Spring 5-3-2023

Music On Canvas: A Quest to Generate Art That Evokes the Feeling of Music

My Linh (Lucy) Tran
Macalester College, lucytranorchestra@gmail.com

Macalester College | Department of Mathematics, Statistics, and Computer Science
Bachelor’s thesis | Computer Science Honors Project
2023-04

Music On Canvas

A Quest to Generate Art That Evokes the Feeling of Music

My Linh (Lucy) Tran

Advisor : Elizabeth Shoop
Committe : Elizabeth Shoop, Susan Fox, Shilad Sen

Macalester College
+1 651-696-6000
www.macalester.edu

Abstract

Although the idea of connecting music and art dates back to ancient Greece, recent advancements in computing have made automating this feasible. This project represents a quest to transform music into art, using three methodologies where each is an improvement towards generating images that convey our feelings and imaginations during music listening. The three methods respectively involve:

1. An element-wise mapping of sound and colors
2. Using song tags
3. Tuning an Artificial Intelligence (AI) model to generate pictorial text captions.

To create artistic images, methods two and three utilize an existing text-to-image generative AI.

Contents

List of Figures

List of Tables

1 Introduction

Music and art have been two of the most primal forms of entertainment. Their materials are raw: essentially, music is made from the sound and speech we hear every day, and art is composed of colors we see in the world around us. It is natural, then, to relate these two forms of art. The thought that there may be some relationship between sounds and colors dates as far back as ancient Greece when Aristotle and Pythagoras speculated that there must be a correlation between the musical scale and the rainbow spectrum of hues [1]. The intuition behind it, I suppose, is that both the number of notes in the heptatonic scale, the most frequently used scale in modern Western music [2], and the number of colors visible to the naked eye in a rainbow are seven. Over time, with the development of art and music, many artists of one medium have used the other medium to inspire their work. Some examples are Jackson Pollock, Paul Klee, Piet Mondrian, and Jean Pederson, artists who confirm using music as an input for their creativity. Vice versa, in [3], we see several instances of paintings turned into classical masterpieces. Yet, the best instance of the harmonious relationship between sound and image lies in our imagination. Some images remind us of a song, and some songs, even instrumental ones, draw pictures in our minds.

It is in light of this intrinsic connection that I started this project. The goal is to find the best computational method that, given a musical piece, will create an artwork that evokes the feeling of the music. As a brief overview, the project covers three different methods, where each is an improvement from the lessons learned in the previous method. The first method is based on an element-wise mapping from music to art, which is similar to the concept of "visual music" explored in the 1990s-2000s. The second approach is inspired by the use of song tags in interpreting music, popularized in the 2010s by the release of the Million Song Dataset in 2011. Finally, the third approach employs modern machine learning models and techniques, including CNN, Transformer, and contrastive learning, trained on a dataset of audio-text pairs to learn music representation in vector format. As part of the third method, I also released a small dataset called Goma's Imagination (or for short, Imagination), which is currently used only as a test set, but demonstrates the ideal training and validation sets for future improvement.

There are many research projects with similar objectives, which I will discuss in the "Background" section of each method. To make it clear, there are three important differences between this project and others that will be mentioned:

1. Despite the changing goals in the three methods, the direction is always "from music to art" and not vice versa. Additionally, considering my background and the scope of this project, the process in each method is generally (1) inputting a piece of music, whether in the form of an audio file, the name of a song, or in binary format, (2) processing the music, which takes some delay time, and (3) outputting a 2D image or images.

2. The output image(s) is generated and is not an existing artwork, which may violate copyright issues, and limit the scope of possible outcomes.

3. Rather than focusing on a specific method, this paper describes a 3-phase journey of experimentation, adjusting, and refining the ultimate goal. As such, each method has a slightly different goal reflecting the interpretative transformation of the research question. Since the time spent on a single method is limited, there is not much fine-tuning in each method, and for the same method, there is space for possible improvement that has not been investigated.

As a result of the nature of this project, the paper is divided into three main parts, each for a method. In each method, I will introduce some background or related work, detail the method's implementation, and show the results followed by a discussion, which leads to the next method and/or future improvements.

2 Method 1: Converting Musical to Art Elements

2.1 Background

Curiosity about the connection between sound and color has existed since ancient Greece. Yet, it was not until 1734 that the first instrument to experiment with such aspiration existed when Louis-Bertrand Castel invented the *Clavecin Oculaire*, known as the first "color organ." Castel's work inspired many successors to create variations of the color organ or other types of color-music instruments. A basic mechanism of these instruments is the augmentation of a keyboard-based instrument with a light projection system so that lights of different colors are projected into the air or some background when someone plays the instrument. Despite their different implementations, the main idea is that variations in luminosity could somehow parallel nuances in music [4].

This trend continues to the computational era, especially after the flourishing of real-time computer graphics. With the increasing capability to control, even animate, the details of graphics shown on the computer screen, engineers have invented various mapping methods and applications to show an animation of colors and shapes varying according to the music. For example, the 21st Century Virtual Color Organ creates a performance space in an interactive, immersive virtual reality environment, with 3D visual images and sound generated based on MIDI files (see Figure 1) [5]. Lee et al. [6] even created a new unit analogous to pixel called soxel, "the smallest discrete component in a given representation of multi-dimensional auditory space." Thus, generating sound from an image and vice versa would be just a matter of converting two units (see Figure 2). There are also papers focusing on image sonification, which is the reverse direction of the focus of this paper, such as [7] and [8]. Leveraging the self-reproductive nature of cellular automata (of which an example would be Conway's Game of Life), several papers have investigated mapping methods to help inspire music composition from the model. Examples of such papers are [9–11].

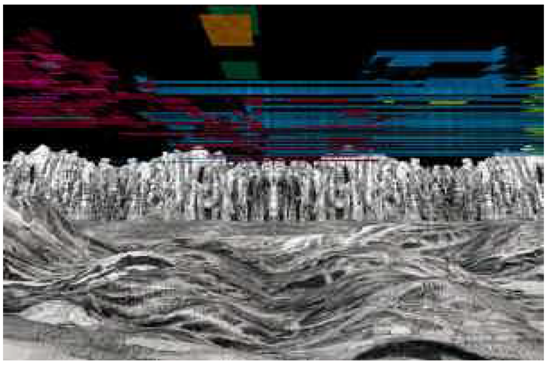
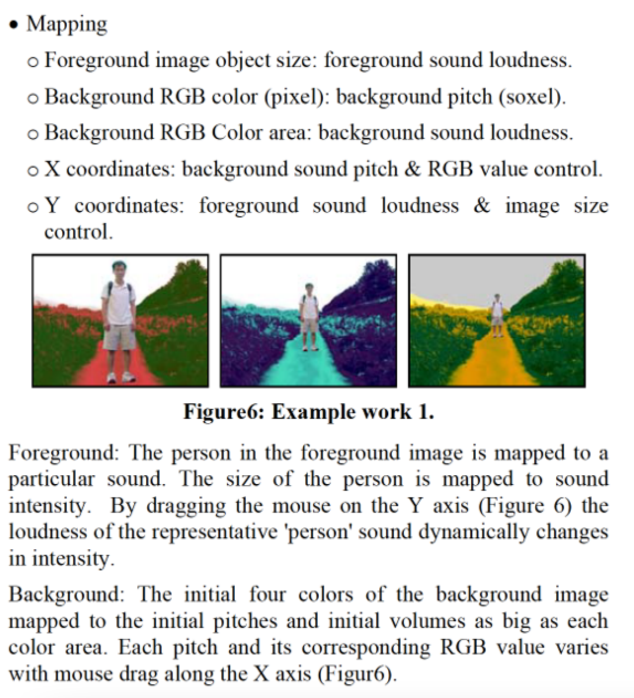


Figure 2.1: A Still-shot with Music from the Virtual Color Organ. Caption adapted from [5].



- Mapping
  - Foreground image object size: foreground sound loudness.
  - Background RGB color (pixel): background pitch (soxel).
  - Background RGB Color area: background sound loudness.
  - X coordinates: background sound pitch & RGB value control.
  - Y coordinates: foreground sound loudness & image size control.

Figure6: Example work 1.

Foreground: The person in the foreground image is mapped to a particular sound. The size of the person is mapped to sound intensity. By dragging the mouse on the Y axis (Figure 6) the loudness of the representative 'person' sound dynamically changes in intensity.

Background: The initial four colors of the background image mapped to the initial pitches and initial volumes as big as each color area. Each pitch and its corresponding RGB value varies with mouse drag along the X axis (Figur6).

Figure 2.2: A Mapping from Image Features to Audio Features [6].

However, my initial inspiration did not come from these papers. The inspiration came from a realization back in 2018 when I looked at Jackson Pollock's paintings:

"I dissected his paintings to unravel the mystery behind this enthralling multilayered tumult of paint: some dominant tones on a deep background, erratic lines of varying width, long straight threads knitted with brisk whirls, spots and dots sprinkled playfully, common patterns hidden delicately. The more I observed his paintings, the more familiar I found them. The words 'motif,' 'unique,' 'thickness,' 'range,' 'emotional,' 'contrapuntal,' and 'improvisational' popped into my head - words that also described music! The paintings really looked like a song!"

Music in Canvas Form, T. My Linh [12]

The experience prompted me to conceptually match the elements of the two media forms.

Figure 2.3: Jackson Pollock, *Convergence*, 1952

"Timbre should coincide with color because if we thought of different instruments as unique colors, we would achieve varied shades by blending and mixing the originals – just like the application of orchestration principles. I paired pitch with the tone of colors. The higher notes of an instrument would produce a brighter tone of a color. Dynamics could be best coupled with brush size – to emphasize a sound or a line, we usually put more force into it. Brushstrokes would be determined by note duration because both have significant impacts on patterns. Long, legato notes would be presented as long lines; shorter notes as curves, and staccato notes as dots."

But at that time, all of this is merely a thought. Past research mentioned above is what motivated me to work on this first method. Among the papers I have read, one particular paper caught my interest. In [13], Ciufo proposes a matching plan that is similar but much more intricate than mine:

"These mappings will vary greatly from piece to piece, depending on the specific concept or intent of the work… Also of great importance is how these mappings might change or evolve over the course of the piece. The following illustration shows a possible mapping matrix with the ability to connect any element to any other element (or elements) in either direction."

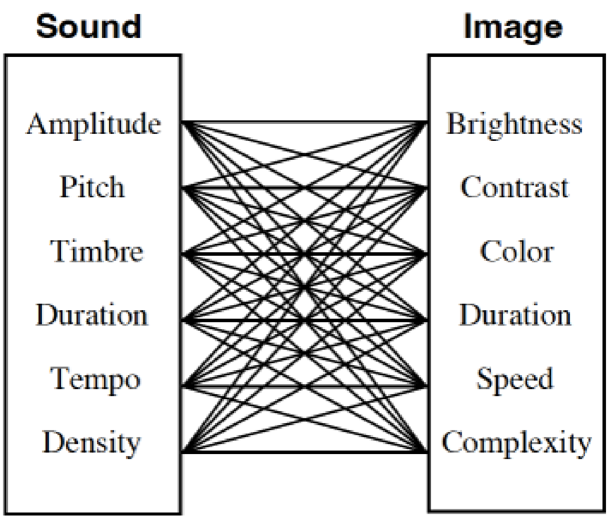


Figure 2.4: A Proposed Mapping Matrix with the Ability to Connect Any Element to Any Other Element (or Elements) in Either Direction [13].

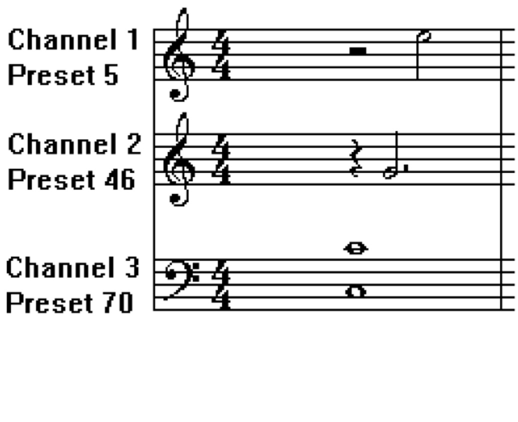
However, the author never actually implemented this matrix in a software application. Driven by curiosity, I started working on a simple, one-to-one mapping as a proof of concept. At this point, there was no clear direction as to how I could transform music into art. Therefore, I planned that if this proof of concept produces desirable images, I would continue to level up the mapping complexity.

2.2 MIDI Basics

One crucial step in this process was learning MIDI, which stands for Musical Instrument Digital Interface, a language for computers to understand and communicate music using "MIDI messages" [14]. A language in binary format, MIDI has solid rules defining the structure of different types of MIDI messages. These messages are not the sound itself, but a sequence of instructions to create the sound in the target synthesizer. The details of these instructions are best described in [15–17]. Here, I will outline some MIDI basics that are necessary for understanding this method:

- A MIDI file consists of chunks, which can be header chunks or track chunks. Header chunks store general information, such as tempo, time signature, file format, etc. about the file, or a section of the file. On the other hand, track chunks store actual musical data, such as note pitch, note start/end time, the musical instrument, the channel, etc.
- Each track chunk is a stream of MIDI events. A MIDI event is a MIDI message combined with a preceding delta-time value that signals when the message is to be sent.
- The delta-time value is how many ticks (60 ticks = 1 second) the machine needs to wait before executing the event. For example, if two events happen simultaneously, the delta-time value of the second event will be 0.
- There are different types of MIDI events. Two important types are Note On and Note Off. For example, when a synthesizer encounters a Note On (or Note Off) event, it will play (or stop playing) the note specified by the event. Both of these events include information about the channel, the pitch, and the velocity (volume) of the note.
- Another important type is the Control Change events, which assign an instrument/sound to a channel. A MIDI program can have up to 16 channels. Each sound is assigned a numeric value designated by General MIDI [18]. Control Change events are typically at the beginning of the stream, before the first Note On event.

