EXTENDING AUDIOLDM AND AUDIOCRAFT MUSICGEN FOR MULTI-MODAL CONDITIONING AND AUDIO-TO-AUDIO GENERATION

A PREPRINT

Selim Elbindary
University of Stuttgart
Stuttgart, Germany
st188997@stud.uni-stuttgart.de

Mohamed Youssef
University of Stuttgart
Stuttgart, Germany
st190193@stud.uni-stuttgart.de

July 14, 2025

ABSTRACT

Generating instrumental music from vocal input—such as singing or humming—offers an intuitive interface for musical expression but remains underexplored in generative modeling. This work investigates two complementary approaches to this task: **AudioLDM**, a latent diffusion model originally designed for text-to-audio generation, and **AudioCraft (MusicGen)**, an autoregressive Transformer model.

AudioLDM is extended to support multi-modal conditioning with both vocal and textual prompts, enabling flexible control over genre and style. MusicGen is fine-tuned to perform voice-to-music generation using melody extracted from vocals, with optional joint conditioning via text. A custom dataset of vocal-instrumental pairs across multiple genres and languages is used for training and evaluation.

Results suggest that joint vocal and text conditioning improves generation quality, with MusicGen demonstrating superior coherence and stylistic alignment. These findings highlight the potential of voice-guided generative systems for creative and accessible music production.

1 Introduction

Generative models for audio synthesis have developed at a quick pace in the recent years, primarily propelled by the invention of *latent diffusion models* and *large-scale autoregressive transformers*, which offer unprecedented levels of control and fidelity in audio generation. These technological improvements have given way to fresh possibilities in music composition, sound design, and voice-based interaction systems.

Some of the most prominent recent architectures are **AudioLDM** [1] and **AudioCraft (MusicGen)** [2], each representing a different paradigm in audio generative modeling. AudioLDM is a diffusion-based architecture designed for *text-to-audio* synthesis. It operates within the latent space of Mel spectrograms, based on the integration of a Variational Autoencoder (VAE), the CLAP audio-text encoder [3], and a UNet-based latent diffusion model. While AudioLDM generates high-quality performance in semantically-guided sound synthesis, it does not handle audio-conditioning, limiting its use in tasks such as voice-to-instrument or audio style change.

On the other hand, AudioCraft (more specifically, its *MusicGen* version) employs a decoder-only Transformer [2] to autoregressively condition sequences of EnCodec [4] audio tokens. MusicGen can condition on text or melody, with the latter being represented as chroma pitch features. This structure is well-suited for being optimally appropriate to controllable music generation, particularly within the instrumental setting.

Despite these advances, one very expressive modality—**human voice**—remains significantly unexplored. Voice, singing, or humming is an intuitive, natural way for humans to communicate musical concepts. Current models, however, do not usually enable direct music generation from voice input often. This requires models that learn in parallel low-level

acoustic features and high-level semantic control so they can produce instrumental accompaniments within a certain genre or emotion based on raw voice input.

In this work, we present a multi-modal variant of AudioLDM to facilitate text and audio conditioning through the employment of double CLAP encoders and the combination of their embeddings in diffusion-based synthesis. We further fine-tune the MusicGen model to accept vocal input as melodic guidance alongside textual genre information to enhance its voice-conditioned music generation ability.

To train and validate these models, we establish a new dataset of 416 vocal-instrumental song pairs for 42 artists across various genres and languages. The dataset includes aligned metadata such as genre, key, BPM, and mood, with fine-grained control and evaluation.

Our project aims to pave the way for a next generation of generative systems that may allow users to *sing to make music*, optionally with the help of text inputs. We compare and contrast AudioLDM and MusicGen along this measure, highlighting their respective strengths and limitations.

2 Background

2.1 AudioLDM

AudioLDM comprises three main components:

- **VAE:** Encodes Mel spectrograms into a compact latent representation. Trained with a perceptual loss for high-quality reconstructions.
- **CLAP:** Contrastive Language-Audio Pretraining model used to extract semantic embeddings from text prompts.
- **Latent Diffusion Model:** A UNet-based denoising model that learns to generate audio in the latent space.

