect to the MIDI-AudioLDM model variations, in this experiment those that have not been trained on the Slakh dataset present the worse results. This seems logical, as they have not learnt from complex mixtures and tracks which contain drums. However, the MIDI-AudioLDM model that shows the best results in a subjective evaluation is `AudioLDM-M-U-S-v1`, which has not learnt from the Slakh drums or mixtures. Finally, all the MIDI-AudioLDM models that have learnt on the three datasets show better results than AudioLDM, which implies that the addition of MIDI conditioning has added value to the model, as well as providing a new feature.

6 Ethical and Social Implications

The development of large-scale machine learning models should always involve the consideration of a series of ethical and social implications. This is especially crucial in the case of generative models, as they can often raise concerns related to ownership and copyright issues, as well as to the potential generation of deep fakes.

In the image generation field, the recent surge of highly realistic generative models has posed a number of unprecedented ethical questions. This is especially evident following the emergence and popularization of text-to-image models like Stable Diffusion (Rombach et al., 2022). As discussed in Heikkilä, 2022, some of the datasets used for training such models have been created by scraping images from the internet, often without obtaining permissions and providing proper attribution to artists. This can lead to systems being able to generate artworks in the style of artists and individuals that are well represented in these datasets, even without their explicit consent. In respond to these concerns, the website “Have I Been Trained” (Herndon & Dryhurst, 2023), allows individuals to check if their works or identity are included in some of the largest public text-to-image datasets.

As described in (Barnett, 2023), advancements in the audio domain usually follow those from the image field, and so do the ethical implications of these models. The introduction of generative audio models like Jukebox

(Dhariwal et al., 2020), which can be conditioned on specific artists or genres, has lead to debates on copyright infringement. This is the case of the music duo DADABOTS, who created a Britney Spears cover of Frank Sinatra (Robitzski, 2020) with the use of this tool. The song was removed from YouTube over copyright claims, although finally restored after a legal argumentation from the artists. On the other hand, the musician Holly Herndon has introduced Holly+ (Herndon & Never Before Heard Sounds, 2021), an AI music tool that can perform style transfer on her own voice. As described by the artist, the tool aims to embrace the utility of deepfake technology rather than to be disempowered by it.

In respect to text-to-audio models, as pointed out by (Barnett, 2023), only a few authors have discussed the possible negative impact of these kinds of models. This is the case of MusicLM (Agostinelli et al., 2023), which highlights the risk of copyright infringement and raises concern about the potential biases present in the training data. These can lead to an under-representation of certain cultures, as well as to issues related to cultural appropriation. In Make-An-Audio (R. Huang et al., 2023), the authors underline the potential risk of misinformation caused by deep fakes, as well as a possible increase in unemployment for related occupations like sound engineering. More recently, MusicGen (Copet et al., 2023) states to have ensured that the training data was covered by legal agreements, and comments on the bias present in these datasets towards Western-style music. Moreover, the authors hope that more advanced controls, such as melody conditioning, can become useful to both amateur and professional musicians.

In the case of MIDI-AudioLDM, we acknowledge that there is a strong bias present in the training data towards music from the Western tradition. As discussed in (Gardner et al., 2022), datasets from other musical traditions are currently low-resource and represent an important area for future work. On the other hand, the current research aims to democratize the music creation process. While DAWs and VST intruments are often expensive and remain unaffordable for a number of people, an open-source tool for text-driven MIDI-to-audio synthesis can be a valuable resource for anyone with access to the internet. Moreover, an interface like Hugging Face can facilitate the music creation process, as very little coding experience is required to utilize the organization's Spaces available in their website.

7 Conclusions

In conclusion, MIDI-AudioLDM introduces a novel task: MIDI-conditional text-to-audio synthesis, for which no prior methods currently exist. Simultaneously, it offers an initial approach to the problem by incorporating MIDI conditioning into AudioLDM. The resulting architecture can perform MIDI-to-audio synthesis by creatively interpreting an input note sequence based on the given text description. This serves as a valuable tool for music production, where preserving aspects like musical key is often crucial during the audio synthesis process.

Furthermore, MIDI-AudioLDM enhances the capabilities of its baseline model AudioLDM. In addition to offering improved quality for purely musical output in direct text-to-audio synthesis, it introduces a new feature that enables conditioning based on MIDI note sequences. This note-level control can complement mood and timbre conditioning provided by text descriptions. On the other hand, MIDI-AudioLDM outperforms previous MIDI-to-audio models like Spectrogram Diffusion in terms of audio quality, and is able to synthesize sounds from certain instruments more realistically.