Page 12-14 of [17] provides an example of how a musical bar is translated into MIDI events:



Delta Time (dec)	Event Code (hex)	Other Bytes (dec)	Comment
0	FF 58	04 04 02 24 08	4 bytes: 4/4 time, 24 MIDI clocks/click, 8 32nd notes/24 MIDI clocks
0	FF 51	03 500000	3 bytes: 500,000 µsec per quarter-note
0	C0	5	Ch. 1, Program Change 5
0	C1	46	Ch. 2, Program Change 46
0	C2	70	Ch. 3, Program Change 70
0	92	48 96	Ch. 3 Note On #48, forte
0	92	60 96	Ch. 3 Note On #60, forte
96	91	67 64	Ch. 2 Note On #67, mezzo-forte
96	90	76 32	Ch. 1 Note On #76, piano
192	82	48 64	Ch. 3 Note Off #48, standard
0	82	60 64	Ch. 3 Note Off #60, standard
0	81	67 64	Ch. 2 Note Off #67, standard
0	80	76 64	Ch. 1 Note Off #76, standard
0	FF 2F	00	Track End

Figure 2.5: A Score Excerpt and Its Corresponding MIDI Stream [1].

2.3 Methods

2.3.1 Conversion Formula

As the first phase, the goal of this method is to try out music visualization using a simple mapping from MIDI data to their corresponding art data. The table below shows an example conversion formula.

Music data	**Art data**	**Explanation**
Instrument	Hue	Each instrument should have a color, so that an orchestral piece when converted to art looks like a mixing of different colors. Example: Piano → Blue vs. Flute → Orange
Pitch (note)	Brightness	Lower notes sound darker, while higher notes sound brighter. Example: Piano C2 → vs. Piano C5 →
Dynamics (essentially volume)	Saturation	Whatever color and brightness a sound possesses, its volume determines a lot how clear and vivid we hear that sound. Example: Piano C5 forte (f) → vs. Piano C5 piano (p) →
NOTE-ON event tick	Starting cell	Tick #96 → Cell #3
NOTE-OFF event tick	Ending cell	Tick #192 → Cell #6

Table 2.1: Method 1's Mapping from Musical Data to Art Data.

Other formulae are also possible. For example, when a track only has 1 instrument, Hue can be defined by Pitch in order to highlight the differences in pitch and make the resulting image look more colorful instead of monochromatic.

2.3.2 Implementation Notes

Because all musical data and art data are numeric, and their values always stay within a range, conversions are done by first scaling the numbers according to their ranges. For example, pitch in MIDI can range from 0 to 127, and brightness in the HSB system can range from 0 to 100. A basic conversion would turn a note at pitch number 96 into a pixel with the brightness level of (96 / 127) * 100 = 76 (rounded up to the nearest whole number). However, in most cases, the pitches and dynamics of a song do not span all possible MIDI values. Of all instruments, the piano has the broadest pitch range, with 88 black and white keys in total. Hence, to maximize the brightness difference, I divide the pitch number by 88 instead of 127, before multiplying it by 100.

Each note in the song will correspond to a pixel on the resulting image. Essentially, the program will draw pixels on an x-y coordinate system based on the musical data. The x-axis indicates time, and the y-axis shows notes of different instruments. Each instrument has its section of row(s). The sections are stacked on top of each other, creating a grid of cells of RGB values.

In MIDI, once instruments are assigned to specific channel numbers at the beginning of the file, the channel number will represent the instrument throughout the rest of the event stream. A channel's width is the least number of rows the channel needs. Because an instrument can play multiple notes simultaneously, assigning one row to each channel will cause overlapping of notes and flawed images. When two (or more) notes are played simultaneously, the note whose Note On event first appears in the event stream will take the lowest row. This ensures that no two notes overlap on the grid.

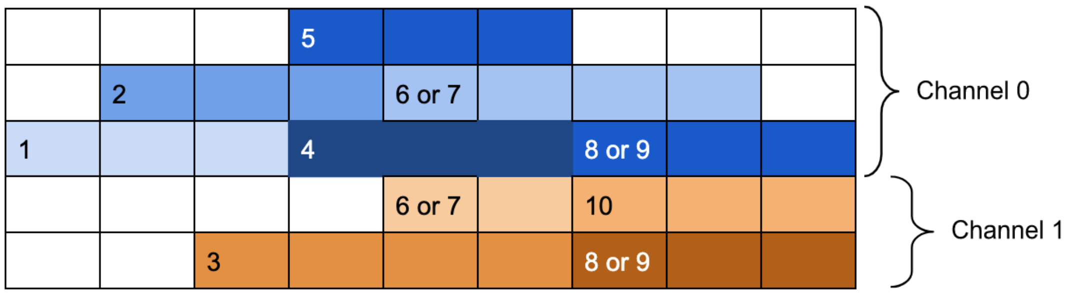


Figure 2.6: Conversion of MIDI Notes to Pixels.

The figure above shows an example of how MIDI notes are displayed on the grid of pixels. The numbers are the inferred order of the Note On events, based on the output image. Even though notes 4 and 5 start at the same time (they start in the same column), note 4 takes the lower row of channel 0. This means that in the event stream, the Note On event of note 4 appears before that of note 5. Hence, note 4 takes the lower row of the same column. Note 5 could have taken the second row (bottom-up) of channel 0. However, because that cell is already occupied by note 2, it must take the third row. For notes that start at the same time but are from different channels, there is no implication as to which one of them appears first in the event stream. Hence, I leave the orders as "6 or 7" and "8 or 9."

The program is written in C++ 11, compiled by GNU g++, and runs on the mscs1 machine. Where there is parallelization (more on this later), I use OpenMP to create more threads and manage them.[1].

[1] Code for this method is available at https://github.com/lucy-tran/MIDI-to-Pixels.

2.3.3 Pseudocode

$channelSounds \leftarrow \{\}$// an empty dictionary
$channelWidths \leftarrow \{\}$
$colorMap \leftarrow [][][]$// a 3D array
$firstNoteOn \leftarrow null$
$fileDuration \leftarrow midifile.fileDuration$

for $event$ in $eventStream$ **do**: //first for-loop
    if $event.type$ is Control Change **then**: // assign a sound to a channel
        $channelSounds[event.channel] \leftarrow event.instrument$
        $channelWidths[event.channel] \leftarrow 1$
    else if $event.type$ is NOTE-ON **then**:
        $firstNoteOn \leftarrow event$
        Break
    end if
end for

//Some MIDI files do not have Control Change events.
if $channelSounds.size$ is 0 || $channelWidths.size$ is 0 **then**:
    $channelSounds[0] \leftarrow 1$ // instrument 1 is "acoustic grand piano"
    $channelWidths[0] \leftarrow 1$
end if

//File duration (in ticks) can be very large, so I scale the file duration down to 500 to avoid running out of memory.
$scale \leftarrow fileDuration/500$
$fileDuration \leftarrow 500$

for $event$ in $eventStream[firstNoteOn:]$ **do**: // second for-loop
    if $event.type$ is NOTE-ON **then**: // all non-NOTEON events are skipped
        $currentChannel \leftarrow event.channel$
        $startingX \leftarrow event.startingTime/scale$
        $endingX \leftarrow startingX + event.duration/scale$

        $y \leftarrow 0$
        while $colorMap[channel][y][startingX] \neq null$ **do**:
            $y \leftarrow y + 1$
            if y is $channelWidths[currentChannel]$ **then**:
                $channelWidths[currentChannel] \leftarrow channelWidths[currentChannel] + 1$
            end if
        end while

// convert musical elements to art elements
        $hue \leftarrow (channelSounds[event.channel]/360) \times 128$ // instrument to hue (H)
        $saturation \leftarrow (event.velocity/127) \times 100$ // velocity (volume) to saturation (S)
        $brightness \leftarrow (event.pitch/88 \times 100)$ // pitch to brightness (B)

// convert HSB to RGB color, because GNU Plot needs RGB
        $rgbColor \leftarrow$ rgb color from (hue, $saturation$, $brightness$)
        $colorMap[channel][y][startingX : endingX] \leftarrow rgbColor$
    end if

end for

// Write the data file for GNU Plot in the x:y:R:G:B format
for *channel* in *colorMap* **do**:
  for *row* in *colorMap*[*channel*][0 : *channelWidths*[*channel*]] **do**:
    for *column* in *row* **do**:
      $dataFile \leftarrow dataFile + column : row : colorMap[channel][row][column]$
    end for
  end for
end for

2.3.4 Parallelization

Because this method was originally a project from my Parallel and Distributed Computing class, I also implemented parallelization on the second for-loop, which takes the majority of the total runtime. Essentially, this for-loop converts musical notes into color values and saves them to a 3D map. Parallelization is performed using parallel loop equal chunks, such that each thread will work on a chunk of size n/k of the event stream, where n is the number of Note On events, and k is the number of threads. The graph below shows how increasing the number of threads speeds up the runtime of the second for-loop at different problem sizes.

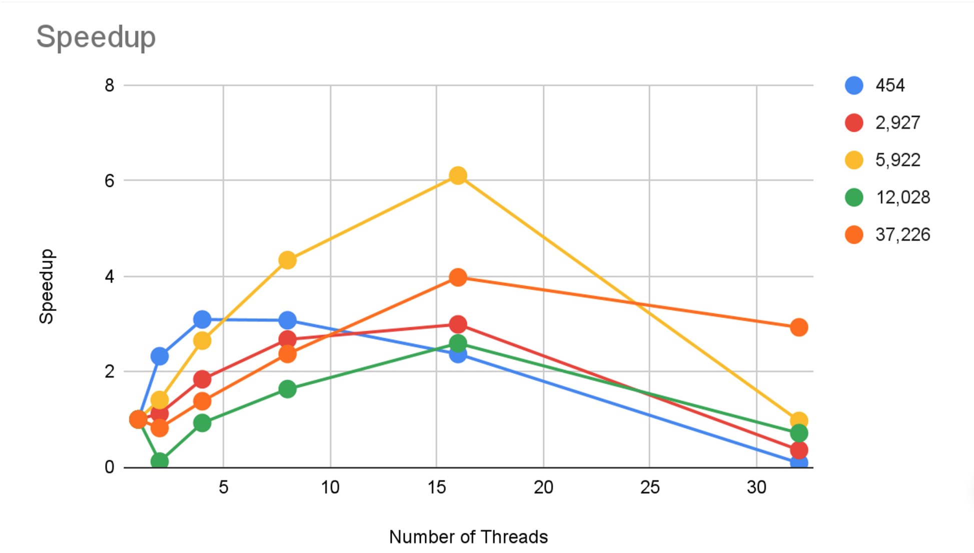


Figure 2.7: Speedup at Different Numbers of Threads for 5 Problem Sizes.

The problem size here indicates the total number of Note On events. When the problem size is too small, having more than four threads counteracts the purpose of parallelism. For large enough problem sizes, however, the speedup still grows as we increase the number of threads up to 16. At 32 threads, most problem sizes only experience worse performance than when run sequentially. This is because of the overhead cost of coordinating work among too many threads.

Increasing the number of threads also hurt accuracy in some cases. Consider a song track of size 100, in which events 50 and 51 are both Note On and have an overlapping section (note 51 starts before note 50 ends). When the event stream is sequentially processed, the resulting image would have event 50 at a lower row than event 51. However, when the program is run by 2 threads in parallel, event 51 would be processed by thread 2 before event 50 by thread

1 (which should be working on event 1), causing the program to draw event 51 at a lower row than event 50. Similarly, for other numbers of threads, accuracy problems can arise at split points. The figure below illustrates how the end visualization differs as the number of threads increases. The red box pinpoints an example point where inaccuracy happens.

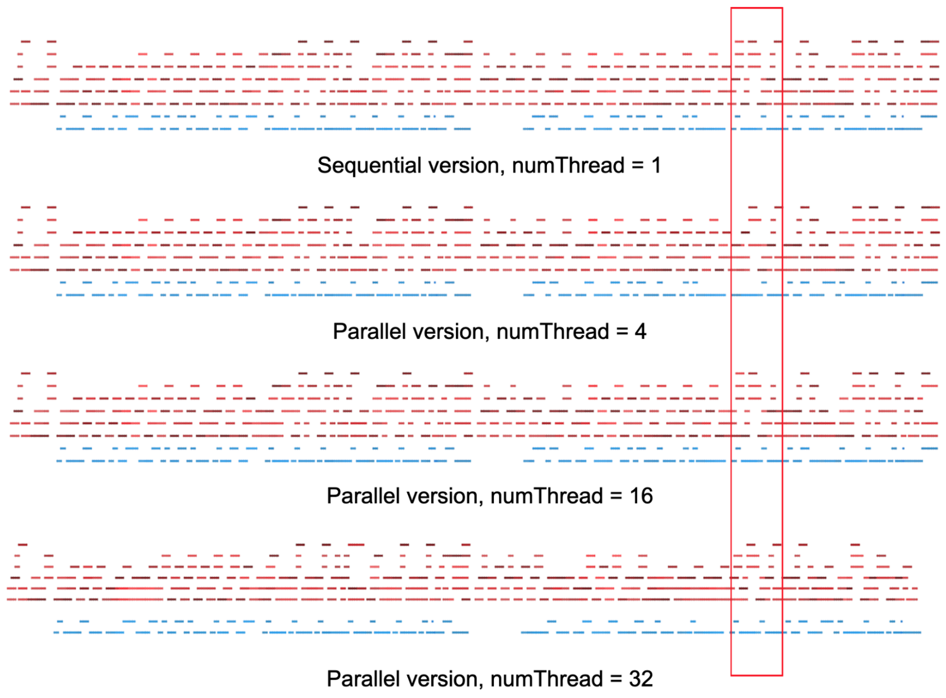
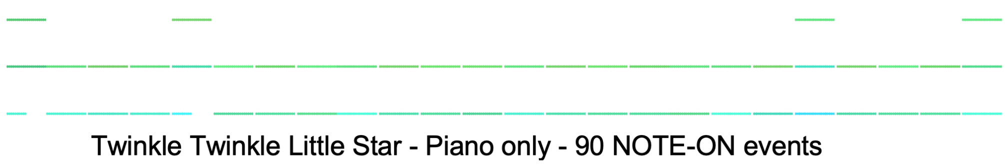


Figure 2.8: Visualizations of the Same Song by Different Program Versions.

