Bridging Paintings and Music – Exploring Emotion based Music Generation through Paintings

Tanisha Hisariya
Student Number - 220929356
Jinhua Liang
MSc Artificial Intelligence, QMUL

Abstract— **The rapid advancements in artificial intelligence, especially in generative tasks of music and images via unimodal or multimodal approaches, have made impactful implementation across various fields and industries. The primary aim of this research paper is to create a generative AI model capable of generating music resonating with emotions perceived by visual arts such as paintings. This paper aspires to contribute to the realm of representing it in latent space and enhancing the cross-modal expression. Furthermore, this research aims to empower visually challenged individuals by enabling them to perceive and understand any form of visual art through emotion-driven music, providing a novel sensory experience. The paper delves deeply into the study of related work, further making a way to implement transfer learning by integrating two different deep learning models- one for image-to-text generation and the other for text-to-music generation. It further extends by exploring the effect of various emotional textual descriptions to generate music and evaluating the effectiveness of generated music. Additionally, it identifies the limitations and suggests possible future work to improve model performance and robustness. This research is expected to contribute towards the applications of multimodal learning.**

Keywords— Music generation, Generative AI, Transformers, Images, Emotions, Music, CNN, Cross modality

I. INTRODUCTION

"Art is not what you see but what you make others see."- Edgar Degas. Visual art is perceiving information and emotions from the artist to the observer. It always encapsulates the perception, style, culture, generation, and innovation. From the ancient to modern periods, how people connect with the art form has changed profoundly. Being deeply ingrained in our culture and tradition, it reflects society's past, present, and future. Though it can reach a wider audience via visual perception, there should be no barrier for visually impaired persons to be excluded from enjoying the culture and emotion of art forms.

On the other hand, Music is another more critical form of creativity and innovation, stimulating a wide range of emotions through its tune and style. It is a separate form of creativity that flows with the same emotions and information as visual art. This can be adapted and lead to bridging a critical gap between paintings and Music. This paper is focused on the innovative intersection of two different art forms, generating music that reflects the emotions perceived by visual arts such as paintings. This method will not only help visually impaired persons perceive the emotions of visual arts through auditory signals and give them a different kind of experience but also open a wider research area in the field of generational models conditioned on images. This intersectional approach promises to enhance the accessibility and scope of creativity in the realm of AI. The rapid growth in technology has created a critical challenge regarding user satisfaction across diverse content forms. The transition of artificial intelligence, particularly in generative models, has always aimed to create new content that aligns perfectly with user expectations and needs. Generative AI models have shown remarkable achievement in various fields, from image generation to audio generation, leveraging the techniques of deep learning algorithms (Briot, 2020). The model's ability to capture the complex patterns of data and the relationship between them generates a new coherent pattern outperforming the traditional algorithms. Music generation is another important application of generative AI models, which use deep learning methods to generate music that is thought of as an unrealistic thing. The common approaches in music generation consist of symbolic and waveform music generation. Symbolic generation depicts events and note sequences (Vinet, 2004), mainly used by musicians or artists to interpret them. On the other hand, waveform music is a representation of the continuous domain of waveforms. It can be interpreted by anyone, making this kind of generation more suitable for daily interaction. However, according to (Colarusso et al., n.d), generating waveform music is highly dependent on the sampling rate to capture and process its structure fully, and this further requires a model with more effective training of a very high-dimensional dataset.

However, the cross-modality approach in Generative AI, such as image to music, is still a topic of discussion. One of the significant challenging aspects of this topic is to find the relation between the two different modalities, and the availability of paired data, which is essential for training the model. Collecting such resource-intensive datasets limits the scope of research. Despite these challenges, with the use of transfer learning, a pre-trained model on music generation tasks which further can be fine-tuned based on requirements. This approach reduces the need for a large dataset by using embedded knowledge on large, pre-trained models. This approach opens broader aspects to infuse different architectures and modalities to create hybrid models facilitating efficient and seamless architecture.

With the motivation of using the pre-trained approach, this paper bridges the gap between two different art forms, making them more accessible and receptive. In this paper, we will explore the described primary objectives:

1. We propose a method to generate music conditioned on the emotions of images.
2. We will work on integrating two different models: first converting the images into emotion-based labels and then using those labels to generate music using deep learning models.
3. We also explore the impact and performance of different textual descriptions on the music generation process in terms of its resemblance and quality.
4. We will also modify the training of base model to assess the improvement by reducing the time consumption and additional parameters.

To evaluate the generated music, it will be qualitatively measured across various metrics using the Fréchet Audio Distance (FAD) (Kilgour et al., 2019), CLAP (Elizalde et al., 2022), Total Harmonic Distortion (THD), Signal to noise ratio (SNR), and KL divergence. By employing advanced deep learning techniques and transfer learning, the aim is to develop a creative tool that will further give people unique experiences.

II. RELATED WORK

The work of this research paper has been associated with three different objectives as discussed below. We will discuss all these objectives respectively ranging from the first experiments done in this field to the recent techniques and their applications across various fields. Furthermore, we will critically analyse this research and provide insights about it.

- Explore the deep learning models for extracting features from images and converting it into words.
- Explore both modern deep learning techniques and the conventional machine learning techniques for evaluation of music generation.
- Exploration of multi-modality in audio and music generation.

A. Image feature extraction and conversion

Extracting an image's features and converting them into word embeddings can use an unimodal or multimodal approach. The unimodal approach will give a single label or categorical classification, while the multimodal approach will use the LLM to understand the visual representation using language processing.

The inception of more profound network architecture can be traced back to the invention of ConvNet based on supervised learning. Following this, the newest models, like LeNet and VGG, faced a prominent gradient degradation problem. Later, Resnet, developed by He et al., (2015) addresses this degradation problem by introducing an additional residual block and utilizing batch normalization. The implementation of shortcut connections helps the network to learn the features automatically and alleviates the disappearance of gradient problems.