2.1.1 Pipeline Overview

1. Text input is encoded using CLAP (text mode) into a semantic embedding.
2. Input audio is transformed into a Mel spectrogram, then encoded into a latent representation via VAE.
3. The diffusion model generates latent audio conditioned on the CLAP embedding.
4. The VAE decoder reconstructs the Mel spectrogram from the latent space.
5. HiFi-GAN vocoder converts the Mel spectrogram to waveform.

2.2 AudioCraft (MusicGen)

MusicGen, introduced by Meta as part of the AudioCraft framework [2], is a state-of-the-art controllable music generation model. Unlike diffusion models, MusicGen formulates audio generation as an autoregressive sequence modeling task. It is based on a decoder-only Transformer model that learns to generate music token by token, conditioned on optional text descriptions and melody references.

The model is configured to produce high-fidelity music of 32 kHz by predicting discrete tokens that have been generated from an audio codec called **EnCodec** [4]. This allows MusicGen to produce coherent musical compositions with consistent structure, instrumentation, and style.

Tokenization with EnCodec

Raw audio waveforms are first compressed using EnCodec, a neural audio codec that transforms waveform audio into a set of discrete token sequences (codebooks) [4]. Each frame of audio (sampled at 32 kHz) is represented using multiple codebooks (typically 4), each emitting one token per 20 ms (i.e., 50 Hz). This tokenized representation drastically reduces the sequence length, making it tractable for Transformers while preserving fidelity.

During training, MusicGen learns to predict these token sequences in an autoregressive fashion—i.e., each new token is generated based on all previous tokens and any available conditioning.

Transformer Architecture

At the heart of MusicGen lies a GPT-style decoder-only Transformer, trained to model the probability distribution over EnCodec tokens [2]. It attends to past token context and optional conditioning inputs. MusicGen is trained using teacher forcing, optimizing a standard cross-entropy loss between the predicted and target token sequences.

The model is available in multiple sizes:

- **Small (300M)**, **Medium (1.5B)**, and **Large (3.3B)** parameters,
- Each with different capacities depending on the target use case (e.g., melody-only vs. full multi-modal generation).

Conditioning Mechanisms

MusicGen can operate unconditionally, but its core strength lies in its controllability via conditioning. Two primary forms of conditioning are supported:

- **Text Conditioning:** Descriptive prompts (e.g., `"a bird chirping in a windy field"`) are embedded using a frozen text encoder such as T5 [5]. The embeddings are prepended or integrated through cross-attention into the Transformer, allowing the model to guide its generation semantically.
- **Melody Conditioning (MusicGen-Melody):** A reference melody (e.g., vocal humming or instrumental sketch) is transformed into a 12-dimensional chroma feature vector over time, representing pitch class intensity. These chroma embeddings act as musical scaffolding, ensuring the generated audio aligns rhythmically and harmonically with the reference [2].

MusicGen Pipeline Overview

The full MusicGen generation pipeline proceeds as follows:

1. A raw audio signal (melody reference) is processed to extract chroma pitch features (optional).
2. A natural language prompt is processed by a frozen encoder (e.g., T5) to obtain semantic embeddings.
3. A decoder-only Transformer autoregressively generates EnCodec token sequences across multiple codebooks, conditioned on the available inputs.
4. The generated token sequence is passed to the EnCodec decoder [4] to reconstruct a high-fidelity waveform (typically 32 kHz).

Strengths and Applications

MusicGen offers fine-grained control over both the content and style of generated music. Its dual conditioning system supports diverse applications, including:

- **Prompt-based composition:** Create music using natural language descriptions alone.
- **Voice-to-instrument synthesis:** Generate instrumentals based on sung or hummed melodies.
- **Melody harmonization:** Extend or arrange a reference melody with multi-instrument backing.
- **Genre transfer:** Render a melody in different genres by modifying the text prompt.

By leveraging EnCodec for discrete audio representation and Transformer-based modeling for sequence prediction, MusicGen achieves high-quality and musically coherent outputs with temporal consistency, outperforming many diffusion-based approaches in terms of generation speed and style control.