It is worth noting that the current largest problem size (37,226), when run sequentially, only takes around 0.07 seconds. Although there is a tradeoff between accuracy and speedup, it is up to subjective preference to decide which `numThread` is best, as both inaccuracy and runtime are little. I decided to run the program with 8 threads, because for all problem sizes, the speedup is still more than 1 at 8 threads, and 8 splits can only cause at most 8 erroneous windows. However, If accuracy is of higher importance, one may opt for simple sequential execution.

2.3.5 Results

The images below show visualizations of songs by this method, run by 8 threads. For each song, I made some minor adjustments in the conversion formula to intensify the expression. Audio samples for these songs are available in the GitHub repo, under the `bin/audio_samples` folder.



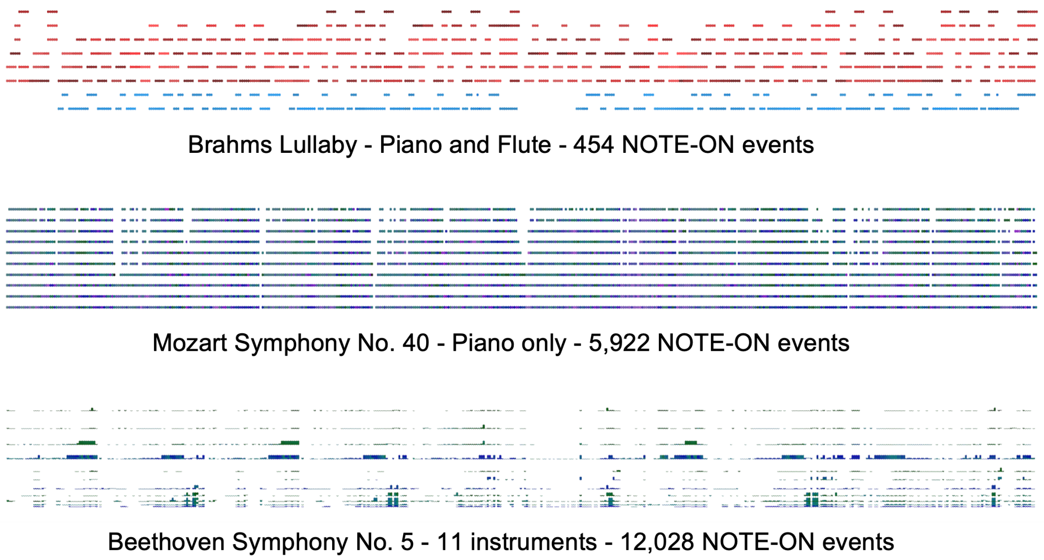


Figure 2.9: Results from Method 1.

2.3.6 Discussion

There are many possible improvements to this method. For example, zooming in and sliding in the image from right to left as we listen to the song, much like a MIDI player, would be a fun experience. The different colors currently do not bring any particular feelings that the notes convey. Instead of letting pitch and volume directly affect brightness and saturation, adding a chord progression analyzer to infer the temporal feelings and adjust these two elements accordingly may be a better approach. I can also define the hue associated with different instruments based on the overarching mood of the song. Most importantly, there needs to be a more artistic, flexible way to control the coordinates of the pixels, such that they do not always stay on a line. This step is important in making the resulting images look less rigid and more like artwork. Currently, the images generally portray the complexity of the musical pieces by their number of notes, instruments, and length. However, I believe that with the mentioned improvements, the images will render more aspects of the input music.

Even so, there are no criteria for which mapping algorithm is the best. An inherent issue in any mapping approach is well-defined by Levin [4]:

"One more theme which has cut across more than four centuries of color-music research, from Castel's to that of the present day, is the question as to whether there are any 'absolute' correspondences between sound and vision. It is one of the deepest issues in the field; some have felt that it is best answered through empirical studies in psychology, while others have denied the possibility of any essential mappings, and instead held that the matter is simply an aesthetic one, best handled by design. The truth is almost certainly somewhere in between"

The fact that no one knows exactly "the truth" means that the assessment of any mapping formula, regardless of how improved it is, is subject to personal taste. Indeed, those that align more with human psychology will generally gain more positive impressions. Yet, to make a person feel certain emotions, the emotions that a piece of music evokes, by only looking at abstract lines, shapes, or colors is an uncontrollable task.

Levin then goes on to point out another problem with existing mapping systems:

"It is a regrettable circumstance that most of the systems which use image and sound together have focused on only one of these dimensions, to the detriment of the other... although the sound may be extremely malleable, the imagery is rigidly constrained by a strict visual language and a pre-determined formal design"

This is certainly the case with my program which converts MIDI to pixels. The one-to-one conversion formula itself is an indication of "a strict visual language." Increasing the complexity of the formula by having each musical component affect multiple art components (as planned in the beginning) only injects more black-box regulations around how the art should appear. Compare this to an artist whose artwork is inspired by music, the inspiration is very subjective, unbounded, and constantly alterable. Even for a layman who does not know any music or art theories, the images that come to their mind during music listening, I suppose, do not come from layers of pre-defined recipes. If there are any abstract visual elements, such as different shapes, colors, or lights, in their imagination, those would come more from their past experiences with these elements. These experiences then unconsciously shape the connection between music and art that many think of as an "absolute correspondence."

Unfortunately, past experiences vary from person to person. I would propose that this is why many attempts to create abstract visualizations of music fail to retain attention: they are not relatable to the audience. Indeed, they are interesting, even mesmerizing, and created by ingenious minds. However, these visualizations are not what the audience would imagine as they listen to music. Take the examples of these eye-catching audio visualizers:

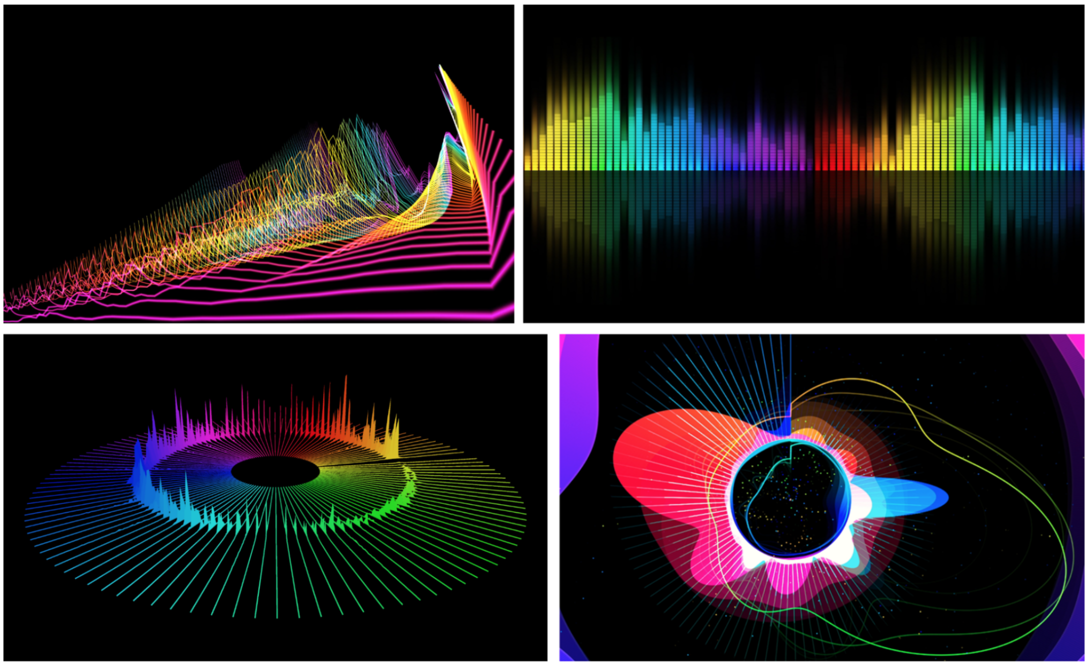


Figure 2.10: Screenshots of Existing Audio Visualizers [19–22].

Or these "audiovisual substances" created by Levin in the same paper [4]:

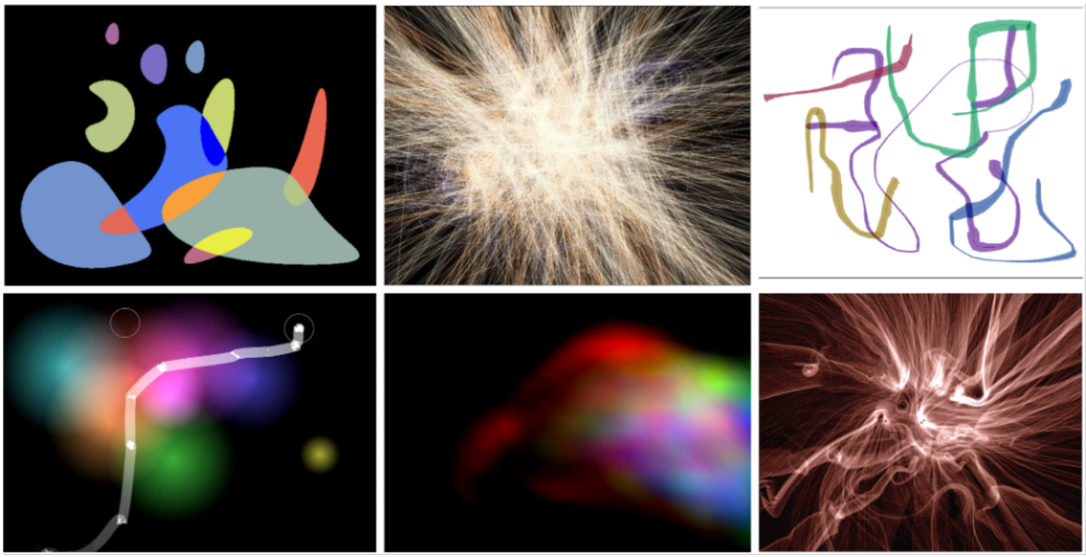


Figure 2.11: Still Captures of the Audiovisual Substances Created by Levine [4]. From left to right, top to bottom: *Polygona Nervosa*, *Directrix*, *Loom*, *Warbo*, *Aurora*, *Floo*.

They are beautiful and artistic. Yet, few people see these visualizations as they close their eyes and listen to music.

Going back to my method, based on the discussion thus far, its results are not desirable. While the images generated are not even close to par with the existing visualizations above, its failure lies in a much deeper problem: the abstraction that is unfamiliar to and unrealistic to expect from a general audience. However, this method is core to redirecting my goal. It prompts me to question whether art that influences our feelings or art that is based on our feelings would better serve the overall goal, that is, to create art that evokes the feeling of music. While each of these two directions includes some aspects of the other, the disparity is in the priority, the inspiration source of the art. While the former fixates on fine-tuning the art, the latter calls attention to the end results, the feeling. As the feeling aspect is prioritized, there is less focus on the meticulous design of formulae and artistic elements. The focus now shifts to *realisticity* and *interpretability*, because these characteristics are what ensure that the art generated is *relatable* to the audience, thus kindling emotions and emphasizing their music-listening experience. This realization shapes the direction of my next steps.

3 Method 2: Generating Art From Song Tags

3.1 Background

The goal of this method is to produce real-life images that match the input music based on its feelings and descriptions. One approach that has been well-studied is recommendation systems, which recommend one or more paintings existing in the database for a given song or vice versa. The architectures behind these systems vary widely. For example, Lee et al. [23] match music and paintings based on their Euclidean distance on A.V. coordinates, where A stands for Arousal, and V stands for Valence. For paintings, the A.V. scores are computed based on emotional adjectives that are extracted from the most-used three-color combination. For music, the authors perform linear regression on the songs' acoustic features to predict the annotated emotional responses (A.V. scores). Li and Shan [24] apply a similar approach but rely on clustering and graph algorithms to suggest fitting music for a slideshow of Impressionism paintings. Another interesting approach utilizes image style transfer to transform original images to the style of emotional reference images so that the transformed images match the emotion categories of the input music [25].

While these methods satisfy the current goal, they all work on existing photos. Although the database of pictures can be up to tens of thousands, the recommendation approach still limits the possibility of finding a picture that users find suitable. More importantly, these methods work on the assumption that colors and styles determine the viewers' reaction to a picture the most. Thus, the models in [23, 24] are trained to extract the emotions of pictures based solely on their color vectors. Similarly, [25] attempts to restyle the images so that they evoke the same feeling as the given music. While there are studies on how colors and styles influence our perception of a view, in reality, many other factors also play a significant role.

Furthermore, matching is done based on a limited set of emotional categories. Theoretically speaking, combinations of these categories can cover the whole spectrum of emotional responses to art. However, emotional responses are not the only way humans identify and differentiate a song or a piece of art. When thinking of a song, people may also think of the lyrics, the artists, the time era, the genre, etc. When looking at a picture, people may also appreciate the artistic talent behind it, the people, objects, events, or stories that are portrayed. Compressing all these factors into rigid boxes of emotions to use them as criteria for matching would be missing out on the richness of music and art. It is a loose connection to make.

In this part of the project, I seek to resolve these underlying issues of existing approaches. First, to expand the possibilities of output images, instead of recommending, this method would produce new artworks. Second, to release the strict association between emotions and colors or styles, generated images would also include contents that match the input music. Lastly, instead of encapsulating all musical features into emotional categories, the process would make use of all available and relevant words associated with the songs.

The first two objectives are implemented with the help of Dall-E 2, a text-to-image generative AI by OpenAI. I chose Dall-E 2 for two reasons: first, it is a popular and well-trusted tool among available models of the same purpose, built on 400 million text-image pairs with 3.5 billion parameters; and second, it has a public API, which would make it straightforward to build an end-user website for showcasing results. For the last objective, I used Last.fm API to get song tags. Last.fm is an online music-listening platform with active users who can freely assign tags for songs. Unlike a static database, new songs and tags are updated live. These tags are also not pre-defined but created by online users, making the tags more diverse and better reflecting the general audience's opinions on each song. Moreover, its database also incorporates tags from the Million Song Dataset (MSD), thus increasing the chance of having higher quality tags for songs that are in MSD.