Based on the basic architecture of these CNN models or transformers, along with the Natural language processing an Image captioning model, came into a pivotal role in capturing the content of an image. There has been recent advancement in this field with the introduction of the CLIP (Radford et al., 2021) model. Open AI came up with CLIP with a focus on the similarity between image and text, which uses a text encoder for textual information and an image encoder for visual details, further calculating the cosine similarity of these two pre-calculated vectors. BLIP (Bootstrapping Language-Image Pre-training) (Li et al., 2022) comes up with flexibility in transferring visual-language understanding and generation tasks. It has been incorporated with a multi modal mixture of encoder and decoder to achieve performance with great accuracy. The features of the input image are extracted by a CNN model and further passed to the transformer encoder. Based on this encoder, a transformer-based text decoder generates the caption. The advancement goes on to the next stage with the introduction of GPT-4 (openAI, 2023) as it is trained with more parameters handling the complexity of contextual information and visual information.

B. Deep Learning methods for Music Generation

In the era of neural networks, Todd (1989) used Recurrent Neural Networks (RNNs) to generate musical melodies. However, RNNs were unable to store the information for a very long period due to the vanishing gradient in their architecture. This challenge made the innovation of Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). This particular feature in the RNN model enabled the model to remember the sequences and their associated information. Lewandowski et al. (2012) came up with the Restricted Boltzmann Machine (RBM) incorporated into the RNN model, which gave a better result in terms of generating melodies; however, long-term memory was still a major drawback. In 2016, Google came up with its RNN model (Magenta, n.d) with improved long-term dependency, marking a significant advancement in the RNN-based music generation model.

The revolution in deep learning models has come up with generative models like Convolutional Neural Networks (CNN), Variational Auto Encoder (VAEs), and Generative Adversarial Networks (GANs). Roberts et al. (2019) proposed MusicVAE, a hierarchical VAE-based model. The model was able to capture the long-term dependencies of musical structure and have generated an efficient music by reconstruction. Yang. et al. (2017) proposed a CNN-based GAN model and melody-RNN called MidiNet. Due to the flexible architecture of this model, it can generate musical chords that follow a sequence. However, the quality and inference of music are not always excellent due to the modal collapsing of GAN. Dong et al. (2017) came up with MuseGAN, which can generate multitrack music based on GAN. Due to the model's ability to capture its interdependent nature, it can produce coherent music. However, interpreting complex musical data remained a limitation. Yu et al. (2017) came up with seqGAN, an RNN-based GAN model. This model enhanced the generation process by leveraging reinforcement learning techniques to optimize the creation of music.

In 2016, Google developed WaveNet (Google DeepMind, n.d., Min et al., 2022), a CNN-based raw audio form generation model. The model encompasses an autoregressive approach and diluted convolutions, which allows the model for long-term dependencies. This model was highly applied in text-to-speech and music generational tasks. However, due to its autoregressive nature, it demanded significant computational resources.

Cheng-Zhi et al. (2018) were the first to generate music using transformers. He proposed a Music Transformer, combining the transformer with a relative attention model to create long-term music. The efficiency of this model surpassed the RNN-based models, but the limitation was that it encompassed too many redundant notes. In 2019, openAI came up with MuseNet (Payne, 2019), a transformer based on the GPT-2 model, able to generate a variety of music and styles due to its ability to capture long-term sequences. Even the model was able to learn the complex patterns and structures producing high-quality music. However, challenges again arise due to the extensive computational cost and the interpretability of the model due to complex architecture.

Jacek and Teodara (n.d) developed a model for generating music based on emotion labels with Conditional VAEs and RNN. The latent space in this model gives a better interpolation, yet due to VAE, there is instability in training, causing a significant drawback.

C. Multi Modality Audio and Music generation

In 2022, cMelGAN (Kaunismaa, Qian and Chung, 2022), a conditional GAN model based on Mel Spectrogram. The feedback from the generator and additional training parameters from the discriminator enhanced efficient music generation. Additionally, the use of the Mel-spectrogram provides better information, but it lacks expressive elements. The limitation of this model lies in the training instability of GAN. A notable model called JukePix was proposed (Wang et al., 2018) for transforming paintings into music segments based on Convolutional GANs. Their architecture was based on the MuseGAN (Dong et al., 2017); it was able to generate multi-track music yet failed as per human evaluation and solo musical segment track. Chen et al. (2023) proposed MusicLDM, a text-to-music generation model based on a strategy of synchronous beat mix-up. The model is based on the incorporation of CLAP, VAE, Hifi-GAN, and diffusion models. The model performs exceptionally well due to its complex architecture; however, there are several limitations, including the sampling rate of trained data and the computational resource constraints. Liu et al. (2023) proposed AudioLDM, a text-to-audio generation model which uses CLAP embeddings and a latent diffusion model to learn about the continuous audio representation in latent space. The model gives a state-of-the-art performance in generating the audio with exceptional quality and fidelity. Leveraging this, AudioLDM 2 (Liu et al., 2023) came, which uses the chatGPT-2 model to work on any input modalities, either text, image, or video, and convert it into language, outperforming most of the models in terms of accuracy. Models like Riffusion (Liu et al., 2023) and Mousai (Schneider et al., 2023) were also introduced using diffusion model techniques for text-conditioned music generation. Andrea et al. developed the MusicLM model (Agostinelli et al., 2023), a text-conditioned music-generating model to generate music at a sampling rate of 24kHz. However, the model outperforms Riffusion and provides consistent music, yet it needed help understanding the precise ordering of negations in the text. Lam et al. (2023) proposed MeLoDy, a diffusion model architecture conditioned on a language model. The model is based on the language model of MusicLM (Agostinelli et al., 2023), for semantic modelling and a VAE-GAN as a decoder. The model applies a basis of a dual path diffusion model that incorporates this information at each step of denoising. However, the training data of this model was biased, limiting the varied music generation. Sheffer and Adi (2022) came up with the im2wav model, a transformer language model based on an imaged conditioned model to generate relevant audio. The first transformer language model will generate low-level audio, which will then be inputted to another model to upsample it, thus generating high-fidelity sound. For image embedding, this model uses pre-trained CLIP architecture. The model provided a baseline model for measuring the efficiency and inference in the same field. Yet the two transformer architectures made it computationally high; also, the model was unable to process the multi-scenic images. Very recently, in 2024, Chowdhury et al. proposed a model called MelFusion to synthesize music conditioned on images as well as text using the latest deep-learning diffusion models. This model architecture consists of extracting the image features using a stable diffusion model, further incorporating it with their textual description and using a diffusion model-based text-to-music model. The diffusion model exceeded the previous model's performance and efficiency, opening new areas for further research.