3 Dataset Generation

One of the main challenges in developing a voice-to-music generation model is the absence of a publicly available dataset that maps isolated vocal inputs to corresponding instrumental accompaniments. To address this limitation, we constructed a custom dataset specifically curated for training and evaluating our models under multi-modal conditioning scenarios.

We collected a total of **416 songs** spanning **42 different artists**, with approximately 10 songs per artist. Each song was manually or semi-automatically processed to extract two key components using Demucs:

- **Vocals:** Clean, isolated vocal tracks used as audio condition inputs.
- **Instruments:** Instrumental components of the same song used as generation targets.

To enrich the dataset and support conditional generation tasks, we tracked the following **metadata** for each track:

- Genre
- Artist name
- Mood descriptors
- Title
- Musical key
- Beats per minute (BPM)

The dataset also features a multilingual composition, supporting research in cross-lingual generation. The **language distribution** of the collected songs is as follows:

- **English:** 69%
- **Arabic:** 19%
- **French:** 12%

To further augment the training data and allow the models to learn from shorter sequences, each track was segmented into **time-aligned chunks**. The chunk duration was adapted to suit each model's training requirements:

- **30-second segments** were used for fine-tuning *MusicGen Melody*, allowing the model to capture long-range harmonic and structural dependencies.
- **20-second segments** were used for fine-tuning *AudioLDM*, optimized for learning in the Mel spectrogram latent space with diffusion models.

This curated dataset forms the foundation for fine-tuning both the AudioLDM and AudioCraft models, enabling them to learn mappings from vocal cues (with or without additional text prompts) to coherent instrumental audio outputs.

4 Methodology

In this section, we describe the architectural limitations of the original models and outline our proposed enhancements for multi-modal music generation from vocal input. We divide our approach into two parallel lines of work: extending **AudioLDM** for audio-based conditioning and leveraging **MusicGen** for melody- and text-guided generation.

4.1 AudioLDM-Based Pipeline

4.1.1 Limitations of the Original Architecture

The original *AudioLDM* pipeline was designed for **text-to-audio** generation. It lacked the ability to accept audio as a conditioning source, limiting its applicability for tasks such as vocal-to-instrument conversion, sound editing, or audio style transfer. Furthermore, it could only leverage semantic cues from text prompts, without modeling the acoustic features of a vocal input.

4.1.2 Proposed Enhancements

To enable multi-modal conditioning, we introduced the following key architectural modifications:

- **Dual-Branch CLAP Encoding:** Both the audio and text branches of CLAP are used simultaneously. Input vocals and text prompts are encoded in parallel to capture both acoustic and semantic context.
- **Embedding Concatenation:** The resulting audio and text embeddings are concatenated to form a unified conditioning vector.
- **Diffusion UNet Modification:** The multi-modal conditioning vector is injected into the UNet either via Feature-wise Linear Modulation (FiLM) layers or direct concatenation, enabling the model to attend to both modalities.

4.1.3 Modified Architecture Pipeline

The enhanced AudioLDM pipeline operates as follows:

1. **Input Audio:** Preprocessed to extract both CLAP audio embedding and latent Mel features via a pretrained VAE.
2. **Text Prompt:** Encoded using CLAP in text mode to produce semantic embeddings.
3. **Conditioning:** Audio and text embeddings are concatenated into a single vector.
4. **Diffusion:** The UNet generates latent audio conditioned on the multi-modal vector.
5. **Decoding:** The VAE decoder reconstructs a Mel spectrogram, which is then converted to waveform using HiFi-GAN.

4.1.4 Benefits of Multi-Modal Conditioning

- Combines low-level acoustic features with high-level semantic guidance.
- Enables use cases such as vocal-to-instrument transfer and style adaptation.
- Supports generation from either pure text prompts, pure audio, or hybrid input.