However, this initial approach remains insufficient and opens up various avenues for possible future work. As text-to-audio synthesis continues to be a focal point of research, several highly successful models have been released during the development of this work. Some examples are MusicGen or AudioLDM 2, both of which use a language modeling approach to achieve text-to-audio synthesis in the music domain, showing a substantial improvement in respect to previous works of this kind. For this reason, it would seem appropriate to adapt the concept behind MIDI-AudioLDM to a range of new models and architectures, in order to achieve optimal results.

Similarly, as described previously, the generation of audio in the frequency domain can lead to a number of inefficiencies. While this approach is often an immediate response to recent advancements that have proven successful in the image generation domain, more suitable model architectures and configurations are usually introduced which take more into account how sound is naturally generated and perceived. For this reason, new strategies to address the task of text-to-audio synthesis are expected to appear in the following months, which could be adapted to incorporate MIDI conditioning as in MIDI-AudioLDM.

Finally, in the current work, MIDI conditioning has been implemented with the use of ControlNet, without conducting a thorough study of the existing alternatives that provide additional control to a latent diffusion model. As demonstrated by the results, this approach has limitations and may prove inadequate for cases that require fine-grained note-level control in the MIDI-to-audio synthesis process. In the same line, using a continuous representation to encode MIDI is not optimal, and a number of alternatives should be considered. As proved by several generative music models in the symbolic domain, learning MIDI representations with the use of language models such as Transformers can result successful, as both share similarities in terms of syntax rules and long-term structure.

For these reasons, MIDI-AudioLDM is presented as a valuable contribution in the field of neural audio synthesis. It takes a first step towards solving the challenge of incorporating MIDI conditioning into a text-to-audio model, and sets the stage for further advancements and refinements in this area of research.

References

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). *Musiclm: Generating music from text* (tech. rep.). https://doi.org/10.48550/arXiv.2301.11325

Barnett, J. (2023). The ethical implications of generative audio models: A systematic literature review. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 146–161. https://doi.org/10.1145/3600211.3604686

Bittner, R. M., Bosch, J. J., Rubinstein, D., Meseguer-Brocal, G., & Ewert, S. (2022). *A Lightweight Instrument-Agnostic Model for Polyphonic Note Transcription and Multipitch Estimation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2203.09893

Briot, J.-P., Hadjeres, G., & Pachet, F.-D. (2019). *Deep Learning Techniques for Music Generation* (1st). Springer Publishing Company, Incorporated. https://link.springer.com/book/10.1007/978-3-319-70163-9

British Broadcasting Corporation. (1997). *BBC sound effects library index.* Films for the Humanities & Sciences. https://books.google.es/books?id=pWcAtAEACAAJ

Caillon, A., & Esling, P. (2021). *RAVE: A variational autoencoder for fast and high-quality neural audio synthesis* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2111.05011

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* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2202.00874

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., & Chan, W. (2020). *WaveGrad: Estimating Gradients for Waveform Generation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2009.00713

Chowning, J. M. (1977). The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. *Computer Music Journal*, *1*(2), 46–54. Retrieved September 4, 2023, from https://www.jstor.org/stable/23320142

Cooper, E., Wang, X., & Yamagishi, J. (2022). *Text-to-Speech Synthesis Techniques for MIDI-to-Audio Synthesis* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2104.12292

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). *Simple and Controllable Music Generation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2306.05284

Cuenca, P., & Apolinário. (2023). Train your ControlNet with diffusers. *Hugging Face*. Retrieved May 17, 2023, from https://huggingface.co/blog/train-your-controlnet

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). *Jukebox: A Generative Model for Music* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2005.00341

Donahue, C., McAuley, J., & Puckette, M. (2019). *Adversarial Audio Synthesis* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1802.04208

Dong, H.-W., Chen, K., McAuley, J., & Berg-Kirkpatrick, T. (2020). *MusPy: A Toolkit for Symbolic Music Generation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2008.01951

DuBreuil, A. (2020). *Hands-on music generation with magenta*. Packt Publishing Ltd. https://www.oreilly.com/library/view/hands-on-music-generation/9781838824419/

Elizalde, B., Deshmukh, S., Ismail, M. A., & Wang, H. (2022). *CLAP: Learning Audio Concepts From Natural Language Supervision* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2206.04769