III. MusicGen

MusicGen (Copet et al., 2023) invented by MetaAI, is a single-language model architecture used to generate music. Because of its key features, it provides exceptional performance for the Music generation area on numerous benchmarks. The MusicGen-Small model has approximately 420 million trainable parameters.

The mode can be conditioned on both text and melody. This means when giving text as a prompt which includes mood, theme, genre, bpm, length, description and more musical characteristics, it can generate an audio closely matching with the described attributes. With melody conditioning, the model can modify the input melody by either adding or removing some features making a piece of creative and dynamic music. The model’s architecture consists of an audio Tokenizer that converts the continuous long audio into short discrete tokens which are further used by the transformer. This conversion makes the model computationally low while still able to capture all important features. Apart from this, it also simplifies the way data is being dealt with. Finally, the transformer model it is using is well-suited for the long-term dependencies of music. This architecture ensures the contextual dependence of the next note on its preceding note, making it perfect to generate resonating melodious music. The attention mechanism used in the transformer will be able to capture the global intricacies of tokens and maintain the importance of weighing all tokens to generate quality-assessed music.

The workflow of MusicGen model consists of converting audio into a sequence of tokens and then modelling them again, with an auto-regressive transformer-based decoder model. It has been built upon the Encodec (Défossez et al., 2022), which has main components such as encoder, quantizer, and decoder for the token generation of audio along with other conditioning models like T5 (Raffel et al., 2020) for encoding the text prompts.

A. Encodec Model

The Encodec Model comprises three components in their architecture – the encoder, the quantizer, and the decoder. This also uses the multiple interleaving codebook patterns in this model to process the input, where each pattern is responsible for reducing the dimensionality and hence removing the need for an upsampling model. The encoder is responsible for transforming the audio into low dimensional latent representation. The quantizer used is based upon residual vector quantization (RVQ), which helps to generate dependent quantized values across different codebooks.

B. T5 Text Encoder

MusicGen model uses a pre-trained T5 text encoder model for processing the text input tokens and passing it to the transformer language model. The T5 model is responsible for converting text into the matrix to explain information about text into tokens. An additional linear layer is added into the architecture, which converts the output of T5 into an acceptable form for the language model along with Encodec model.

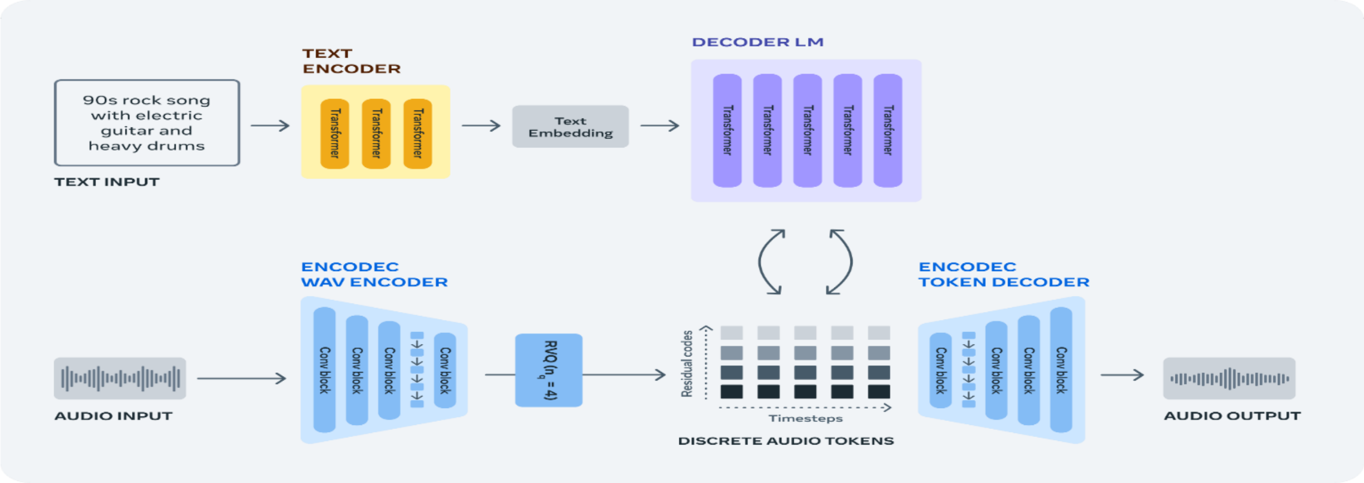

Fig. 1. Architecture of MusicGen model explaining the flow of working along with encoder, transformer and codebook interleaving patterns

Note: The given image is taken from the [https://docs.openvino.ai/2023.3/notebooks/250-music-generation-with-output.html] in order to illustrate the architecture

C. Transformer

A transformer-based decoder model then comes into play, which takes the input from codebooks and text condition embeddings. The decoder model is built of L layers and D dimensions, where each layer is composed of a cross-attention mechanism and various linear layers to decode, thus generating the required audio. The tokens from the conditioning model are mapped with audio tokens in latent representation. Thus, during inference, the model chooses the token with the most resemblance to the condition and puts it through a decoder for a generation. Besides, it includes an adversarial reconstruction loss, to improve the fidelity and realism of the generated audio by training the system to produce outputs indistinguishable from actual audio samples.

IV. METHODOLGY

The method of generating music from images based on emotions consists of integrating deep learning models.

Firstly, we will convert the images into their textual format, and then text along with its associated music will be used to fine-tune the MusicGen (Copet et al., 2023) model to generate the required music. By doing this, we are not only generating the music conditioned on emotions from images, but we are also exploring the effect of various textual descriptions during musical generation. In this research paper, we are exploring four models: - an emotion labelling model, an image description model, a Large Language model, and a Music Generation model. The overview of our methodology can be seen in Figure 3.