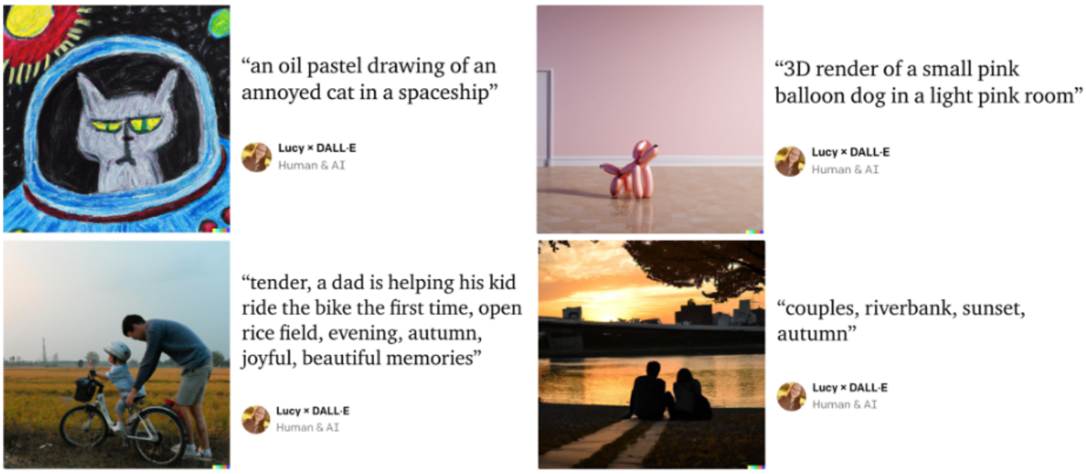


Figure 3.1: Images Generated by Dall-E 2 Based on Text Prompts. Generally, Dall-E 2 does not need a grammatically correct sentence. Phrases usually suffice.

From a high level, this method is similar to past attempts in that language serves as a bridge connecting music and art. However, while "language" only means emotional terms in others, in this case, language is all that is available.

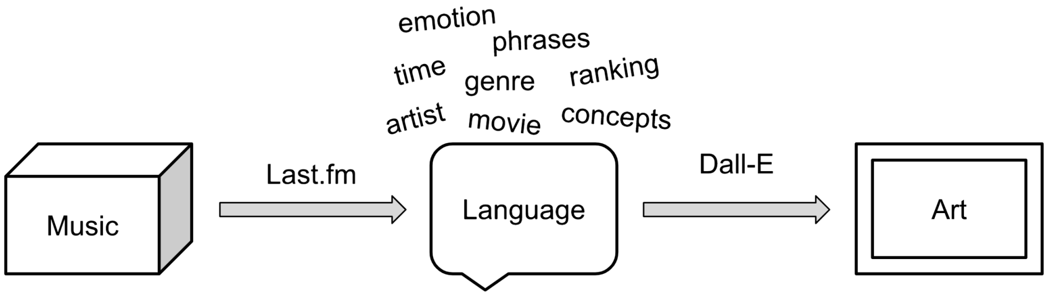


Figure 3.2: Method 2 From a High-Level Perspective.

3.2 Method

3.2.1 Procedure

Metaphorically speaking, this method is like putting existing Lego pieces into place. In this phase of the project, I built a website using Next.js that allows users to enter a song name and receive four images that match the song. The underlying process is as follows:

1. Validate that the song title and artist are correct and exist in the Last.fm database. Then, check if the song has any tags. If either of these two steps fails, show the user an error message.[1] Otherwise, move to the next step.

2. Last.fm API returns all the tags associated with the songs in the form of JSON arrays. Besides the tag name itself, each JSON object also has a "count" property, which is the percentage of users that use this tag over all users that have tagged this song. In other words, this number indicates the popularity of the tag for the song. Generally, the higher it is, the more chance there is that the tag is valid (not a junk tag). Therefore, I filter out tags whose "count" is less than 10. This threshold is arbitrary, as the best threshold wildly differs from song to song. However, this threshold generally ensures that the tags to be included in the next steps are at least not chosen by too few people.

3. The tags are filtered again using an existing language model to make sure there are no junk tags, and that the tags passed to Dall-E are relevant for image generation (more on this later).

4. The remaining tags are included in a prompt that the website sends to the Dall-E 2 API. The website then shows the image responses to users.

3.2.2 Prompt Design for Tag Filtering

While the wide range of tags is helpful in describing the songs from many angles, not all tags are helpful to draw images. Take the example of the tags for these songs. The numbers in parentheses are the tags' "count." All the tags have a "count" of more than 10.[2]

- I'm Not the Only One by Sam Smith: soul (100), pop (84), melancholy (34), 2014 (25), songs to kill yourself to (25), honest (21), sad (17), british (13), Ballad (13), love at first listen (13)

[1] This is usually the case for non-English songs.
[2] Tags are last updated on April 18, 2023.

- Symphony No. 5 by Beethoven: Classical (100), instrumental (60), favorites (20), haunting (20), intense (20), halloween (20), Movies (20), the best eclectic tag ever - period - dec-05 top-30s (20), epic maneuver (20), slgdmbestof (20), slgdm (20), Epic Maneuver music (20)
- My Heart Will Go On by Celine Dion: pop (100), Soundtrack (92), Celine Dion (87), female vocalists (78), titanic (59), Love (32), 90s (29), Ballad (29), My Heart Will Go On (26), romantic (23), love songs (18), beautiful (13), Canadian (12), easy listening (11), female vocalist (11)

These tags show that even after filtering out tags whose "count" is less than 10, there are still many tags that are not suitable as descriptions for Dall-E 2 to draw images from. Generally, adjectives related to emotions should be kept, because the images need to retain the same feelings and vibes that the songs convey. However, emotions alone are not enough to describe an image, especially when the objective of this method is to make the image results realistic, and interpretable. There needs to be some main character, object, place, or concept in the pictures. Therefore, I decided that the criteria for tags would be adjectives related to emotions, and nouns.

With these two criteria, I chose text-davinci-003, the most capable GPT 3.5 language model that is not optimized for chat [26], to filter the tags. To minimize randomness in results and prioritize accuracy, the temperature is set to 0. After experimenting with different prompts, the prompt "Find common adjectives related to emotions, and common nouns, from these: ", followed by the tags produces the most accurate results. I added the word "common" to remove some non-sense, lengthy or abbreviated tags that the language model may classify as nouns.

3.2.3 Prompt Design for Image Generation

I also experimented with two different types of prompts for Dall-E 2: tags only, and full-sentenced image descriptions. The former is constructed with "Draw a digital art painting based on these keywords: ", followed by the filtered tags. The latter is obtained by passing the filtered tags to text-davinci-003 with the prompt: "Describe a digital art painting based on these keywords: ". The results from text-davinci-003, which usually start with "This digital art painting", are then used directly as prompts for Dall-E 2. Generally, images created from tags are more conceptual, ambient, and symbolic, while images drawn from full sentences are more concrete, vivid, and specific. I did not include this step in the procedure, because there is no definitive answer whether it would be "better" to include it. At this point, it is up to personal preferences regarding the dichotomies mentioned above.

The `max_tokens` parameter is set to 256, which not only allows text-davinci-003 space to harness all tags in a detailed, expressive manner without having to cut off the captions mid-sentenced, but also limits the length of the prompt for Dall-E 2. Because Dall-E 2 processes the prompt as a single stream of data, the sooner a word appear in the description, the more prominent it is expressed in the resulting images. A prompt with more than 256 tokens will produce somewhat the same images as a prompt with 256 tokens. Therefore, choosing 256 as the token limit also ensures efficient captions for Dall-E 2. Although not confirmed for Dall-E 2, this number is also the maximum BPE-encoded tokens that the first Dall-E model uses to represent input captions [27].

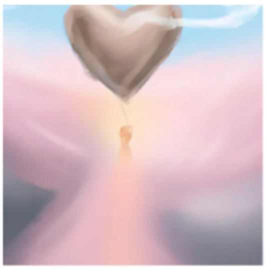
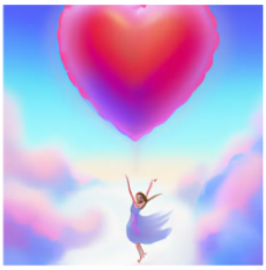
Temperature plays an important role in the generated descriptions, and ultimately the output images. A temperature of 0.01 guarantees that the description for a song is the same across different API calls, and is thus beneficial for comparison purposes. However, a temperature of 0.1 produces different versions of the description, some of which may yield more interesting and fitting images for the input song. Take the example of *My Heart Will Go On* by Celine Dion. After filtering for the "count" threshold, adjectives related to emotions, and nouns, the remaining tags are: pop, Soundtrack, Celine, Dion, female, vocalists, titanic, Love, 90s, Ballad, My, Heart, Will, Go, On, love, songs, Canadian, related, to, emotions, romantic, beautiful, easy, listening. With the same tags and prompt fed into text-davinci-003, when the temperature is set to 0.01, the generated description is always similar to

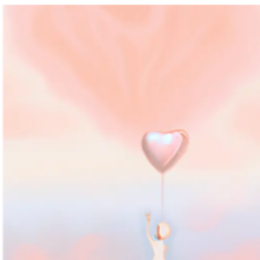
"This digital art painting is inspired by the iconic Canadian female vocalist Celine Dion and her 90s ballad "My Heart Will Go On" from the soundtrack of the movie Titanic. The painting is a beautiful and romantic representation of the song, with a pop art style. The painting is composed of a bright blue background, with a large heart in the center, representing the love and emotions related to the song. The heart is filled with a collage of images of Celine Dion, as well as images of the Titanic and other symbols of love. The words "My Heart Will Go On" are written in a bright pink font, with a bright yellow outline, making the words stand out. The painting is easy to listen to, with a soft and beautiful soundtrack of Celine Dion's love songs playing in the background."




Whereas when the temperature is 0.1, the captions can be

"This digital art painting is inspired by the iconic Canadian female vocalist Celine Dion and her 90s ballad "My Heart Will Go On" from the Titanic soundtrack. The painting is a romantic and beautiful representation of the song, with a soft pastel color palette. The painting features a female figure in the center, with a heart-shaped balloon in her hands, representing the love and emotions related to the song. The background is filled with a collage of images related to the song, such as a boat, a sunset, and a starry night sky. The painting is easy to listen to and evokes a feeling of nostalgia for the 90s love songs."



or

"This digital art painting is inspired by the iconic Canadian female vocalist Celine Dion and her song "My Heart Will Go On" from the 90s. The painting is a pop art style with a bright and vibrant color palette. The painting features a portrait of Celine Dion in the center, surrounded by a collage of images related to the song and the movie Titanic. The images include a heart, a boat, a rose, and a sunset. The background is a mix of blues and purples, with a few stars and a moon. The painting is a tribute to the romantic and beautiful love songs of the 90s, and to the emotions that they evoke. It is an easy listening piece that will bring back memories of the movie Titanic and the song "My Heart Will Go On"."

or

"This digital art painting is inspired by the iconic Canadian female vocalist Celine Dion and her iconic 90s ballad "My Heart Will Go On" from the soundtrack of the movie Titanic. The painting is a beautiful and romantic representation of the song, with a pop art style. The painting is composed of a bright pink background with a silhouette of Celine Dion in the center, with the words "My Heart Will Go On" written in a bold font. The painting also features a variety of other elements related to the song and Celine Dion, such as a heart, and a rose."

Again, results vary from song to song, and there are no objective metrics to judge which temperature is better. For the website, I chose 0.1 to allow for some variety, which can make the process of creating images from songs appear more human-like to end users.

The detailed process of this method is summed up in the figure below: [3]

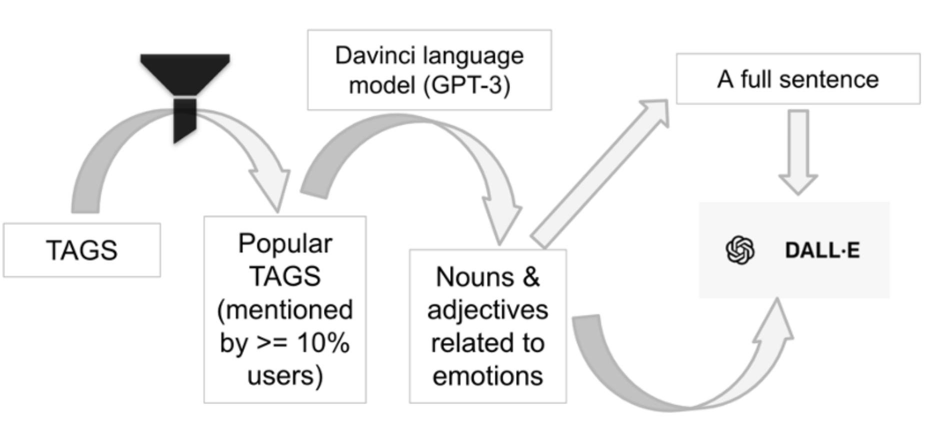


Figure 3.3: Detailed Procedure of Method 2.

[3] Code for method 2 is available at https://github.com/lucy-tran/Music-On-Canvas.

3.3 Results

The following images show some example outputs. For each piece of music, I generated two sets of images: one based on just tags, and one based on a full-sentenced description.