4.1.5 Implementation Notes

- **Backbone:** HTSAT-tiny (shared across CLAP text/audio modes)
- **Audio embeddings:** Pretrained CLAP (`clap_htsat_tiny.pt`)
- **VAE:** `vae_mel_16k_64bins.ckpt`
- **Diffusion model:** `audioldm-s-full.ckpt`

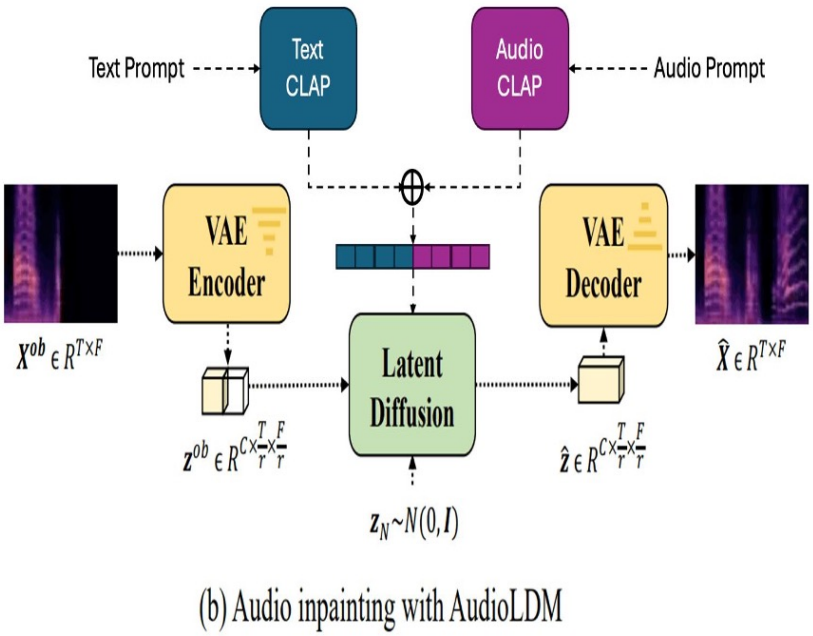

Figure 1: Modified AudioLDM pipeline with multi-modal conditioning from both CLAP text and audio encoders.

4.2 MusicGen-Based Pipeline

4.2.1 Architecture Overview

MusicGen (part of the AudioCraft framework) uses a transformer-based decoder-only architecture to autoregressively generate high-quality music at 32 kHz. Unlike diffusion-based models, it generates music by predicting sequences of EnCodec audio tokens.

- Decoder-only Transformer (GPT-style)
- Fully autoregressive modeling of instrumental audio
- Predicts EnCodec audio tokens, which are decoded into waveforms

4.2.2 Conditioning Mechanisms

MusicGen supports both text-based and melody-based conditioning:

- **Text Prompt:** Encoded using a frozen text encoder (e.g., T5) and used as contextual guidance via cross-attention.
- **Melody Audio (Optional):** Transformed into 12-dimensional chroma pitch features, used to guide harmonic structure without passing through the token encoder.

4.2.3 Generation Pipeline

1. Text prompt → text embeddings
2. Melody audio → chroma features
3. Transformer → EnCodec token prediction
4. EnCodec decoder → high-quality 32 kHz waveform

4.2.4 Key Benefits

- Provides fine-grained control over melody and genre
- Captures both semantic intent (text) and musical structure (melody)
- Produces high-fidelity, coherent, and stylistically rich instrumental music