A. Image emotion Labelling Model.

To effectively perceive the emotions of images, we have introduced a classification model to label the emotions. The model will play an important part in determining the emotions during inference, helping to improve the consistency and relevance of generated music. Pre-trained ResNet50 (He et al., 2015) has been chosen as the best model for this approach due to its ability to manage diverse and complex datasets with dense layers. Leveraging the technique of transfer learning, the model has been adapted to a given dataset with the additional two GRU layers along with a multi-head Attention layer before the fully connected layer. The output layer of ResNet50 architecture has been flattening out to meet the output classes of the dataset. Further, to enhance the performance and to prevent overfitting, some dropping out of non-essential neurons, have been incorporated before fully connected layers.

GRU Layers – We included bi-directional GRU layers after extracting the features from initial ResNet layers.

$$\hat{h}_t = \tan h\,(W_h F_t + U_h (r_t \circ h_{t-1})) \qquad (1)$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \hat{h}_t \qquad (2)$$

Where h_t is the final hidden state and r_t is reset gate.

Attention Layers – This makes the main focus on important features only, making it more robust with query Q, key K, and value V metrices.

$$A_t = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

Given an input image I during inference, the framework classifies the image into emotion labels to further map the generated music with emotion.

$$E_{emotion} = f_{ResNet50}(I) \qquad (4)$$

Here, $f_{ResNet50}$ denotes the ResNet50 model generating $E_{emotion}$ label, which will also serve as input to further models.

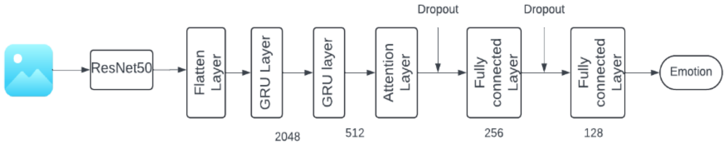

Figure 2. Architecture of Emotion Labelling model with pretrained ResNet50 and additional two bidirectional GRU and one Attention layer proceeding with dropout along fully connected layer.

B. Image Description Model

The Image Captioning Model is very crucial as it is responsible for generating the captions of images reflecting emotions perceived by them. By doing this, we aimed to enhance the description of images by extracting more word tokens. We employed BLIP (Li et al., 2022), a current state-of-the-art model for Multimodal Image Captioning due to its superior performance in generating diverse and descriptive texts aligning closely with visual information.

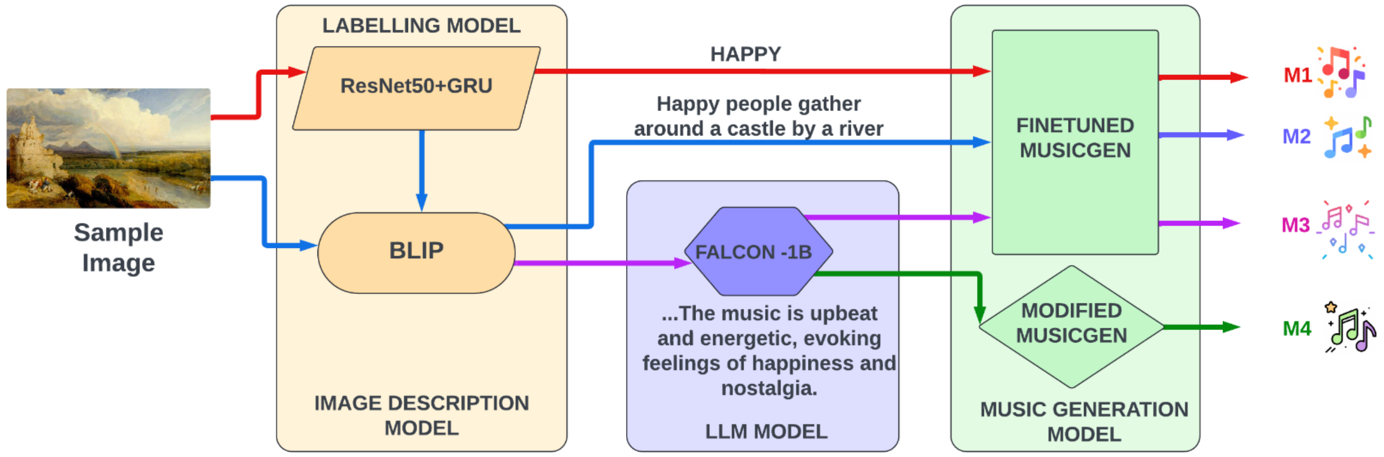

Fig. 3. Overview architecture of our model , which encompasses working flow of all the four models represented as different colour giving the output as M1 (single label text from ResNet50+ MusicGen), M2 (image descritive text from BLIP+MusicGen), M3 (Enhanced description from Falcon+ MusicGen), M4 (Enhanced description from Falcon+enhanced finetuning method of MusicGen).

The model is being conditioned on the emotion labels obtained from the emotion classification model (A) enabling it to give better relevant emotional description. The model is trained on a large amount of highly diversified data so we can directly incorporate the BLIP Large Captioning model to generate the caption.

Given an input image I, and emotion label E (extracted from model A), the model f_{BLIP} is generating captions that are further going to be used in our generation models

$$C_{caption} = f_{BLIP}(\{I|E\}) \qquad (5)$$

C. *LLM Model*

This model plays an integral role in the evolution of visual and musical information. It further enhances the description generated by the captioning model by incorporating some musical terms that reflect the mood, themes, and musical understanding terms that are very useful for generating the music. There is a need to enhance the description because the Image Captioning Model gives us descriptions based on image features such as objects, colors, and more, while the models for generating music need some music related component details in that to optimally perform. Providing this type of description leads to a better quality and resemblance of music. That’s why we are integrating the LLM model into our framework to fill the gap between the provided description and the expected inputs. The model not only works with the optimization of description but also ensures that the information about the visual image, especially the essence of emotion, is not lost.

The LLM model we have incorporated here is FalconRW-1B (Penedo et al., 2023) because it is fully open-source, and able to perform well even with restricted computational power. It consists of language modelling decoder-based architecture that is only incorporated with many advanced techniques like Attention. The input requirement of this model is characterized into three parts as shown in Table 1– system message (intent to tell the behavior of the model), instructions (intent to give the proper input), and response (the response model is providing).