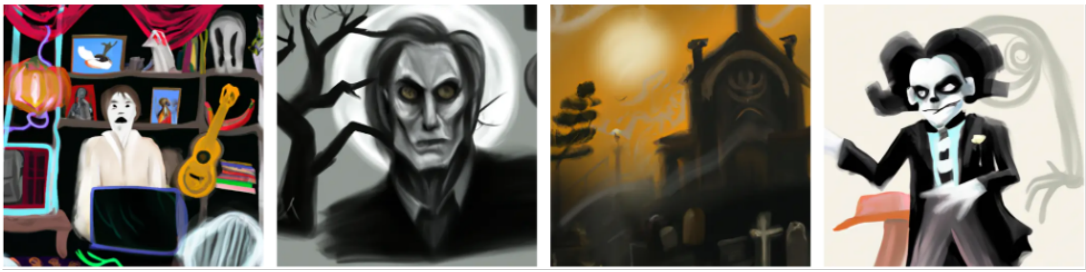
- *Lullaby* by Brahms
  - Draw a digital art painting based on these keywords: Classical, Sleepy, Mellow, Gentle, Soft, Johannes Brahms, Instrumental, Guitar, Harp, Brahms Lullaby, Ave Maria, Sleep Baby

  - This digital art painting is a gentle and soft landscape of a classical music-inspired dreamscape. The painting is set in a meadow of lush green grass and wildflowers, with a sleepy, mellow atmosphere. In the center of the painting is a small stage, with a classical guitar and a harp. On the stage, a musician is playing a gentle instrumental version of Johannes Brahms' "Brahms Lullaby" and "Ave Maria". The music is soft and soothing, and the painting is a peaceful reminder of the beauty of classical music. The painting is a reminder to take a moment to relax and enjoy the beauty of the music, and to let it lull you into a peaceful sleep.

- *Liebestraum* by Franz Liszt
  - Draw a digital art painting based on these keywords: Romantic, Classical, Piano, Liszt, Classic, Harp

  - This digital art painting is a romantic and classic scene of a grand piano in a beautiful, candlelit room. The piano is being played by a figure of Franz Liszt, the famous composer and pianist. He is wearing a black suit and a white shirt, and his hands are gracefully playing the keys of the piano. Behind him is a harp, which is also being played by a figure of a woman. She is wearing a long, white dress and her hands are delicately plucking the strings of the harp. The room is filled with a warm, romantic atmosphere, and the light from the candles is reflecting off the piano and the harp.

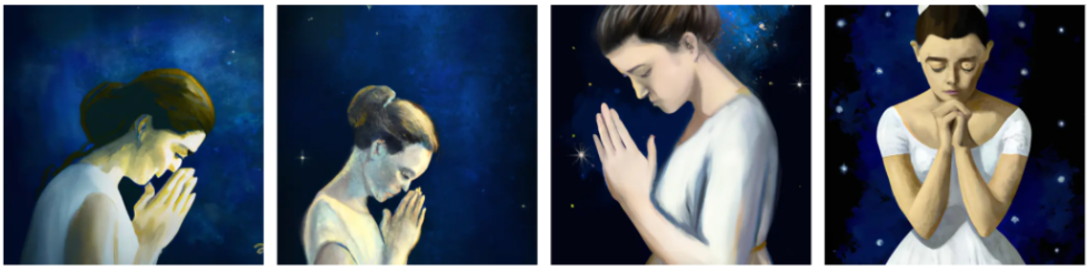
- *Symphony No.4* by Beethoven
  - Draw a digital art painting based on these keywords: haunting, intense, favorites, halloween, movies, music, period, top-30s



  - This digital art painting is a hauntingly intense collage of some of the most beloved Halloween movies, music, and period pieces from the top 30s. The painting is composed of a variety of iconic images from the era, including a black and white photo of a spooky graveyard, a still from the classic horror movie Frankenstein, a vintage record player playing a classic Halloween song, and a sepia-toned image of a woman in a period costume. The colors are muted and dark, creating a moody atmosphere that is both eerie and nostalgic. The painting is a perfect way to celebrate the best of the top 30s Halloween season.

- *I'm Not the Only One* by Sam Smith
  - Draw a digital art painting based on these keywords: Sad, Honest, Melancholy, Soul, Pop, Songs, Ballad, Love, Listen

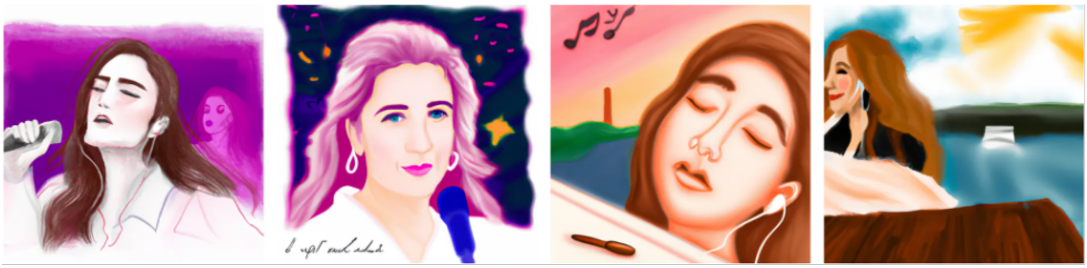
  - This digital art painting is a melancholic portrait of a woman with her head bowed and her eyes closed. She is wearing a white dress and her hair is pulled back in a bun. Her hands are clasped together in a prayer-like gesture. The background is a deep blue, with a few stars twinkling in the night sky. The woman is surrounded by a halo of light, as if she is being illuminated by a divine presence. The painting is meant to evoke a sense of sadness and honesty, as if the woman is in the midst of a soul-searching journey. The colors are muted and the atmosphere is somber, yet there is a hint of hope in the air. The painting is set to the tune of a pop ballad, with lyrics about love and longing. The painting is meant to be a reminder to take the time to listen to our innermost thoughts and feelings, and to be honest with ourselves.



- *Symphony No. 40* by Mozart
  - Draw a digital art painting based on these keywords: Joyful, passionate, powerful, emotional, uplifting, Melody, harmony, rhythm, composition, conductor, orchestra.

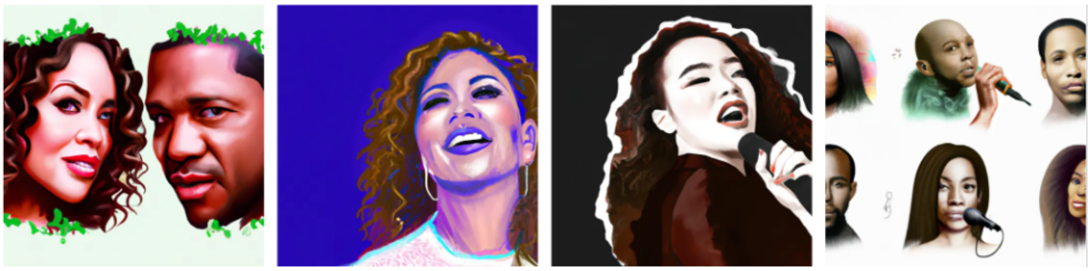
  - This digital art painting is a vibrant and powerful scene of a conductor leading an orchestra in a passionate and emotional performance. The conductor stands in the center of the painting, baton in hand, leading the orchestra in a joyful and uplifting melody. The orchestra is in perfect harmony and rhythm, creating a beautiful composition of sound. The conductor's passionate movements and the orchestra's powerful performance create an uplifting and emotional atmosphere. The painting is a celebration of the joy and power of music.

- *My Heart Will Go On* by Celine Dion
  - Draw a digital art painting based on these keywords: romantic, beautiful, pop, soundtrack, Celine Dion, female vocalists, titanic, love, 90s, ballad, song, Canadian, listening.

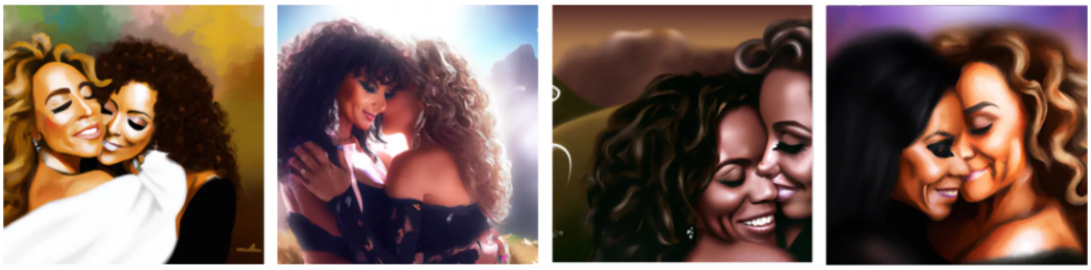


  - This digital art painting is a romantic and beautiful tribute to the iconic Canadian female vocalist Celine Dion and her song "My Heart Will Go On" from the movie Titanic. The painting is in a pop art style, with a bright blue background and a silhouette of Celine Dion in the center. The silhouette is surrounded by a collage of images from the 90s, including a cassette tape, a CD, a record player, and a radio. The painting also features a few lyrics from the song, such as "Near, far, wherever you are" and "Love can touch us one time and last for a lifetime". The painting is a reminder of the power of love and music, and how it can bring us together, no matter where we are.

- *When You Believe* by Whitney Houston and Mariah Carey
  - Draw a digital art painting based on these keywords: romantic, soft, loving, passionate, emotional, vocalists, soul, RB, pop, soundtrack, 90s, ballad, lovesongs, easy listening, duet, Whitney, Prince of Egypt, jazz, Mariah Carey, divas, best ballads.



- This digital art painting is a romantic and soft portrait of two of the most iconic vocalists of the 90s: Whitney Houston and Mariah Carey. The two divas are depicted in a passionate embrace, their faces close together and their eyes closed in a moment of deep emotion. The background is a soft, dreamy landscape of the Prince of Egypt soundtrack, with a hint of jazz and RB. The painting is a tribute to the best ballads of the 90s, from easy listening lovesongs to powerful duets. It captures the soul of the era, and the beauty of two of the greatest vocalists of all time.



3.4 Discussions

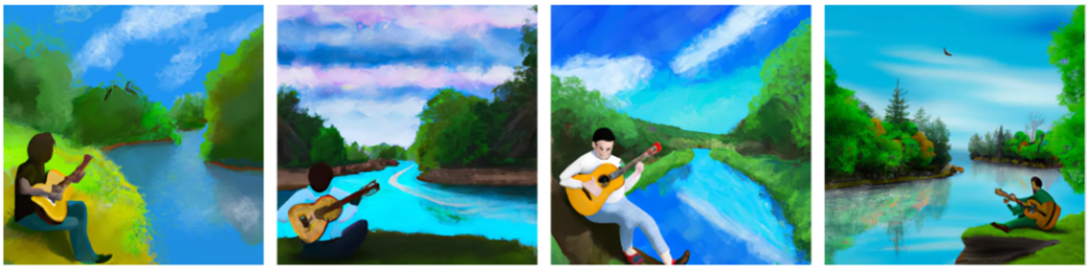
Thanks to the power of existing AI models (text-davinci-003 and Dall-E 2) and the availability of Last.fm API, this method is able to satisfy the three objectives I aim for: (1) generating new artworks instead of recommending or editing available ones, (2) the outcome images not only feature the emotions, but also the contents (people, objects, symbols) that fit the input music, and (3) all available and relevant ideas, including but not limited to feelings, are utilized to understand songs. Overall, the images are realistic, and interpretable, and thus are more relatable to the general audience than the abstract visualizations that method 1 and its related works aim for.

However, this method has three limitations. First, tags from Last.fm can be inaccurate, especially regarding instruments. For example, although one can play Brahms' *Lullaby* or Liszt's *Liebestraum* on harp, the original instruments for these pieces do not include the harp that the tags suggest. Second, in generating full-sentenced image descriptions based on the tags, text-davinci-003 may wrongly interpret the tags. For example, the tag "soul" in *I'm Not the Only One* by Sam Smith means the soul music genre. However, text-davinci-003, when reading the tag along with other tags like "honest" and "love", understands it as in "soul-search." Hence, the model describes a woman in a "prayer-like gesture."

The third limitation happens in two cases: (1) when the songs requested do not exist in the Last.fm database, and (2) the tags only include generic instruments, genres, or emotions. In the first case, there is no information to produce pictures about the musical piece. In the second case, there is not enough information to create pictures that are specific and authentic to the song. Take the example of *All I Ask of You* by Piano Instrumental:

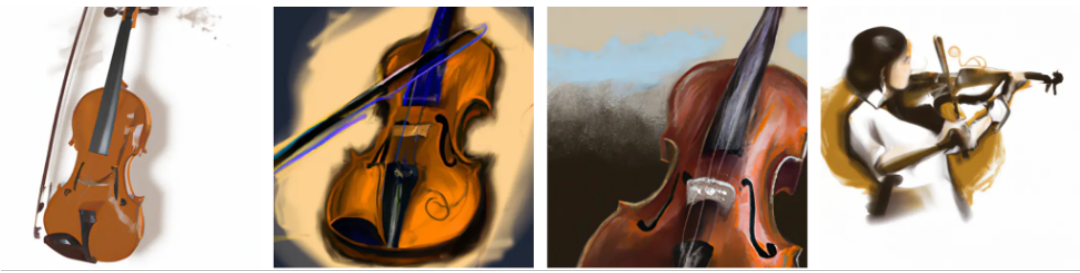
"Draw a digital art painting based on these keywords: Beto1, instrumental, relaxing"

"This digital art painting is inspired by the musician Beto1 and his instrumental music. It features a tranquil landscape with a winding river and lush green trees. The sky is a deep blue with a few wispy clouds. In the foreground, a figure is playing a guitar, creating a peaceful and relaxing atmosphere. The colors are muted and calming, with a hint of warmth from the sun. The painting is a reminder to take a moment to relax and enjoy the beauty of nature."