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). *GANSynth: Adversarial Neural Audio Synthesis* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1902.08710

Engel, J., Hantrakul, L., Gu, C., & Roberts, A. (2020). *DDSP: Differentiable Digital Signal Processing* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2001.04643

Font Corbera, F., Roma Trepat, G., & Serra, X. (2013). Freesound technical demo. *MM '13: Proceedings of the 21st ACM international conference on Multimedia*, 411–2. https://doi.org/http://dx.doi.org/10.1145/2502081.2502245

Forsgren, S., & Martiros, H. (2022). *Riffusion - Stable diffusion for real-time music generation*. https://riffusion.com/about

Fourier, J.-B.-J. (1822). *Théorie analytique de la chaleur, par M. Fourier*. Chez Firmin Didot, père et fils. https://doi.org/10.1017/CBO9780511693229

Gardner, J., Simon, I., Manilow, E., Hawthorne, C., & Engel, J. (2022). *MT3: Multi-Task Multitrack Music Transcription* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2111.03017

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. *Proc. IEEE ICASSP 2017*. https://doi.org/10.1109/ICASSP.2017.7952261

Gugger, S., Debut, L., Wolf, T., andZachary Mueller, P. S., Mangrulkar, S., Sun, M., & Bossan, B. (2022). Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate

Guzhov, A., Raue, F., Hees, J., & Dengel, A. (2021). *AudioCLIP: Extending CLIP to Image, Text and Audio* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2106.13043

Hawthorne, C., Simon, I., Roberts, A., Zeghidour, N., Gardner, J., Manilow, E., & Engel, J. (2022). *Multi-instrument Music Synthesis with Spectrogram Diffusion* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2206.05408

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J., & Eck, D. (2019). *Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1810.12247

Hayes, B., Shier, J., Fazekas, G., McPherson, A., & Saitis, C. (2023). *A review of differentiable digital signal processing for music & speech synthesis* (tech. rep.). https://arxiv.org/abs/2308.15422

Heikkilä, M. (2022). The Algorithm: AI-generated art raises tricky questions about ethics, copyright, and security. *MIT Technology Review*. Retrieved August 29, 2023, from https://www.technologyreview.com/2022/09/20/1059792/the-algorithm-ai-generated-art-raises-tricky-questions-about-ethics-copyright-and-security/

Herndon, H., & Dryhurst, M. (2023). Have i been trained? *Spawning*. https://haveibeentrained.com/

Herndon, H., & Never Before Heard Sounds. (2021). *Holly+*. https://holly.plus/

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). *CNN Architectures for Large-Scale Audio Classification* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1609.09430

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2006.11239

Holz, D., Keller, J., Friedman, N., Rosedale, P., & Warner, B. (2022). *Midjourney*. https://www.midjourney.com/home/

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). *Music transformer* (tech. rep.). https://arxiv.org/abs/1809.04281

Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., & Ellis, D. P. W. (2022). *MuLan: A Joint Embedding of Music Audio and Natural Language* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2208.12415

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* (tech. rep.). https://arxiv.org/abs/2301.12661

Kadlcík, M., Hájek, A., Kieslich, J., & Winiecki, R. (2023). *A Whisper transformer for audio captioning trained with synthetic captions and transfer learning* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2305.09690

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A. v. d., Dieleman, S., & Kavukcuoglu,

K. (2018). *Efficient Neural Audio Synthesis* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1802.08435

Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2019). *Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1812.08466

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

Kim, J. W., Bittner, R., Kumar, A., & Bello, J. P. (2018). *Neural Music Synthesis for Flexible Timbre Control* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1811.00223

Kingma, D. P., & Welling, M. (2022). *Auto-Encoding Variational Bayes* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Kong, J., Kim, J., & Bae, J. (2020). *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2010.05646

Kong, Q., Cao, Y., Liu, H., Choi, K., & Wang, Y. (2021). *Decoupling Magnitude and Phase Estimation with Deep ResUNet for Music Source Separation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2109.05418

Kong, Q., Li, B., Song, X., Wan, Y., & Wang, Y. (2021). *High-resolution Piano Transcription with Pedals by Regressing Onset and Offset Times* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2010.01815

Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2021). *DiffWave: A Versatile Diffusion Model for Audio Synthesis* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2009.09761