Given a description C (extracted from Model B) of the image I, the model is intended to enhance the description.

$$S_{description} = f_{falcon}(C_{caption}) \qquad (6)$$

Role	Content
System Message	You are an enhanced description generator. You will be given with image description and you have to enhance those in musical terms.
Instruction	Generate a musical theme description for the following image description: "{description}". Include details like mood, genre, tempo, and melody in 2 lines.

Table 1. Description of input format given to falcon 1B model, emphasizing the system message and instruction, followed by a description provided.

D. *Music Generation*

The music generation model encompassed fine-tuning the MusicGen-small model for generating music conditioned on various text inputs. The model will work with the generated text files from each image to text model and its audio files. These respective files will be given to their respective enecoder model to encode them into tokens. Then condition_fuser of the model prepend the text vector along with applying the Attention mechanism on that. Later codebook interleaving patterns will be used by applying masking technique further employed by transformer model to generate the most resembling tokens. These tokens will be in the form of logits and masks. Further used to calculate the loss and refine the parameters and weights of model.

As reflected in Fig 3. Our experiment process was meticulous and thorough, involving the following versions of models to generate music.

Model 1: We used a single-label emotion description of images derived from the Image Emotion labelling model and fine-tuned all parameters of MusicGen.

$$M_1 = f_{MUSICGEN}(E_{emotion}) \qquad (7)$$

Model 2: We conditioned the image descriptions on emotional cues using an Image Description model, again fine-tuning all MusicGen parameters, thus embedding richer emotional context into the music.

$$M_2 = f_{MUSICGEN}\left(C_{caption}\right) \qquad (8)$$

Model 3: We enhanced the image descriptions using the LLM model further to have some musical knowledge as well and applied fine-tuning to all parameters of MusicGen.

$$M_3 = f_{MUSICGEN}\left(S_{description}\right) \qquad (9)$$

Model 4: Here, we introduced significant architectural innovations to optimize the MusicGen model. The training pipeline for fine-tuning initially involved a redundant and intensive process where each epoch required the pre-processing of music and associated text files using T5 and EnCodec models. This repeated pre-processing introduced instability in training and increased memory requirements due to constant data loading, a major reason contributed by our small dataset size. To look at this and provide a solution for this approach, we proposed a way to optimise the model by providing a revised training approach that streamlines the pre-processing of audio and its associated text inputs beforehand and stores it for further use. Additionally, precomputing these tensors reduces the dynamic computation, giving a fixed parameter to our transformer model. This way, we freeze the initial layer, making the training more stable and optimised. Additionally, the model works by randomly taking a chunk of 30s music to deal with it but fixing it in our code for the text description, making the model learn better about the data and consistent. One more notable architectural difference we made is that the original model works with dropping out some tokens from descriptions before pre-processing happens, but due to our exploration of text description, we are avoiding this so that we can analyse the model's performance more accurately on our dataset.

Thus, this variation in our experiment consisted of an enhanced description with a modified MusicGen version.

$$M_4 = f_{Mod-MUSICGEN}\left(S_{description}\right) \qquad (10)$$

V. Experimenets

A. DATASETS

Data collection and preparation are elementary steps in developing any model. Due to the lack of an existing paired dataset encompassing both art and music, which share the same emotional attribute, we proposed to make our own bespoke paired dataset integrating two different art forms-painting and music- while conveying the same emotion.

For the image paintings dataset, we utilized WIKIART EMOTION DATASET (Mohammad and Kiritchenko, n.d.), an openly available dataset of wiki art depicting various emotions. Wikiart is a collection of various paintings that evolved from different eras of past till present, symbolizing different art forms and meanings. These paintings have been analyzed and further categorized into more than 10 emotions in the emotion dataset. Based on these, I have manually analyzed datasets for particular 5 emotions - Happy, Angry, Sad, Fun, and Neutral and collected 1200 such paintings, which will serve as one part of our paired dataset. The manual selection process depicts the accurate representation of each dataset conveying the emotion.

The next part contains the collection of music datasets that should convey the same emotion. For this purpose, we selected MIREX EMOTION DATASET, a dataset of 193 MIDI files of music depicting various emotions. Initially, we preprocess the musical dataset from its raw form of MIDI and convert it into .wav form at 32KHz frequency making them compatible frequency for the MusicGen model. We further combined the emotional aspects of several parts to categorize it into five similar emotions as with paintings finally. Furthermore, as we are using a fixed 30s audio segment chunk in our music generation model, the music has been trimmed into various 30s without any overlapping between two different audios. By doing these preprocessing steps, we can generate more music samples that are uniquely identified from each other.

Later, to form the final paired dataset- the music and paintings have been associated with each other randomly, depicting the same emotions. In this way, we can have 1200 different pairs of paintings and music depicting five different emotions. Furthermore, for the training purpose of the model, we took around 80% of data from each section of emotion, with the remaining 20% split evenly between evaluation and testing sets.

B. EVALUATION METRICES

To evaluate the quality and resemblance of generated music, we evaluated it using a set of objective evaluation metrics. These metrics enable us to measure the quality, smoothness, noise, and distortion in the generated music further making us evaluate the model's performance on unseen data. Initially, we planned to conduct a subjective evaluation for the analysis of emotions of music, but due to the unforeseen situation, we were unable to perform the subjective evaluation part, thus leaving us to have a baseline qualification on objective evaluation only.

1. Frechet Audio Distance (FAD)

FAD (Kilgour et al., 2019), is based on the idea of FID which is calculated by comparing the statistical distribution of the generated music with a reference set in the embedding space. To extract the features of both generated and reference music sets, it uses the VGGish model. The mathematical function is:

$$F(N_r, N_e) = \left|\left|\mu_r - \mu_e\right|\right|^2 + tr\left(\sum r + \sum e - 2\sqrt{\sum r}\sum e\right)$$

Where represents the mean vectors of the reference and generated music feature distributions, respectively, and are the covariance matrices of the reference and generated music feature distributions. The lower the FAD score the better it is and more similar to the reference set.