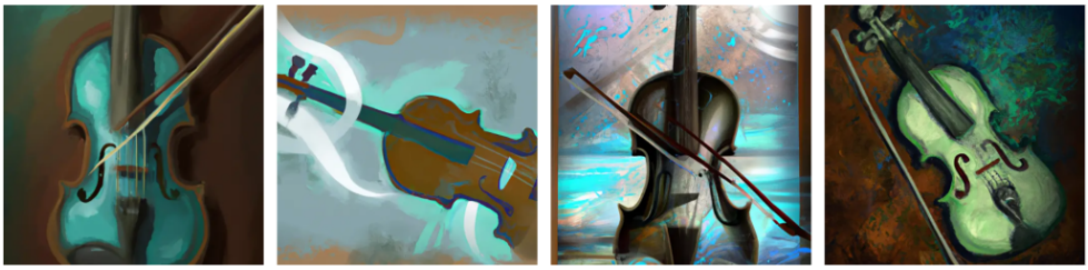


This song is an instrumental piano piece. However, because the tags only reveal that it is instrumental, Dall-E 2 and text-davinci-003 infer different instruments that are incorrect to the musical piece. Or, take the example of *Winter* by Antonio Vivaldi:

"Draw a digital art painting based on these keywords: music, Klassik, classic, instrumental, baroque, violin, classical"

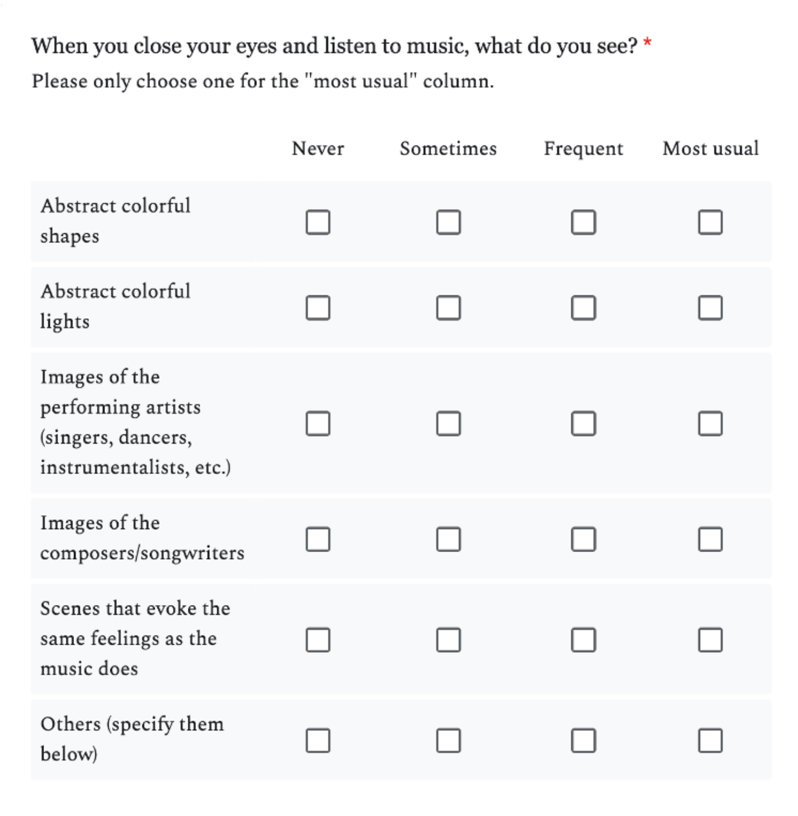


"This digital art painting is inspired by the classic music of the Baroque period. It features a violin playing a beautiful instrumental piece of Klassik music. The painting is in a classic style, with a muted color palette of blues, greens, and browns. The violin is the focal point of the painting, with the strings of the instrument curving gracefully around the body of the instrument. The background is a simple landscape of rolling hills and trees, with a few birds flying in the sky. The painting conveys a sense of peace and tranquility, and the music of the violin is a reminder of the beauty of classical music."



These images are generic and can apply to any violin piece of music. They do not clearly evoke the fierceness of the piece instrumentally, or the winter imagery that Vivaldi depicts through music.

Besides, after assessing the image results and how they enhance the music-listening experience, I realized that visions of the composers, singers, or instruments are usually not what I have in mind when listening to their music. In most cases, I have vague visions of a typical scene, whose general vibe is the same as the vibe that the music evokes. To evaluate this realization, I did a survey on what others see when they listen to music. The survey received 22 responses, with 4 levels of frequency for each possible vision. Scores are then summed up for each vision, with "never" as 0 points, "sometimes" as 1, "frequent" as 2, and "most usual" as 3.



When you close your eyes and listen to music, what do you see? *

Please only choose one for the "most usual" column.

	Never	Sometimes	Frequent	Most usual
Abstract colorful shapes	☐	☐	☐	☐
Abstract colorful lights	☐	☐	☐	☐
Images of the performing artists (singers, dancers, instrumentalists, etc.)	☐	☐	☐	☐
Images of the composers/songwriters	☐	☐	☐	☐
Scenes that evoke the same feelings as the music does	☐	☐	☐	☐
Others (specify them below)	☐	☐	☐	☐

Figure 3.4: The Main Question of Method 2's Ending Survey.

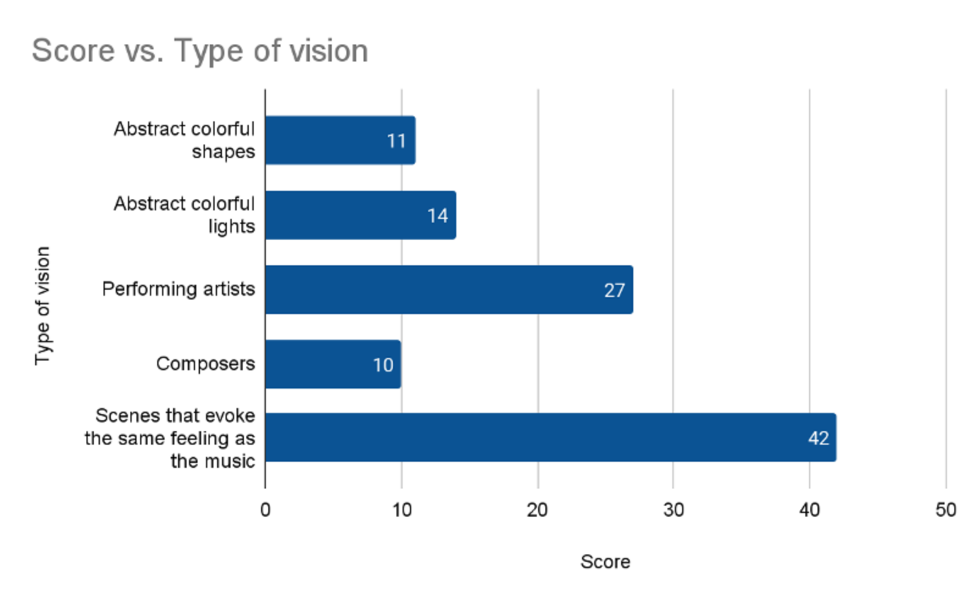


Figure 3.5: Five Types of Vision a Person May Experience During Music Listening and Their Popularity Scores.

Although the sample size is small, the survey result shows that "scenes that evoke the same feeling as the music" are common, if not the most popular, visions that come to the human mind during music listening. In other words, this type of vision is a realistic one to expect from users and thus has a high chance of being relatable to them. Visions of performing artists and/or composers are also common and are what this method has been able to create, but it is missing if the website cannot produce scenic pictures that convey the same feeling as the music.

In light of the three limitations and the discovery above, I decided to experiment with another method. This new method is not a replacement for method 2, but a complement that tackles its limitations discussed. Because the three limitations are inherent to the public accessibility of Last.fm, and because Last.fm tags are not sufficient to generate scenic (as opposed to performing artists or composers) descriptions that are authentic to input songs, this new approach eliminates the use of Last.fm in understanding the music. Instead, I will utilize machine learning to infer the characteristics of the input songs. To narrow down the scope, this method will focus on instrumental music, partly because this is the genre of music that Last.fm lacks information on the most. The new objective is to generate descriptive captions for instrumental songs such that the output images are scenic.

4 Method 3: Generating Scenic Art for Instrumental Music

4.1 Background

For a machine to recognize the characteristics of a piece of music without reading directly from its tags, it needs to be trained on existing data. This task is core to the field of Music Information Retrieval (MIR). Early efforts in this field use text as an isolated modality, applying Natural Language Processing techniques to build knowledge of the musical world based on artist biographies [28], album reviews [29], or a collection of texts related to music [30]. Thanks to recent advancements in Computer Vision, studies in MIR have shifted focus to multimodal input data, typical audio, and text. For example, *musicnn* [31] and *Harmonic CNN* [32] are two Convolutional Neural Networks (CNN) used for automatic music tagging, trained on audio-text pairs where the audio modality is represented as spectrograms of the raw audio inputs. Other applications of Deep Learning techniques in MIR are music emotion recognition [33], mood detection [34], and music classification and recommendation [35].

However, the text modality of all these works only considers genres, emotions, instruments, moods, themes, or sonic descriptions of music. There exists no scenic caption dataset for songs. It is possible that scenic caption generation or recommendation is not even thought of before. But, based on previous research on music captioning and retrieval, which have successfully assigned tags for songs, I hypothesize that if there is a large enough dataset for scenic captions, then these models are also able to write, or at least recommend, scenic captions.

Given the time and resource constraints at this point of the project, it is infeasible to annotate thousands of scenic captions to prove this hypothesis. Therefore, I decided to find a workaround that serves as a proof of concept. In this approach, I tune an existing model on a music-mood dataset to recognize the mood of the songs. Then, I build a small dataset called Imagination that includes scenic captions for music.[1] These captions should have the same mood vocabulary as the training set. During testing, the Imagination dataset is used as the test set to see if the model can retrieve *suitable* captions in the Imagination dataset for the input audio based on the emotional vocabularies that are common between the training and

[1] Dataset and code for preprocessing is available at https://github.com/lucy-tran/Imagination-Dataset.

test set. Suitability is assessed by listening to the music and viewing the images generated from the recommended captions simultaneously.

Among several systems for multimodal music learning, I selected MusCALL, a framework for Music Contrastive Audio-Language Learning [36]. One task that this model is able to do well is cross-modal retrieval of music. For audio-to-text retrieval, which is the target task for the current method, the model performs considerably better than two baseline models of a MIR-related challenge (with recalls at 1, 5, and 10 being 25.8, 53.0, 63.0).

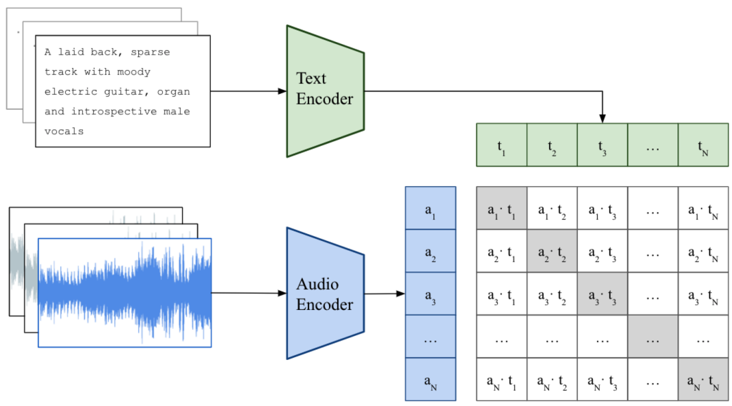


Figure 4.1: Overview of MusCALL. An audio encoder and a text encoder are trained via a contrastive loss to maximise the similarity between representations of N aligned (audio, text) pairs within a mini-batch. At test time, the similarity between embeddings in the learnt multimodal space is used to rank database items and perform cross-modal retrieval [36].

For model configurations, such as parameters related to the audio encoder and the text encoder, I use the original values in MusCALL's codebase. Tuning is done on learning rate and batch size to find the optimal combination of these parameters. The metric for determining the best model is recall-at-10 (R@10).[2]

4.2 Method

4.2.1 Preprocessing

For training data, I chose the MTG-Jamendo dataset. The dataset contains over 55,000 full audio tracks with 195 tags from genre, instrument, and mood/theme categories. While genre and instrument are helpful, because the current objective is only to learn the mood/theme of songs, only mood/theme tags are used. In addition, because MusCALL learns music representation using the ResNet-50 CNN on the Mel spectrogram of the audio files, I suspect that including songs of different instruments would make it harder for the model to learn the mood/theme apart from the instrument. This is because even for the same note playing, the Mel spectrogram of different instruments may look different due to harmonic overtones. Therefore, considering the time limit at this point of the project, I decided to narrow down the scope by only including songs with the tag "piano."[3]

Due to memory and time constraints, only a part of the full MTG-Jamendo dataset is

[2]Experiments with MusCALL is available at https://github.com/lucy-tran/MusCALL.

[3]A "piano" tag does not mean the piano is the only instrument in the song. Although for most songs with the "piano" tag, this is the case, I later realize that there are many songs which have a mix of piano, vocal and other instruments. This discovery came much later in the project, so I did not have the time to amend this mistake.

downloaded. After preprocessing, this part contains a total of 1477 audio files, each with 1 to 7 mood/theme tags.

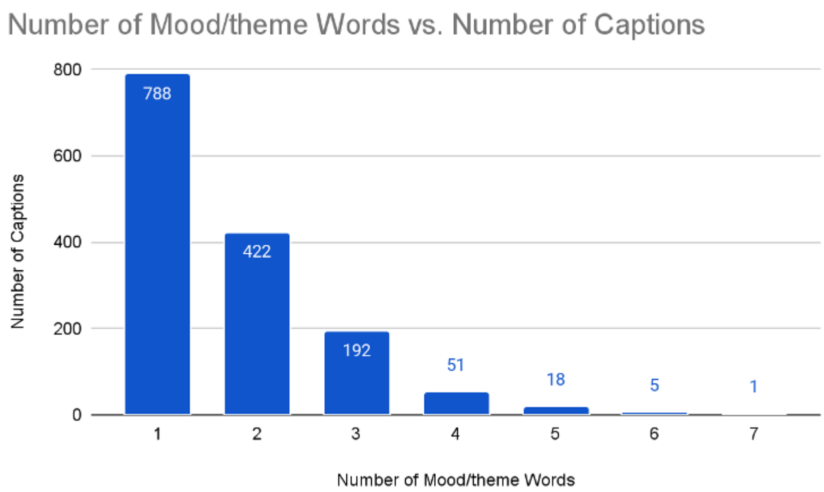


Figure 4.2: Number of Mood/Theme Words in a Caption Versus Number of Captions in Method 3's Training Set.