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., & Adi, Y. (2023). *AudioGen: Textually Guided Audio Generation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2209.15352

Li, B., Liu, X., Dinesh, K., Duan, Z., & Sharma, G. (2019). Creating A Multitrack Classical Musical Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. *IEEE Transactions on Multimedia*, *21*(2), 522–535. https://doi.org/10.1109/TMM.2018.2856090

Likert, R. (1932). *A technique for the measurement of attitudes* (Vol. 22). Archives of Psychology.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., & Plumbley, M. D. (2023). *AudioLDM: Text-to-Audio Generation with Latent Diffusion Models* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2301.12503

Liu, H., Tian, Q., Yuan, Y., Liu, X., Mei, X., Kong, Q., Wang, Y., Wang, W., Wang, Y., & Plumbley, M. D. (2023). *AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2308.05734

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Manilow, E., Wichern, G., Seetharaman, P., & Roux, J. L. (2019). *Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity* (tech. rep.). arXiv. Retrieved July 24, 2023, from http://arxiv.org/abs/1909.08494

Manzelli, R., Thakkar, V., Siahkamari, A., & Kulis, B. (2018). *Conditioning Deep Generative Raw Audio Models for Structured Automatic Music* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1806.09905

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., & Bengio, Y. (2017). *SampleRNN: An Unconditional End-to-End Neural Audio Generation Model* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1612.07837

MIDI Manufacturers Association. (1996). *The complete midi 1.0 detailed specification: Incorporating all recommended practices.* https://books.google.es/books?id=cPpcPQAACAAJ

Moebert, T., Hanappe, P., Berhörster, C., Schmitt, A., López-Cabanillas, P., Green, J., & Henningsson, D. (2018). *Fluidsynth: A soundfont synthesizer.* https://www.fluidsynth.org/

Moog, R. A. (1964). Voltage controlled electronic music modules. *Journal of The Audio Engineering Society*, *13*, 200–206. https://api.semanticscholar.org/CorpusID:108780667

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A Generative Model for Raw Audio* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1609.03499

Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. (2017). *FiLM: Visual Reasoning with a General Conditioning Layer* (tech. rep.). arXiv. 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* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2103.00020

Raffel, C. (2016). *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching* [Doctoral dissertation]. https://colinraffel.com/projects/lmd/

Raffel, C., & Ellis, D. P. W. (2014). Intuitive analysis, creation and manipulation of midi data with pretty$_m$*idi*. *Proceedings of the 15th International Conference on Music Information Retrieval Late Breaking and Demo Papers.* https://github.com/craffel/pretty-midi

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). *Zero-Shot Text-to-Image Generation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2102.12092

Renault, L., Mignot, R., & Roebel, A. (2022). Differentiable piano model for midi-to-audio performance synthesis. *Proceedings of the 25th International Conference on Digital Audio Effects.*

Robitzski, D. (2020). A bot made frank sinatra cover britney spears. youtube removed it over copyright claims. *Futurism.* Retrieved August 30, 2023, from https://futurism.com/bot-frank-sinatra-britney-spears-youtube-copyright

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2112.10752

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1505.04597

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2205.11487

Schneider, F., Jin, Z., & Schölkopf, B. (2023). *Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2301.11757

Shih, Y.-J., Wu, S.-L., Zalkow, F., Müller, M., & Yang, Y.-H. (2022). *Theme Transformer: Symbolic Music Generation with Theme-Conditioned Transformer* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2111.04093

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). *Deep Unsupervised Learning using Nonequilibrium Thermodynamics* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1503.03585

Steinberg. (1996). Steinberg cubase 3. *Sound On Sound*.

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2018). *Neural discrete representation learning* (tech. rep.). https://arxiv.org/abs/1711.00937

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1706.03762

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., & Wolf, T. (2022). Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers

Weidenaar, R. (1995). *Magic music from the telharmonium.* Metuchen, N.J. : Scarecrow Press. Retrieved September 4, 2023, from http://archive.org/details/bub_gb_Gr2kq-598-YC

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., . . . Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://github.com/huggingface/transformers

Wu, H.-H., Seetharaman, P., Kumar, K., & Bello, J. P. (2022). *Wav2CLIP: Learning Robust Audio Representations From CLIP* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2110.11499

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2022). *Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2211.06687

Wu, Y., Manilow, E., Deng, Y., Swavely, R., Kastner, K., Cooijmans, T., Courville, A., Huang, C.-Z. A., & Engel, J. (2022). *MIDI-DDSP: Detailed Control of Musical Performance via Hierarchical Modeling* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2112.09312