2. Contrastive Language Audio Pretraining (CLAP)

CLAP score (Elizalde et al., 2022) is a metric that calculates the similarity between text descriptions and audio. We leveraged the pre trained LAION CLAP (Wu et al., 2022) model with its text encoder and audio encoder. Then, based on the embeddings extracted with the help of these encoders, we calculate the cosine similarity score between them. The formula for CLAP score is given:

$$CS(a,t) = \frac{\overrightarrow{e_a} \cdot \overrightarrow{e_t}}{\max\left(||\overrightarrow{e_a}||\ ||\overrightarrow{e_t}||, \in\right)}$$

Where $\overrightarrow{e_a}$ refers to the embeddings of the CLAP audio encoder and $\overrightarrow{e_t}$ refers to the embeddings of the CLAP text encoder. The CLAP score ranges between 0 and 1 and the greater the value of the CLAP score the better it is.

3. *Total Harmonic Distortion (THD) score*

THD is used to assess the quality of generated music contributing as one of the essential criteria. It is measured as the extent of harmonic distortion present in the generated signal. The formula is given by:

$$THD = \sqrt{\frac{\sum_{K=2}^{n+1} H_K^2}{H_1^2}}$$

Where H_k denotes the amplitude of the kth harmonic, the fundamental, and N is the number of harmonics.

4. *Signal-to-Noise Ratio (SNR)*

SNR is a metric, measured in decibels (dB), contrasts the strength of the intended signal with the strength of the background noise. When SNR is greater, the signal is more pronounced with less noise but when it is lower the noise iss more dominant. The formula for SNR is:

$$\mathrm{SNR\ (dB)} = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$$

Where P_{signal} is the signal power and P_{noise} is the noise power. It helps to assess the quality of signal by determining the noise present in it.

5. *Kullback-Leibler divergence (KL)*

KL score, is used to evaluate the difference between the probability distribution of two metrics. It further tells about the divergence of generated music features from the real music features. The lower value of KL tells us about the resemblance so the lower the value the better it is. The formula for the same is :

$$KL_{div}(P||Q) = \sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right)$$

Where P(i) is the probability distribution of the reference audio features, Q(i) is the probability distribution of the generated audio feature.

C. *TRAINING*

All our experiments were performed on the powerful JupyterHub platform, leveraging the capabilities of an NVIDIA A40 GPU with 48 GB of memory. This setup allowed us to perform training with a batch size of 16, using the small version of the MusicGen model over 40 epochs. We finetuned the model with the given hyperparameters which we determined after several runs and found these to be optimal. We have used an AdamW optimizer and a learning rate of 10e-5, a cosine scheduler with warmup steps of 100. During inference, we are taking the top_k as 250, selecting the top 250 most resembling audio tokens at a temperature of 1. The temperature parameter was changed to control the creativity of the model.

For the loss function during training, we observed the cross-entropy loss across logits for both the training and validation datasets. The cross-entropy loss for the training dataset exhibited significant fluctuations across the first three models. This behavior could be attributed to several factors. Firstly, the limited capacity of the small model might stop it from completely capturing the complex details during fine-tuning. Secondly, the variability of the dataset, which includes a variety of new data types that the model was not pre-trained on, may have contributed to these fluctuations. Thirdly, the descriptions provided are not in alignment with MusicGen's expectation. For this study, we limited the descriptions to simpler forms, which could explain the observed behavior from Model 1 to Model 4. However, due to our modified method, Model 4 showed a notable decrease in cross entropy loss with minimal fluctuations, setting it apart from the other models. This improvement highlights the importance of calibrated training and finetuning process when working with such large models. However, the decrease in validation loss across all models shows the model's ability to perform better on unseen data.

VI. Results and Discussion

Model	*FAD*↓	CLAP↑	KL↓	THD↑	SNR↓
Model 1	7.02	0.075	0.054	1.79	-4.29
Model 2	5.22	0.096	0.045	1.73	**-2.47**
Model 3	**5.06**	0.11	0.046	**1.92**	-3.43
Model 4	5.54	**0.13**	**0.012**	1.75	-3.57

Table 2. Objective comparison of test set data for emotion based image to music generation across all the models. Here the best results are made bold.

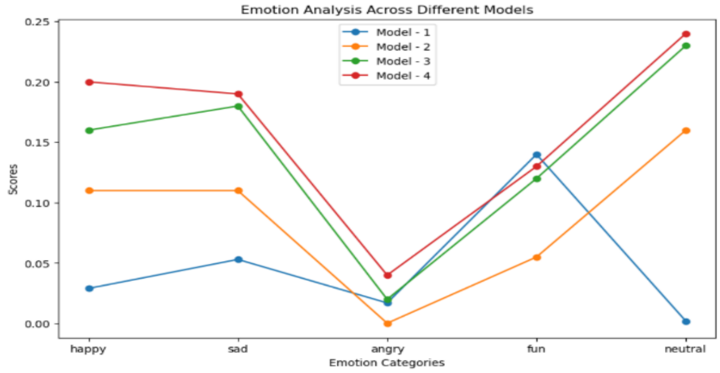

Fig. 4. CLAP analysis of generated song acroos model with their emotions. It is being meausred with providing "{emotion} song" as text during CLAP calculation

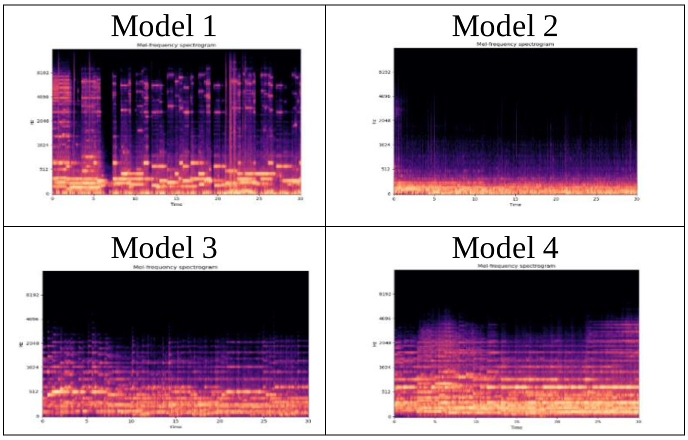

Fig. 5. Detailed spectogram analysis for example song generated from test set across all models.

The architectural experiments of all four models provide us with a critical evolution of model's behaviour on input description and the fine-tuning approach considered in MusicGen, as reflected by Table 2.