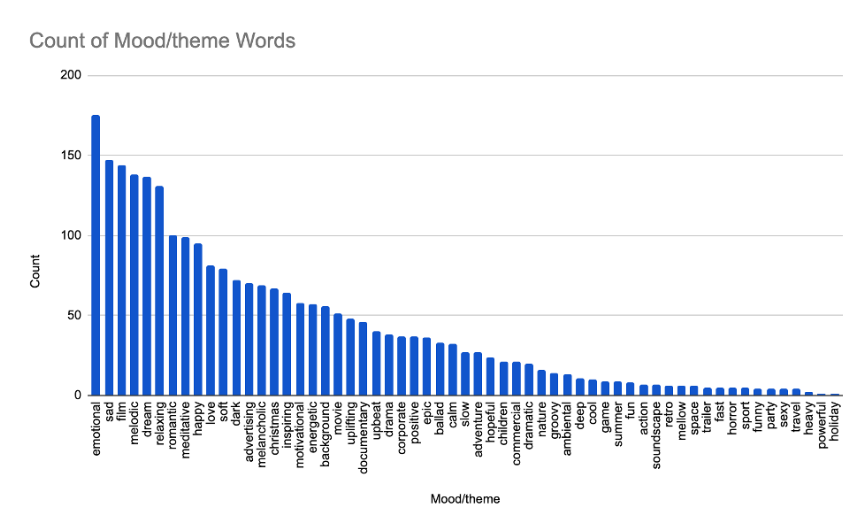


Figure 4.3: Counts of Mood/Theme Words in Method 3's Training Set.

4.2.2 Designing the Imagination dataset

The design of the Imagination dataset is based on the MTG-Jamendo dataset and the desired image results. First, because I only include piano music from the MTG-Jamendo dataset, I also only include piano instrumental music in Imagination. Wav files are downloaded copyright-free from pixabay.com and chosic.com. Then, the mood/theme words from MTG-Jamendo are prepended to the captions for each song, in proportion to the words' distribution in the training set. The remaining parts of the captions consist of an object, person or people, followed by an action, a location, a time of day, and a season. Some tracks also have addition feelings and scenic details at the end of their captions.

Mood/theme	Objects/People	Action	Location	Time of Day	Season	Other feelings	Additional #1	Additional #2

Figure 4.4: Contents of Captions in the Imagination Dataset. In practice, a caption is a string that consists of these columns separated by commas.

4.2.3 Tuning: First Experiment

In the first experiment, the MTG-Jamendo dataset is used for both training and validation. A split of 85:15 is made to divide the 1477 items of MTG-Jamendo into training and validation sets. Because MusCALL authors use a learning rate of 5e-5 for their dataset of private production music, I started tuning with learning rates 5e-4, 5e-5, and 5e-6. For each learning rate, batch sizes 8, 16, and 32 are used.[4] For each pair of learning rate - batch size, I ran 10 trials to avoid false comparison due to random initialization of parameters.

Training using the MTG dataset as the validation set does not yield the best results. Models with better metrics do not consistently have better R@10 on the test set. I also did manual testing to see if the models with the best R@1 give suitable first-ranked captions for the testing songs. The table below shows an example of these recommendations from a model with an R@1 of 15 on a test set that includes 20 instances.

id	Original caption	1st-ranked caption
0	**happy**, **hopeful**, man, helping his child to ride a bike, open field, evening, autumn, joyful, beautiful memories	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around
1	**happy**, **dream**, girl, riding horse, across a park, noon, spring, joyful, dreamy, sun shining	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around
2	**happy**, couple, looking down river, from a bridge, evening, autumn, emotional, autumn leaves flying, beautiful scene	**retro**, **upbeat**, gladiator, making an entry, in circuit, night, spring, joyful, crowd cheering
3	**ambiental**, **uplifting**, couple, chasing each other, across a park, morning, autumn, joyful, dreamy, autumn leaves flying	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around
4	**relaxing**, girl, playing piano, on blue abstract background, morning, spring, joyful, music notes flying around	**inspiring**, **dream**, a girl, holding a championship cup, in a photo frame, morning, spring, joyful, nostalgia, sun shining

[4]When run with batch size 64 and larger, Macalester lab computer throws an error related to memory.

5	**documentary**, shepherd, leading a flock of sheep, on green pasture, morning, spring, joyful, sun shining, beautiful scene	**inspiring**, **dream**, a girl, holding a championship cup, in a photo frame, morning, spring, joyful, nostalgia, sun shining
6	**calm**, **relaxing**, canoe, floating, on river, night, spring, peaceful, moon shining, beautiful scene	**inspiring**, **dream**, a girl, holding a championship cup, in a photo frame, morning, spring, joyful, nostalgia, sun shining
7	**relaxing**, **dream**, butterflies made by lights, flying, open field, night, spring, nostalgia, dreamy, beautiful scene	**relaxing**, **dream**, butterflies made by lights, flying, open field, night, spring, nostalgia, dreamy, beautiful scene
8	**retro**, **upbeat**, gladiator, making an entry, in circuit, night, spring, joyful, crowd cheering	**retro**, **upbeat**, gladiator, making an entry, in circuit, night, spring, joyful, crowd cheering
9	**relaxing**, **dream**, girl, looking at lanterns floating, on an open sky, night, spring, emotional, dreamy, beautiful scene	**relaxing**, **dream**, girl, looking at lanterns floating, on an open sky, night, spring, emotional, dreamy, beautiful scene
10	**emotional**, **hopeful**, girl, looking up to the sky, on a green pasture, morning, spring, joyful, hopeful, sun shining	**relaxing**, **dream**, girl, looking at lanterns floating, on an open sky, night, spring, emotional, dreamy, beautiful scene
11	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around	**retro**, **upbeat**, gladiator, making an entry, in circuit, night, spring, joyful, crowd cheering
12	**retro**, **uplifting**, guy, playing with a dog, on an open road, noon, autumn, nostalgia, autumn leaves are flying, sun shining	**retro**, **upbeat**, gladiator, making an entry, in circuit, night, spring, joyful, crowd cheering
13	**melancholic**, **sad**, a glass heart, broken, on blue abstract background, night, winter, regretful, sorrows, emotional	**happy**, **dream**, girl, riding horse, across a park, noon, spring, joyful, dreamy, sun shining
14	**melancholic**, **sad**, guy, looking down a river, from a bridge, evening, autumn, regretful, sorrows, autumn leaves are flying	**calm**, **relaxing**, canoe, floating, on river, night, spring, peaceful, moon shining, beautiful scene
15	**happy**, **dream**, group of people, having dinner together, in a garden, evening, spring, joyful, nostalgia, memories	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around
16	**inspiring**, **dream**, a girl, holding a championship cup, in a photo frame, morning, spring, joyful, nostalgia, sun shining	**happy**, couple, looking down river, from a bridge, evening, autumn, emotional, autumn leaves flying, beautiful scene
17	**inspiring**, **uplifting**, butterflies made by lights, are flying, across a starry sky, night, spring, dreamy, nostalgia, moon shining	**retro**, **upbeat**, gladiator, making an entry, in circuit, night, spring, joyful, crowd cheering
18	**fun**, an old couple, playing a see saw, in a garden, evening, spring, joyful, nostalgia	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around

19	**happy**, **energetic**, a ballerina, dancing, on blue abstract background, night, spring, joyful, nostalgia, music notes flying around	**retro**, **upbeat**, gladiator, making an entry, in circuit, night, spring, joyful, crowd cheering

Table 4.1: First-ranked Captions by a Model In the First Experiment. The tags in bold are mood/theme tags that are common between the training MTG-Jamendo dataset and the testing Imagination dataset. The matchings in blue color are exact matchings. Those in red are clear mismatches because the mood/theme of the original captions is quite opposite to the mood/theme of the first-ranked captions. The matchings in yellow are reasonable based on their mood/theme tags, but the images generated from the recommended caption do not fit the input music. Other matchings are acceptable ones, with possibly fitting outcome images for the audio.

Tested on only 20 instances, this model has already produced 2 clear mismatches in a simple comparison between the mood/theme tags of the original captions and those of the first-ranked captions, and up to 6 recommendations whose outcome images are not suitable for the music.

One possible reason for these mismatches is the differences between the validation set and the test set. During training, the best models are chosen based on the metric on the validation set. But, because the validation set and the test set are very different on both the audio and text modalities, the metric does not extend well to the Imagination dataset.

4.2.4 Tuning: Second Experiment

Learning from the first experiment results, in the second experiment, I use the Imagination dataset as the validation set. However, with only 100 songs, there needs to be a split between the validation set and the test set, such that the best R@10 on the validation set leads to the best R@10 from the whole population (R@10 on 100 captions). To find the optimal split, making use of the generated tests on different test set sizes 20, 40, 60, and 80, for each model, I calculated the proportion of the R@10 at each test set size to the R@10 at 100 tests. Then, the standard deviations of these proportions across all models are computed. The result shows that R@10 from 80 tests has the least varied proportion to R@10 from 100 tests. On average, the R@10 from 80 tests is 1.986 +/- 0.434 the R@10 from 100 tests. Therefore, I chose 80 items from the Imagination dataset as the validation set.

This reduces the largest test set size to 20. Although this is very small, these 20 tests are enough to show end users how well the model recommends, for example, the best 5 captions for a song when given only 20 captions to select from.

Retraining was done on a training set of size 1477, with 3 trials for each learning rate - batch size pair. The outcome is still the same: the models with higher R@10 on the new validation set do not always yield the highest R@10 on the 20 tests. This may be due to the difference between the 80 vs. 20 partitions of Imagination, or the difference between the training and testing sets. However, there is a model with a significantly higher metric (R@10 on the 80-partition of Imagination) than others. This model also has the highest R@1 on the 20-partition. Thus, I chose this model, which has a learning rate of 5e-4 and a batch size of 8, to try listening to the songs and comparing them with the captions the model ranks first.

4.3 Results

The table below shows the original captions side-by-side with the captions ranked first by the selected model.

id	Original caption	1st-ranked caption
0	**happy**, **hopeful**, man, helping his child to ride a bike, open field, evening, autumn, joyful, beautiful memories	**ambiental**, **uplifting**, couple, chasing each other, across a park, morning, autumn, joyful, dreamy, autumn leaves flying
1	**happy**, **dream**, girl, riding horse, across a park, noon, spring, joyful, dreamy, sun shining	**ambiental**, **uplifting**, couple, chasing each other, across a park, morning, autumn, joyful, dreamy, autumn leaves flying
2	**happy**, couple, looking down river, from a bridge, evening, autumn, emotional, autumn leaves flying, beautiful scene	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around
3	**ambiental**, **uplifting**, couple, chasing each other, across a park, morning, autumn, joyful, dreamy, autumn leaves flying	**happy**, love, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around
4	**relaxing**, girl, playing piano, on blue abstract background, morning, spring, joyful, music notes flying around	**documentary**, shepherd, leading a flock of sheep, on green pasture, morning, spring, joyful, sun shining, beautiful scene
5	**documentary**, shepherd, leading a flock of sheep, on green pasture, morning, spring, joyful, sun shining, beautiful scene	**documentary**, shepherd, leading a flock of sheep, on green pasture, morning, spring, joyful, sun shining, beautiful scene
6	**calm**, **relaxing**, canoe, floating, on river, night, spring, peaceful, moon shining, beautiful scene	**melancholic**, **sad**, a glass heart, broken, on blue abstract background, night, winter, regretful, sorrows, emotional
7	**relaxing**, **dream**, butterflies made by lights, flying, open field, night, spring, nostalgia, dreamy, beautiful scene	**melancholic**, **sad**, a glass heart, broken, on blue abstract background, night, winter, regretful, sorrows, emotional
8	**retro**, **upbeat**, gladiator, making an entry, in circuit, night, spring, joyful, crowd cheering	**ambiental**, **uplifting**, couple, chasing each other, across a park, morning, autumn, joyful, dreamy, autumn leaves flying
9	**relaxing**, **dream**, girl, looking at lanterns floating, on an open sky, night, spring, emotional, dreamy, beautiful scene	**documentary**, shepherd, leading a flock of sheep, on green pasture, morning, spring, joyful, sun shining, beautiful scene
10	**emotional**, **hopeful**, girl, looking up to the sky, on a green pasture, morning, spring, joyful, hopeful, sun shining	**documentary**, shepherd, leading a flock of sheep, on green pasture, morning, spring, joyful, sun shining, beautiful scene
11	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around
12	**retro**, **uplifting**, guy, playing with a dog, on an open road, noon, autumn, nostalgia, autumn leaves are flying, sun shining	**ambiental**, **uplifting**, couple, chasing each other, across a park, morning, autumn, joyful, dreamy, autumn leaves flying

13	**melancholic**, **sad**, a glass heart, broken, on blue abstract background, night, winter, regretful, sorrows, emotional	**melancholic**, **sad**, a glass heart, broken, on blue abstract background, night, winter, regretful, sorrows, emotional
14	**melancholic**, **sad**, guy, looking down a river, from a bridge, evening, autumn, regretful, sorrows, autumn leaves are flying	**melancholic**, **sad**, a glass heart, broken, on blue abstract background, night, winter, regretful, sorrows, emotional
15	**happy**, **dream**, group of people, having dinner together, in a garden, evening, spring, joyful, nostalgia, memories	**ambiental**, **uplifting**, couple, chasing each other, across a park, morning, autumn, joyful, dreamy, autumn leaves flying
16	**inspiring**, **dream**, a girl, holding a championship cup, in a photo frame, morning, spring, joyful, nostalgia, sun shining	**happy**, **dream**, girl, riding horse, across a park, noon, spring, joyful, dreamy, sun shining
17	**inspiring**, **uplifting**, butterflies made by lights, are flying, across a starry sky, night, spring, dreamy, nostalgia, moon shining	**emotional**, **hopeful**, girl, looking up to the sky, on a green pasture, morning, spring, joyful, hopeful, sun shining
18	**fun**, an old couple, playing a see saw, in a garden, evening, spring, joyful, nostalgia	**happy**, **love**, an old couple, dancing, on blue abstract background, morning, winter, joyful, heart shape stars in the background, music notes flying around
19	**happy**, **energetic**, a ballerina, dancing, on blue abstract background, night, spring, joyful, nostalgia, music notes flying around	**ambiental**, **uplifting**, couple, chasing each other, across a park, morning, autumn, joyful, dreamy, autumn leaves flying

Table 4.2: First-ranked Captions Recommended by the Selected Model (learning rate = 5e-04, batch size = 8, R@1 = 15) for 20 Piano Instrumental Tracks from the Imagination Dataset.