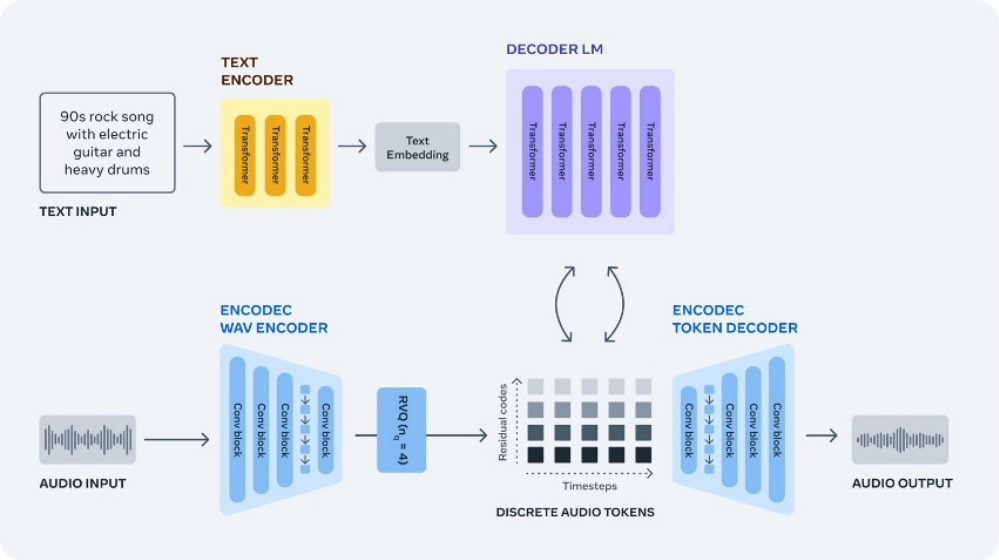

Figure 2: MusicGen pipeline with multi-modal conditioning

5 Experimental Settings

All experiments were conducted on the **AISA GPU cluster** at the University of Stuttgart, using a single **NVIDIA A40 GPU** with 48 GB of VRAM.

5.1 MusicGen Fine-Tuning

We fine-tuned the *MusicGen-Melody (medium)* model, a transformer-based architecture with approximately **1.5 billion parameters**. Fine-tuning was carried out to enable both melody-based and multi-modal music generation from vocal input.

Our approach involved two stages of fine-tuning:

1. **Audio-Only Conditioning (Voice to Music)**
   The model was initially trained to generate instrumental music from isolated vocal input. This taught the model to capture melody, rhythm, and timbre characteristics from raw voice signals and synthesize coherent musical accompaniment.
2. **Audio + Text Conditioning (Voice+Text to Music)**
   In the second stage, we introduced genre-specific text prompts alongside the vocal input. This allowed the model to align the instrumental output with both the musical content of the vocals and the stylistic semantics of the prompt. Example prompt: `"Generate Disco music that matches input vocals"`.

The fine-tuning for 2. was performed with the following hyperparameters:

- **Model:** MusicGen-Melody Medium (1.5B parameters)
- **Batch size:** 2
- **Total epochs:** 48
- **Training duration:** Each 12-epoch segment took approximately 12 hours, totaling **48 hours** of training time.

5.2 AudioLDM Fine-Tuning

We fine-tuned the *AudioLDM-s* model, which has approximately **330 million parameters**. The model was adapted for voice-to-music generation in the latent Mel spectrogram domain using diffusion-based synthesis.

Training was conducted on the same AISA GPU infrastructure. In a typical 12-hour session, the model was able to complete approximately **10,000 training steps**.

We fine-tuned three variants of the model to explore different conditioning strategies:

1. **Audio-Only Conditioning**
   The model was conditioned solely on vocal input (Mel spectrogram), allowing it to reconstruct or translate the input audio into instrumental form.
2. **Text + Audio Conditioning**
   The model received both the vocal input and a textual prompt describing the desired genre or mood. This dual conditioning setup was designed to guide the generation process semantically and acoustically.
3. **Adjusted Text + Audio Conditioning**
   In this configuration, we refined the textual prompts to more closely resemble those used with MusicGen (e.g., `"Generate Pop music that matches input vocals"`). The goal was to unify the prompting style and enable better comparison between model families.

Each variant was trained using **20 or 30-second audio chunks**, which matched the architectural and memory requirements of the latent diffusion process.

6 Results

Evaluating generative models for music synthesis remains an open research challenge, as there are no widely accepted objective metrics that can capture perceptual qualities such as musicality, emotional alignment, or stylistic fidelity. As such, our evaluation was conducted primarily through **qualitative listening tests** and manual inspection of the generated audio.

6.1 MusicGen-Melody Results

Fine-tuning the *MusicGen-Melody* model on our custom dataset produced highly promising results. In our qualitative tests, the generated music closely matched the original instrumentals in both **theme** and **rhythm**. The model was able to effectively capture the essence of the vocal melody and generate coherent instrumental backings that aligned with the vocal phrasing and dynamics.