Model 1 represents the baseline approach in our study, whose architecture relies mainly on the use of ResNet50 combined with GRU architecture to extract the emotion labels from input images. Further, these single-word descriptions are used to finetune the MusicGen model. This straightforward approach showed significant limitations in the model's potential to depict the details of text descriptions, aligning them with music. The high FAD and KL scores suggested that the model struggled with distribution alignment, indicating a massive gap in the model's performance on simpler one-word text description files. Additionally, the average CLAP score implies the weak alignment of the text input and the music produced by the model; the reason is mainly contributed by the absence of emotional content in the input. Furthermore, the music generated contains more blank space and noise than other models, as evidenced by the spectrogram analysis, which indicates a lack of clarity and coherence.

Model 2 introduced a more sophisticated approach to captioning the image based on emotions with the pre-trained BLIP model. This gives an additional layer of semantic understanding and enables the MusicGen model to fine-tune the more detailed textual description. The enhancement in description is clearly reflected in improved CLAP score showcasing a better alignment between textual descriptions and generated music. Furthermore, the improved and more satisfactory SNR score above all the Models tells about the significantly less noise presence. However, it faces several challenges even after fine-tuning the model with more complex input sequences, as seen by moderate improvements in FAD and KL. However, the model was able to capture the emotional context of some music very well but lacked in the capturing of "Angry" emotion, indicating limitations in handling specific emotional tones.

Model 3 progressed our architecture with the incorporation of LLM model to enrich the descriptions with some musical contents adding one more layer of detailed relevance of description before the finetuning of MusicGen model. This includes prompt engineering method to provide a better prompt and output to the model. This enhancement in our architecture allowed the model to be conditioned on a more semantically suitable representation. With the incorporation of the LLM model, there has been a significant increase in FAD and CLAP scores, making the model's performance and capturing of tone and details more aligned with inputs. However, the increased complexity in descriptions required a more nuanced fine-tuning approach, as we can see a very good trade-off between KL and THD scores. While the model achieved higher alignment, it introduced some distortion, reflecting the challenges of balancing complexity and fidelity in generated music.

Model 4 represents the most advanced architecture in our study, combining the more contextually enriched descriptions along with a modified fine-tuning MusicGen pipeline. With this notable improvement, performance has not only been enhanced, but it also reduces the training time as it goes from 240s to 70s per epoch, which is approximately threefold. With the streamlined process, the model minimized the distortion and noise and maintained an excellent trade-off relationship with the description by attending the highest CLAP score among all the models. Talking about the performance of individual emotions, the model outperformed all by giving a better score as reflected in Figure 4. The spectogram analysis for this model as reflected in Fig 5. further highlights high fidelity and minimal distortion, making it the most effective architecture.

VII. Future work

The study indicates a comprehensive discussion addressing all our objectives, which we mentioned earlier. Although we are able to successfully generate music resembling closely to our specified emotions of paintings, the findings highlight the most crucial gap between the descriptions used to condition the Music Generation process. The first problem came from the availability of datasets used for training while we focused on enriching the diversity in the datasets; yet for fine-tuning such large models like MusicGen, we need vast number of datasets. Further, the investigation of behaviours with different labels gave us insight into how poorly the Model performs with a single label and non-musical descriptions, emphasising a massive gap between how end users convey information and the understanding of the Generative AI model. Addressing these limitations and reflecting on our study not only suggests room for improvement but also opens a new research area in text-conditioned generation models. Additionally, the evaluation of generated music in relation to its corresponding image or emotional content remains an area of exploration. Further developing specific evaluation metrics tailored for this context could significantly advance the field and help further in providing more concise assessments where we are dealing with a multimodality approach. The model is still a challenge for real time image to music generation due to its high inference time.

VIII. Conclusion

In conclusion, we have successfully developed a generative AI model capable of delivering music conditioned on the emotions conveyed in paintings. This work describes a remarkable step towards bridging the gap between two distinct modalities—visual art and music—while underscoring the societal importance of such models. Moreover, this paper shows a foundational baseline in the emerging field of generative AI models specially conditioned on emotional content.

By evaluating the generated music in terms of quality, diversity and noise present our study also talks about the critical gap between the prompts that models require to perform optimally, and the types of inputs typically provided by users. Addressing this gap is important for acquiring more diverse and high-quality outputs. Additionally, for optimal results we also suggested some improvements need to done in model process. Besides, the research underlines the concern of the limited availability of datasets, which are required for extending the creative prospect of art-music generation models.

Finally, our work points the potential of generative AI and deep learning models to drive innovation in creative fields, presenting a meaningful contribution to the ongoing development of AI technologies.

Acknowledgment

I would like to thank my supervisor, Jinhua Liang, for his expert direction and valuable insights throughout the duration of this project. His mentorship and motivation helped me to gain a very deep understanding of the subject and helped me to face the challenges of this project.

References

Agostinelli, A., Denk, T., Borsos, Z., Engel, J., Verzetti, M., Caillon, A.. Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N. and Frank, C. (n.d.). MusicLM: Generating Music From Text. [online] Available at: https://arxiv.org/pdf/2301.11325.

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

Briot, J.-P. (2020). From artificial neural networks to deep learning for music generation: history, concepts and trends. Neural Computing and Applications. doi:https://doi.org/10.1007/s00521-020-05399-0.

Briot, J.-P., Hadjeres, G. and Pachet, F.-D. (2019). Deep Learning Techniques for Music Generation -- a Survey. *arXiv:1709.01620 [cs]*, [online] 1. Available at: https://arxiv.org/abs/1709.01620.

Chen, K., Wu, Y., Liu, H., Nezhurina, M., Berg-Kirkpatrick, T. and Dubnov, S. (2023). *MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2308.01546.

Cheng-Zhi, A., Huang, Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A., Hoffman, M., Dinculescu, M., Eck, D. and Brain, G. (n.d.). *MUSIC TRANSFORMER: GENERATING MUSIC WITH LONG-TERM STRUCTURE*. [online] Available at: https://openreview.net/pdf?id=rJe4ShAcF7.