Yamamoto, R., Song, E., & Kim, J.-M. (2020). *Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.1910.11480

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., & Yu, D. (2023). *Diffsound: Discrete Diffusion Model for Text-to-sound Generation* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2207.09983

Zeng, M., Tan, X., Wang, R., Ju, Z., Qin, T., & Liu, T.-Y. (2021). *MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2106.05630

Zhang, L., & Agrawala, M. (2023). *Adding Conditional Control to Text-to-Image Diffusion Models* (tech. rep.). arXiv. https://doi.org/10.48550/arXiv.2302.05543

Appendix A Survey

In the initial stage of the research, a survey titled "Text-driven deep learning music production tools" was realized. The survey was hosted in Google Forms and was addressed to both amateur and professional music producers. The purpose of this survey was to find out which text-driven audio synthesis tools could serve as a useful tool during the music production process. Figure 8 shows the main questions from the survey, as well as a summary of the provided answers. Even though the participation was low ($\approx$ 15 participants) and the results are far from conclusive, a series of interesting deductions can be made from them.

In the first place, respondents were asked how useful they would find an AI music production tool for the following purposes: sample generation, melody generation, loop generation, accompaniment generation, and whole song generation. As seen by the provided answers, most music producers prefer tools for specific tasks (sample generation), in contrast to having less control over the musical output (whole song generation).

Secondly, respondents were told to score the usefulness of the following applications: text-to-audio generation, text-to-MIDI generation, text-conditional MIDI-to-audio synthesis, text-conditional audio-to-audio translation, and text-conditional MIDI-to-MIDI translation. Text-to-audio tools are seen as the most useful to music producers, followed by text-driven applications involving MIDI and, finally, text-conditional audio-to-audio translation.

A third question asked how likely the respondents were to use a tool to generate the following types of sounds: environmental sounds, synthetic sounds, acoustic instrument sounds, singing voice, and drums. As seen by the answers, most producers are interested in generating environmental and synthetic sounds. Many respondents did not find reproducing sounds of acoustic instruments particularly useful.

Finally, the respondents were asked to rate the usefulness of the following types of text conditioning: name or type of instrument, timbric features, low-level, mid-level or high-level audio features, or style of specific musicians. The results provided show that most music producers are interested in describing timbric features, while the remaining features are considered equally useful.