Moreover, when we tested the model's ability to generate music in **different genres** from the same vocal input, the output correctly followed the genre-specific instructions provided in the text prompt. This demonstrated that the model had successfully learned to respond to both semantic and acoustic conditioning.

Compared to the **pretrained MusicGen** baseline, our fine-tuned model produced audio that was more **clear, consistent**, and **stylistically aligned** with the input vocals.

6.2 AudioLDM Results

Unlike MusicGen, AudioLDM was not originally designed for audio-to-audio generation. Therefore, we implemented several model modifications to enable this capability and trained three model variants using our dataset.

- **Audio-Only Conditioning:** The results were poor, with generated outputs being barely audible or musically coherent. The model struggled to produce meaningful audio when conditioned solely on vocal spectrograms.
- **Audio + Text Conditioning:** Introducing a genre prompt significantly improved output quality. The generated audio was more structured and reflected characteristics of the input vocals.
- **Adjusted Prompting:** Further refining the prompts to align with the structure used in MusicGen (e.g., `"Generate Rock music that matches input vocals"`) led to noticeably better results. The generated outputs exhibited clearer genre characteristics and stronger correlation with vocal input.

While the improvements to AudioLDM were effective, the quality and musical coherence of its outputs appeared to fall short of those produced by the fine-tuned MusicGen model. The transformer-based architecture and EnCodec representation of MusicGen appeared to better capture the complexity and structure required for instrumental music synthesis.

6.3 Qualitative Evaluation via Survey

To evaluate perceptual quality, we conducted a user study involving **10 participants**, each of whom listened to music generated by three different models: **MusicGen Pretrained**, **MusicGen Fine-tuned**, and **AudioLDM Fine-tuned**. The study included **4 songs**, with **3 questions per song**, resulting in a total of 120 responses. Participants were asked to rate:

1. Which model best aligned with the input vocals?
2. Which model produced the best audio quality?
3. Which model best matched the intended genre?

Vocal Matching Preference

Survey responses indicated a preference for the **fine-tuned MusicGen model**, which achieved the highest win rate of **58.57%** across evaluated tracks. In comparison, the **pretrained MusicGen** and **fine-tuned AudioLDM** models received lower win rates of **43.33%** and **40.00%**, respectively, highlighting the impact of fine-tuning and the relative strength of the transformer-based approach for vocal-conditioned music generation. It is important to note that the win rates among the models are relatively close, suggesting that while fine-tuning offers noticeable improvements, the overall performance gap is not overwhelming. A more detailed study with more participants would need to be conducted in order to get better insights.

Genre Matching Per Track

Genre alignment results were more balanced across models. Figure 4 shows the breakdown of genre-fit preferences for each test track.

Survey Insights

- **MusicGen Fine-tuned** was the most preferred model for vocal alignment and overall audio quality.