Chowdhury, S., Nag, S., J, J.K., Srinivasan, B.V. and Manocha, D. (2024). *MeLFusion: Synthesizing Music from Image and Language Cues using Diffusion Models*. [online] arXiv.org. Available at: https://arxiv.org/abs/2406.04673 [Accessed 19 Aug. 2024].

Colarusso, P., Kidder, L., Levin, I. and Lewis, N. (n.d.). *Raman and Infrared Microspectroscopy*. [online] Available at: https://nanoqam.ca/wiki/lib/exe/fetch.php?media=raman_and_infrared_microspectroscopy.pdf [Accessed 18 Aug. 2024].

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. The Journal of Machine Learning Research, 21(1):5485–5551, 2020

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

Dong, H.-W., Hsiao, W.-Y., Yang, L.-C. and Yang, Y.-H. (2017). *MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment*. [online] arXiv.org. Available at: https://arxiv.org/abs/1709.06298.

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

Google DeepMind. (n.d.). *WaveNet*. [online] Available at: https://deepmind.google/technologies/wavenet/.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). *Deep Residual Learning forImageRecognition*.[online]Availableat:https://arxiv.org/pdf/1512.03385.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, [online] 9(8), pp.1735–1780. doi:https://doi.org/10.1162/neco.1997.9.8.1735.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K. (n.d.). *Densely Connected Convolutional Networks*. [online] Available at: https://arxiv.org/pdf/1608.06993.

Ji, S., Yang, X. and Luo, J. (2023). A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges. *ACM Computing Surveys*, 56(1), pp.1–39. doi:https://doi.org/10.1145/3597493.

Kaunismaa, J., Qian, T. and Chung, T. (n.d.). *cMelGAN: An Efficient Conditional Generative Model Based on Mel Spectrograms*. [online] Available at: https://arxiv.org/pdf/2205.07319 [Accessed 19 Aug. 2024].

Kilgour, K., Zuluaga, M., Roblek, D. and Sharifi, M. (2019). *Fr\'echet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms*.[online]arXiv.org.doi:https://doi.org/10.48550/arXiv.1812.08466.

Lam, M.W.Y., Tian, Q., Li, T., Yin, Z., Feng, S., Tu, M., Ji, Y., Xia, R., Ma, M., Song, X., Chen, J., Wang, Y. and Wang, Y. (2023). *Efficient Neural Music Generation*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2305.15719.

Li, J., Li, D., Xiong, C. and Hoi, S. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. [online] Available at: https://arxiv.org/pdf/2201.12086.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W. and Plumbley, M.D. (2023). AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *arXiv:2301.12503 [cs, eess]*. [online] Available at: https://arxiv.org/abs/2301.12503.

Magenta. (n.d.). *Generating Long-Term Structure in Songs and Stories*.[online]Availableat:https://magenta.tensorflow.org/2016/07/15/lookback-rnn-attention-rnn.

Min, J., Liu, Z., Wang, L., Li, D., Zhang, M. and Huang, Y. (2022). Music Generation System for Adversarial Training Based on Deep Learning. *Processes*,10(12),p.2515.doi:https://doi.org/10.3390/pr10122515.

Mohammad, S. and Kiritchenko, S. (n.d.). *WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art*. [online] Available at: https://aclanthology.org/L18-1197.pdf.

Muradyan, H. and Muradyan, K. (n.d.). *Efficiently Fine-Tuning MusicGen For Text Conditioned Armenian Music Generation*. [online] [Accessed 19 Aug. 2024]

OpenAI (2023). GPT-4 Technical Report. *arXiv:2303.08774 [cs]*. [online] Available at: https://arxiv.org/abs/2303.08774.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E. and Launay, J. (2023). *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only*. [online] arXiv.org. Available at: https://arxiv.org/abs/2306.01116.

Payne, C. MuseNet. OpenAI Blog 2019, 3

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*. [online] Available at: https://arxiv.org/abs/2103.00020.

Riffusion.(n.d.). *Riffusion*.[online]Availableat:https://www.riffusion.com/.

Roberts, A., Engel, J., Raffel, C., Hawthorne, C. and Eck, D. (2019). A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. *arXiv:1803.05428 [cs, eess, stat]*. [online] Available at: https://arxiv.org/abs/1803.05428.

Schneider, F., Kamal, O., Jin, Z. and Schölkopf, B. (2023). *Mo\^usai: Text-to-Music Generation with Long-Context Latent Diffusion*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2301.11757.

Sheffer, R. and Adi, Y. (2022). *I Hear Your True Colors: Image Guided Audio Generation*. [online] arXiv.org. Available at: https://arxiv.org/abs/2211.03089 [Accessed 19 Aug. 2024].

Todd, P.M. (1989). A Connectionist Approach to Algorithmic Composition. *Computer Music Journal*, 13(4), p.27. doi:https://doi.org/10.2307/3679551.

Vinet, H. (2004). The Representation Levels of Music Information. Lecture notes in computer science, pp.193–209. doi:https://doi.org/10.1007/978-3-540-39900-1_17.

Wang, X., Gao, Z., Qian, H. and Xu, Y. (2018). Jukepix: A Cross-Modality Approach to Transform Paintings into Music Segments. doi:https://doi.org/10.1109/robio.2018.8665063.

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

Xu, T., Li, J., Chen, X., Yao, X. and Liu, S. (n.d.). Mozart's Touch: A Lightweight Multi-modal Music Generation Framework Based on Pre-Trained Large Models Input Text Image Video LLM Understanding Bridging Module BLIP Sampling Image Set Image description Multi-modal Captioning Module Prompt Builder LLM Semantic Optimization Music description Music Generation Module MusicGen Generated Music. [online] Available at: https://arxiv.org/pdf/2405.02801 [Accessed 19 Aug. 2024].

Yang, L.-C., Chou, S.-Y. and Yang, Y.-H. (2017). MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation. *arXiv:1703.10847 [cs]*. [online] Available at: https://arxiv.org/abs/1703.10847.

Yu, L., Zhang, W., Wang, J. and Yu, Y. (2017). SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *Proceedings of the AAAI Conference on Artificial Intelligence*, [online] 31(1). Available at: https://ojs.aaai.org/index.php/AAAI/article/view/10804.