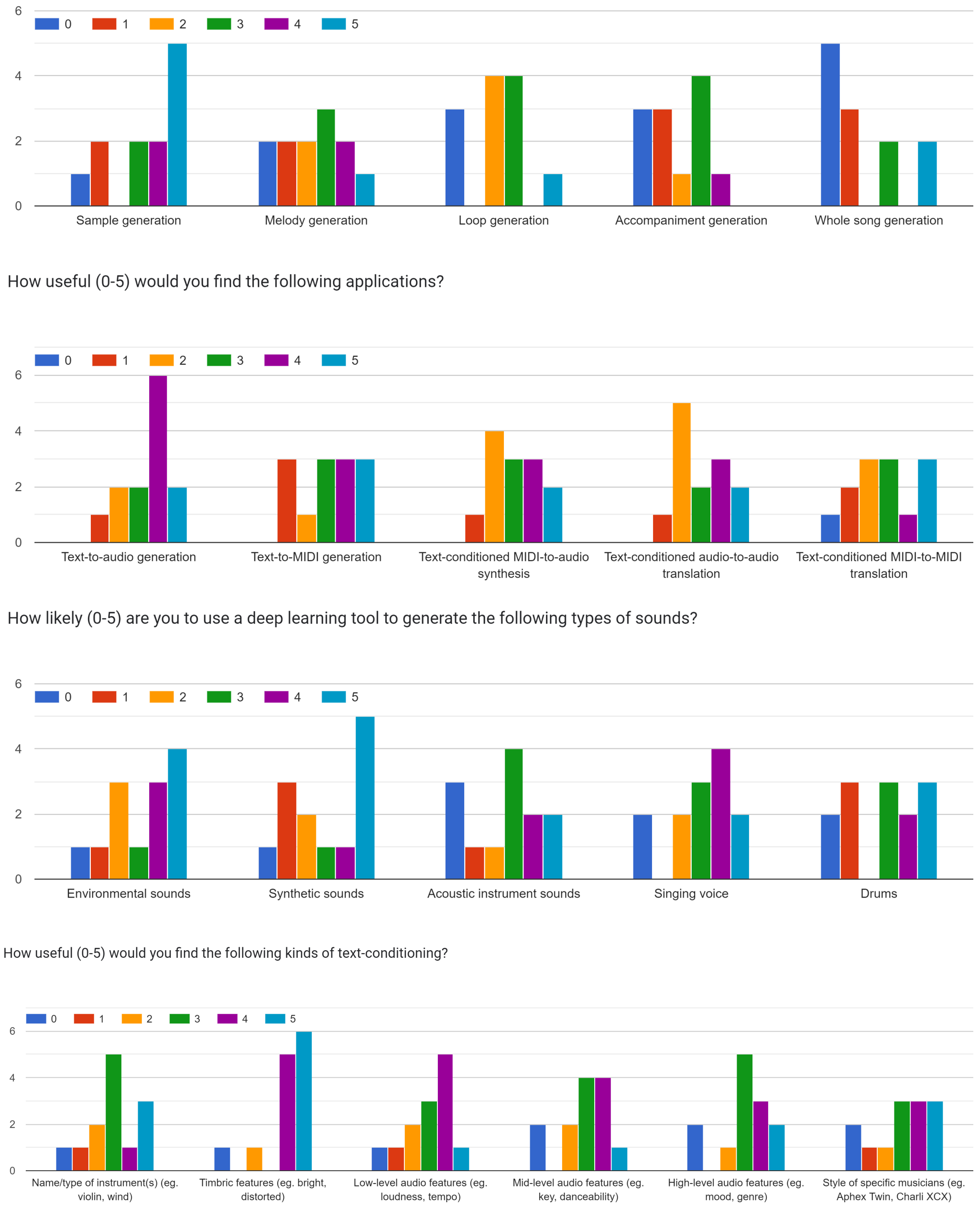
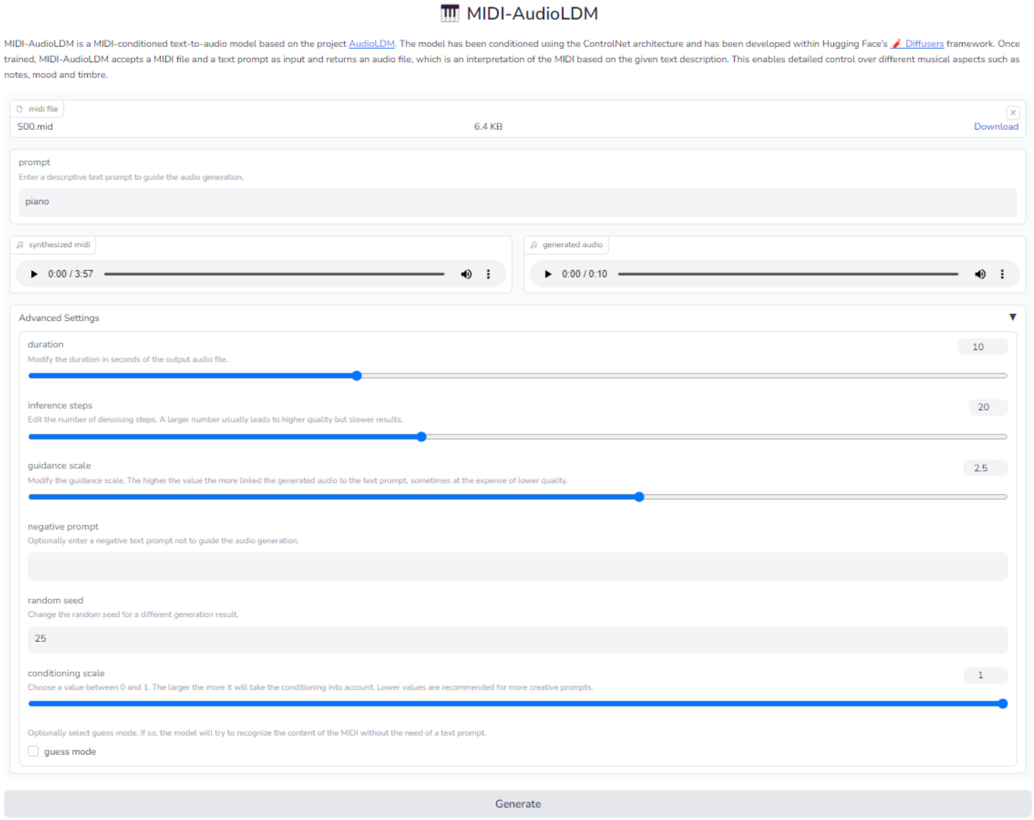


Figure 8: Results from "Text-driven deep learning music production tools" survey, hosted in Google Forms.