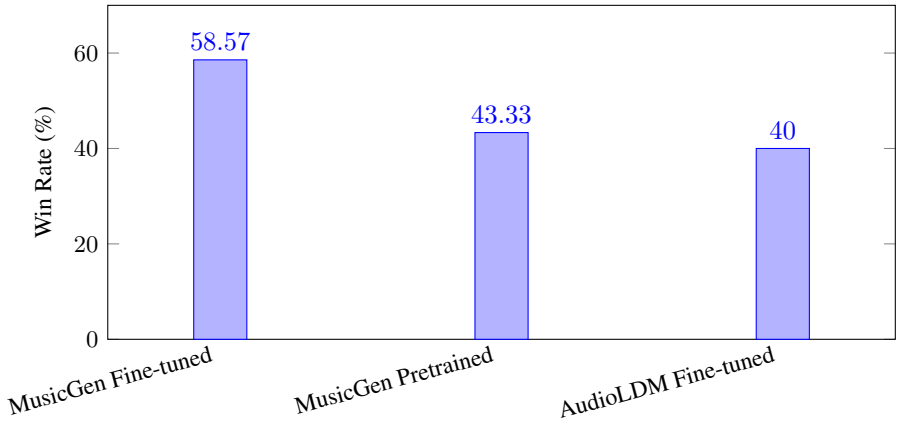

Figure 3: Win rate comparison across models based on survey responses

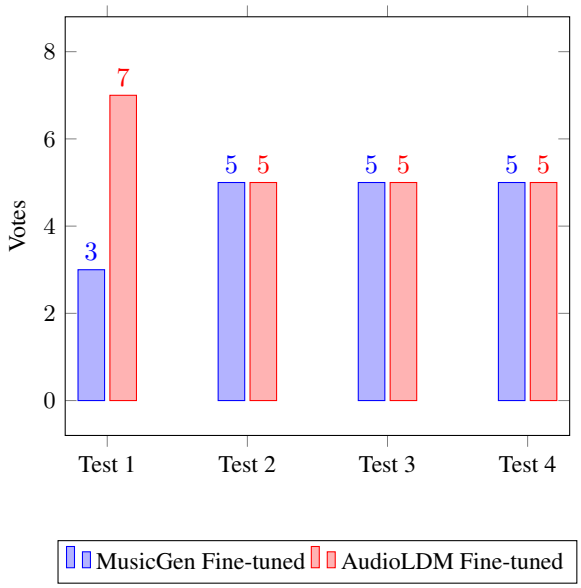

Figure 4: Genre fit preferences across the four test tracks.

- **Genre fit** was relatively balanced between MusicGen and AudioLDM models.
- Fine-tuning improved both models over their pretrained counterparts, particularly for vocal-matching tasks.

The **pretrained MusicGen** model was generally rated slightly lower in melody alignment, while **AudioLDM** received good feedback, performing reasonably with text+audio conditioning. These results underscore the importance of both *fine-tuning* and *multi-modal input* for high-quality music generation.

6.4 Quantitative Evaluation

While subjective listening tests are crucial for evaluating perceptual quality and musicality, we also performed quantitative analysis using two metrics: **Fréchet Audio Distance (FAD)** and the **CLAP score**. These metrics allow us to assess the fidelity and semantic alignment of generated audio, particularly in relation to genre and prompt consistency.

6.4.1 Fréchet Audio Distance (FAD)

Fréchet Audio Distance (FAD) [6] is a commonly used metric to evaluate the fidelity of generated audio. It measures the statistical distance between the embeddings of generated and reference audio using a pretrained VGGish network.

Lower FAD scores indicate better alignment with the distribution of real audio, and thus higher perceptual quality. FAD is analogous to the Fréchet Inception Distance (FID) used in image generation.

- We used ground-truth instrumental tracks as the reference distribution.
- Generated outputs from each model (AudioLDM Fine-tuned, MusicGen Pretrained, and MusicGen Fine-tuned) were used as the comparison set.
- FAD was computed per test track and averaged across all test examples.

Test Track	AudioLDM Fine-tuned	MusicGen Fine-tuned	MusicGen Pretrained
Test 1	**7.12**	15.20	10.98
Test 2	13.98	**6.62**	10.37
Test 3	9.69	**8.48**	9.03
Test 4	**7.12**	12.52	10.16

Table 1: FAD scores across four sample test tracks. Lower is better.

Model	Average FAD Score ↓
AudioLDM Fine-tuned	**9.48**
MusicGen Pretrained	10.64
MusicGen Fine-tuned	10.70

Table 2: Average FAD scores by model. Lower scores indicate better statistical alignment with reference instrumentals.

Interestingly, while qualitative evaluations and survey results favored the **fine-tuned MusicGen** model, the **AudioLDM Fine-tuned** model achieved the best average FAD score. This suggests that AudioLDM generated audio with statistical properties closest to the real instrumentals used for training, despite its relatively lower perceptual quality in human evaluations.

The fine-tuned MusicGen model performed slightly worse in terms of FAD, but offered better stylistic control and genre alignment—factors not captured by FAD. These results highlight the need for multiple evaluation perspectives in generative audio tasks, combining both perceptual feedback and statistical metrics.

6.4.2 CLAP-Based Evaluation (Genre and Prompt Consistency)

To evaluate how well generated outputs aligned with their intended **textual prompts and genre labels**, we used the **Contrastive Language-Audio Pretraining (CLAP)** model [3]. CLAP embeds audio and text into a shared space, enabling the computation of a similarity score between generated audio clip and corresponding text prompt.

- We computed **cosine similarity** between the audio embedding of each generated track and the embedding of the original text prompt or genre label.
- Higher CLAP scores indicate stronger semantic alignment between the generated audio and the intended description.

Model	CLAP Score ↑
MusicGen Fine-tuned	0.0974
MusicGen Pretrained	0.1266
AudioLDM Fine-tuned	0.1364

Table 3: CLAP scores between generated audio and textual prompts. **Higher is better.**

Important note: Since AudioLDM was trained using CLAP embeddings as part of its conditioning, CLAP-based evaluation is not used for cross-model comparison. Instead, it primarily serves to assess *relative prompt and genre alignment within the same model family*.

The results in Table 3 reveal that the **AudioLDM Fine-tuned** model achieved the highest CLAP score of **0.1364**, indicating a strong alignment between generated audio and textual prompts. This suggests that the introduction of multi-modal conditioning in AudioLDM, along with improved prompting, contributed positively to its performance.

In contrast, the **MusicGen Pretrained** model yielded a CLAP score of **0.1266**, while the **MusicGen Fine-tuned** model obtained the lowest score at **0.0974**. These results suggest that fine-tuning MusicGen on vocal-guided generation may have introduced minor variations in semantic consistency with the textual prompt.

It is important to interpret these scores cautiously. While the CLAP metric provides insight into text-audio alignment, it doesn't account for other dimensions such as melody preservation, musicality or genre fidelity. Moreover, lower CLAP scores for the finetuned MusicGen doesn't contradict the listening results, which showed a user preference for its outputs. This highlights that semantic alignment alone is not a sufficient indicator of perceptual quality.

Overall, these findings indicate that:

- AudioLDM, with proper multi-modal conditioning, can achieve competitive semantic alignment.
- MusicGen's autoregressive structure excels in perceptual quality and user preference, even when CLAP alignment drops slightly post-finetuning.

7 Conclusion

In this work, we explored the problem of generating instrumental music from vocal input, a relatively underexplored direction in music generation research. We investigated two distinct generative models—*MusicGen* and *AudioLDM*—and extended their capabilities to support **multi-modal conditioning** using both audio and text prompts.

To support this task, we created a **custom dataset** comprising 416 songs across 42 artists, with cleanly separated vocal and instrumental tracks, along with metadata such as genre, key, BPM, mood, and language. We chunked each song into shorter segments tailored for training both transformer-based and diffusion-based models.

Our contributions include:

- Fine-tuning **MusicGen-Melody (1.5B)** model on vocal-to-music tasks, with and without textual prompts. The model successfully generated genre-aligned, rhythmically coherent music that closely matched input vocals.
- Extending the **AudioLDM-s (330M)** model to support audio-to-audio and multi-modal conditioning. We introduced dual-branch CLAP encoders and prompt formatting strategies that significantly improved output quality over the baseline.
- Performing qualitative evaluations that demonstrated the effectiveness of multi-modal guidance, with MusicGen consistently producing superior results in clarity, musicality, and genre control.

Overall, our results suggest that transformer-based models like MusicGen are currently better suited for vocal-to-instrumental generation tasks. However, the extended AudioLDM architecture also showed promising behavior when augmented with the right conditioning and training strategies. Together, these models pave the way toward more expressive and controllable voice-driven music synthesis.

References

[1] Haohe Liu, Qiuqiang Kong, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[2] Jade Copet, Alexandre Défossez, Morgane Riviere, Gabriel Copet, Gabriel Synnaeve, and Yossi Adi. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

[3] Benjamin Elizalde, Bryan Krug, Eduardo Fonseca, Amin Ghias, Kevin Li, Shuayb Z Yang, Bhiksha Raj, and Beat Gfeller. Clap: Learning audio-text joint embedding for audio retrieval and tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

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

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[6] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019.