Appendix B Demo

A demo for MIDI-AudioLDM[24] is hosted in Hugging Face Spaces. The resulting interface is shown in Figure 9. In this case, one of the predetermined examples has been loaded. The selected MIDI file can be removed, and a local MIDI file can be uploaded instead. The MIDI file is automatically synthesized using a basic synthesis tool, and shown on the left for comparison purposes and to facilitate the selection of the desired duration. The Advanced Settings, hidden by default, offer detailed control over the generation process. This includes audio duration, negative prompt and conditioning scale among other parameters that are described accordingly. Once the 'Generate' button is pressed, the audio is synthesized and displayed on the right.



MIDI-AudioLDM

MIDI-AudioLDM is a MIDI-conditioned text-to-audio model based on the project AudioLDM. The model has been conditioned using the ControlNet architecture and has been developed within Hugging Face's Diffusers framework. Once trained, MIDI-AudioLDM accepts a MIDI file and a text prompt as input and returns an audio file, which is an interpretation of the MIDI based on the given text description. This enables detailed control over different musical aspects such as notes, mood and timbre.

midi file: S00.mid 6.4 KB

prompt
Enter a descriptive text prompt to guide the audio generation.
piano

synthesized midi: 0:00 / 3:57

generated audio: 0:00 / 0:10

Advanced Settings

duration: 10
Modify the duration in seconds of the output audio file.

inference steps: 20
Edit the number of denoising steps. A larger number usually leads to higher quality but slower results.

guidance scale: 2,5
Modify the guidance scale. The higher the value the more linked the generated audio to the text prompt, sometimes at the expense of lower quality.

negative prompt
Optionally enter a negative text prompt not to guide the audio generation.

random seed
Change the random seed for a different generation result.
25

conditioning scale: 1
Choose a value between 0 and 1. The larger the more it will take the conditioning into account. Lower values are recommended for more creative prompts.

Optionally select guess mode. If so, the model will try to recognize the content of the MIDI without the need of a text prompt.
□ guess mode

Figure 9: MIDI-AudioLDM demo hosted in Hugging Face Spaces.

[24] https://huggingface.co/spaces/lauraibnz/midi-audioldm

Appendix C Conferences

MIDI-AudioLDM has been presented at Sónar+D 2023 and will be presented at Volumens Festival 2023 in the following weeks. A brief description of both of these is given next:

- **Sónar+D. 15-17 July 2023. Barcelona, Spain.** Sónar+D[25] is an international conference and festival that explores the intersection of creativity, technology and art, with a special focus on music. MIDI-AudioLDM was selected along with 34 other projects out of more than 450 applications. The project was showcased during the three days of the festival as part of the Project Area, in the section of Music & Sound[26]. A large number of people were able to ask questions about MIDI-AudioLDM and test the Hugging Face demo. The project was featured in the Sónar+D pre-summer recap[27] as part of the festival's newsletter.

- **Volumens Festival. 23 September 2023. Valencia, Spain.** Volumens[28] is an annual festival that encompasses and explores the field between contemporary art, music, science, and technology. With music as its main focus, the current edition of the festival will be dedicated to artificial intelligence and its applications to art. A hybrid talk/workshop will be presented[29] about text-to-audio models and MIDI-AudioLDM.

In addition, the Hugging Face Space for MIDI-AudioLDM was awarded a GPU Grant[30] by the Hugging Face community. This Grant is provided to a select number of "innovative Spaces" to assist in covering the costs of GPU hardware upgrades.

[25] https://sonar.es/en/programme/sonar-d

[26] https://sonar.es/es/actividad/project-area-music-and-sound

[27] https://r.contact.sonar.es/mk/mr/sh/1t6AVsd2XFnIGF9UytoK2Oain1CyiD/6lo-RHi1DsPm

[28] https://volumens.es/en/home-2/

[29] https://volumens.es/es/project/laura-ibanez/

[30] https://huggingface.co/docs/hub/spaces-gpus#community-gpu-grants