Overall, this model returns better first-ranked recommendations. Among the 20 songs it is tested on, there is no clear mismatch based on the mood/theme tags. However, there are 5 caption recommendations whose end images do not closely fit the input songs. Since these are all instrumental songs, and there are no numeric metrics to objectively evaluate the matchings, these numbers may change depending on the beholder.[5]

4.4 Discussions

Although there are some unfitting recommendations from the best model, the fact that updating the validation set to be more similar to the test set improves the overall performance proves that the hypothesis in the beginning is true. If the Imagination dataset is large enough to be the training set, a model like MusCALL will be able to recommend more suitable scenic captions for instrumental music. These captions will then be prompts for Dall-E to create pictures that can enhance the music-listening experience.

Compare method 3 with method 2. In a previous discussion section (see pages 18-19), I came to the conclusion that due to its relatability, art that is based on our imaginations better evokes the feeling of music than art that influences our imagination. While images generated from method 2 can be inaccurate or too generic, for images whose tags are true and authentic

[5]For a slideshow of music along with images generated from the first ranked captions, visit https://github.com/lucy-tran/Scenic-Immersion.

to the input songs, it is safe to say that the images reflect common listeners' imagination, because they are produced based on listeners' tags. As a matter of fact, the total score for "performing artists" and "composers" types of vision from the survey (see Figure 16) is 37, only 5 points lower than the score for "scenes that evoke the same feeling as the music." Moreover, performing artists and composers are usually traceable and verifiable, making it easier to ensure the relatability of the output images. On the other hand, while the objective of method 3 is based on the most popular audience's response, and while pictures from method 3 can evoke the same feeling as the music, they may not be the same scenes that music listeners have in mind. The same music can be connected to different scenes, and different scenes can kindle the same emotions, depending on the individuals' experience and preferences. Thus, the pictures produced by method 3 can easily suffer from the same problem of unrelatability discussed in method 1. It would be best to understand these scenes as suggestions for, rather than reflections of, the audience's scenic imagination.

Compare both methods 2 and 3 with method 1. Both methods 2 and 3 produce much more realistic, understandable, and ultimately relatable images than method 1 does. Thanks to their relatability, these images should be able to reach a higher number of people than the abstract visualizations in method 1. However, as each person has their own experience and identity that can alter their imaginations differently, either in the details or in the entirety, compared to the majority, the users should have full autonomy over their imagination and perception of the music. In practice, this means that for songs that (1) exist in Last.fm and (2) have enough specific tags, the end product will show images produced by method 2; for songs that do not satisfy both criteria, the website will generate scenic captions following method 3, and show the pictures that Dall-E 2 produces based on the captions. In either case, the users should be able to tailor the description so that it matches their imagination and feelings.[6] In this way, I make sure to serve all audiences on the spectrum of relatability with the image results.

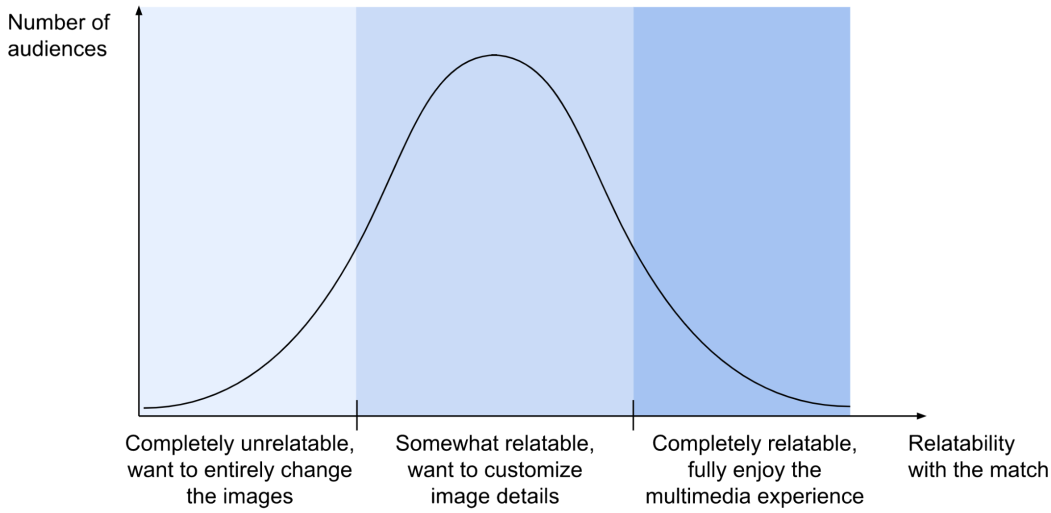


Figure 4.5: Hypothetical Number of Audiences on the Range of Relatability with Images Generated by Methods 2 and 3. By adding customization features, the website will serve all audiences regardless of their reaction to the match.

[6]This feature has not been implemented as of the end of this project, but will be in further research.

Attribute	Comparison	Basis
Interpretability	1 << 2 = 3	Common sense
Realisticity	1 << 2 < 3	Survey results (unconfirmed due to small sample size)
Chance of results' relatability	1 << 3 < 2	Realisticity and verifiability

Table 4.3: A Summary of Comparison of Methods 1, 2, and 3 Regarding Their Realisticity, Interpretability, and Chance of Results' Relatability to the Audience. "«" denotes "much less."

5 Future Directions

With all the achievements that method 2 has been able to accomplish, it would be a waste to entirely discard this simple yet effective approach. Although it cannot generate scenic captions because of the lack of tags, there can still be small improvements in what it can do with Last.fm tags. Specifically, both the inaccuracy of Last.fm tags, as text-davinci-003's wrong interpretations of the tags can be tackled with generated knowledge [37] before giving the final answers. For example, to avoid wrong instrument tags of Brahms' *Lullaby*, the model can be prompted to obtain basis knowledge of the song, before asked to filter out inaccurate tags:

```
Q: How many instruments are usually played in Brahms' Lullaby?

A: Brahms' Lullaby is typically performed as a solo piano piece.

Q: Based on that knowledge, omit the tags that are not true to
Brahms' Lullaby. Tags: Classical, Sleepy, Mellow, Gentle, Soft,
Johannes Brahms, Instrumental, Guitar, Harp, Brahms Lullaby, Ave
Maria, Sleep Baby

A: Classical, Sleepy, Mellow, Gentle, Soft, Johannes Brahms,
Instrumental, Brahms Lullaby, Sleep Baby Sleep.
```

Or for *I'm Not the Only One* by Sam Smith, to avoid the language model to wrongly interpret the "soul" tag in as in "soul-searching," I can prompt it to find basic information of the song first:

```
Q: Find the following information about the song "I'm Not the Only
One" by Sam Smith: instruments played in the song, vocalist, main
genre, and related genres.

A: Instruments Played: Piano, guitar, drums, bass, strings
Vocalist: Sam Smith
Main Genre: Pop
```

```
Related Genres: Soul, R&B, Adult Contemporary

Q: With the information above, describe a digital art painting
based on these keywords: Sad, Honest, Melancholy, Soul, Pop, Songs,
Ballad, Love, Listen.

A: This digital art painting depicts a lonely figure standing in
a dark room, illuminated only by a single lightbulb. The figure
is wearing a white dress, and their head is bowed in a sad and
melancholic pose. The walls of the room are lined with records,
each one representing a different genre of music, from soul to pop.
In the center of the room is a record player, playing a slow and
honest ballad about love. The figure stands in the center of the
room, listening to the song, and feeling the emotions it brings.
The painting is a reminder of the power of music to evoke strong
emotions, and to bring comfort in times of sadness.
```

The exact chain of prompts for generated knowledge that apply to every song is to be studied.

For method 3, collecting more data for the Imagination dataset is not the only necessary improvement toward the ideal model. So far, the pictures generated (by both methods 2 and 3) have been track-level ones, meaning that they are meant to fit the input song as a whole, not specific sections of the song. However, for songs of long duration, there can be different emotions and contents expressed at different times. Given that MusCALL processes each audio by taking an arbitrary 20 seconds as a representative, I suppose that targeting specific sections and giving each its own scenic caption is a feasible task in future development of method 3.

In addition, MusCALL does not have the ability to recognize or understand lyrics. One straightforward way to add this capability is to fetch song lyrics from online sources. However, to give each section a caption, as proposed above, there needs to be speech-to-text AI to recognize the lyrics at specific timestamps. Online song lyrics can then be used to check against and improve this model's output. Besides, rankings of captions, a task that MusCALL is able to perform well, should not be the end goal, because this limits the possibility of output captions. Instead, the recommended captions and their weights can be used as references for a decoder to generate new captions. To ultimately generate fitting scenic captions, the recognition of lyrics within framed sections, and the infusion of these lyrics with the similarity weights of reference captions are all interesting areas for further research. Finally, throughout this paper, I have assessed each method on three criteria: interpretability, realisticity, and chance of results' relatability. However, there are no metrics to quantify them, which easily leads to bias in the evaluation of results. At heart, these criteria reflect my desire to create matchings that are aligned with users. Past research in language models has used reinforcement learning using human feedback to align the models with user intent, gauging this "alignment" metric directly with the responses from human evaluators [38, 39]. If resources allow, this approach can be applied to evaluate results, especially the generated captions once a decoder is developed.[1]

[1]Human feedback can also help evaluate method 2's results, and compare methods 2 and 3 more objectively.

6 Conclusion

In this paper, I attempted at three different methods to generate art that evokes the feeling of music. Method 1, which converts MIDI data to art data, helps redirect me to focus on the interpretability, realisticity, and ultimately relatability of the visualizations to the general audience. Built upon this realization, method 2, which generates images based on song tags, was able to produce concrete images that incorporate different aspects of a musical piece such as the performing artists, genres, time eras, and emotions. Finally, method 3 aims to create typical scenes that evoke the same feelings as the music, a common type of vision that music listeners have in mind. Through a workaround approach, I proved that this objective is possible, but needs a lot more research and development to implement. Overall, this project can be considered a solid starting point for a new field of computational research that encapsulates human imagination during music listening, thereby connecting music and art in a naturalistic way.

Bibliography

[1] William Moritz. “The Dream of Color Music, and Machines that made it Possible”. In: *Animation World* 2.1 (Apr. 1997).

[2] *Scale (music)*. Apr. 2023. URL: https://en.wikipedia.org/w/index.php?title=Scale_(music)&oldid=1148180738.

[3] Helena Asprou. *5 artworks that inspired classical masterpieces*. URL: https://www.classicfm.com/discover-music/artworks-inspired-classical-music/.

[4] Golan Levin. “Painterly interfaces for audiovisual performance”. PhD thesis. Massachusetts Institute of Technology, 2000.

[5] Jack Ox. “2 performances in the 21st century virtual color organ”. In: *Proceedings of the 4th Conference on Creativity & Cognition*. 2002, pp. 20–24.

[6] Zune Lee, Jonathan Berger, and Woon Seung Yeo. “Mapping Sound to Image in Interactive Multimedia Art”. In: *Online document* (2004).

[7] RS Wilson, Jonathan Berger, and WS Yeo. “A flexible framework for real-time sonification with sonart”. In: Georgia Institute of Technology. 2004.

[8] Woon Seung Yeo and Jonathan Berger. “Application of image sonification methods to music”. In: *ICMC*. 2005.

[9] Peter Beyls. “The musical universe of cellular automata”. In: *Proceedings of the 1989 International Computer Music Conference, San Francisco, September 1989*. 1989.

[10] Dale Millen. “Cellular automata music”. In: *Proceedings of the 1990 International Computer Music Conference, San Francisco, September 1990*. 1990.

[11] Alexis Kirke and Eduardo Reck Miranda. “Capturing the aesthetic: Radial mappings for cellular automata music”. In: *J. ITC Sangeet Res. Acad* 21 (2007), pp. 15–23.

[12] Lucy Tran. *Music in Canvas Form*. URL: https://docs.google.com/document/d/1MpFFxEt8dtbxLp5uhjKlgN-D5OL18JBOdboSZnlQ8Hc/edit?usp=sharing.

[13] THOMAS Ciufo. “Real-time sound/image manipulation and mapping in a performance setting”. In: *Proc. MAXIS Festival of Sound and Experimental Music*. 2002.

[14] John Gibson. *The MIDI Standard: Introduction to MIDI and Computer Music*. Center for Electronic and Computer Music: Jacobs School of Music. URL: https://cecm.indiana.edu/361/midi.html.

[15] John Elligner. *MIDI Basics*. Carleton College, MN, USA. URL: https://people.carleton.edu/~jellinge/m208w14/pdf/02MIDIBasics_doc.pdf.

[16] Dominique Vandenneucker. *MIDI Tutorial For Programmers*. Arpege Music. URL: https://www.cs.cmu.edu/~music/cmsip/readings/MIDI%20tutorial%20for%20programmers.html.

[17] MIDI Manufacturers Association et al. "Standard MIDI files 1.0". In: *RP-001* (1996).

[18] Midi Association. *GM 1 Sound Set*. URL: https://www.midi.org/specifications-old/item/gm-level-1-sound-set.

[19] en. URL: https://apkpure.com/music-visualizer/be.tmde.musicvisualizer.

[20] en. URL: http://splusopm.weebly.com/1/post/2022/11/advanced-audio-visualizer-software.html.

[21] en. URL: https://wallpapercave.com/audio-visualizer-wallpapers.

[22] Dorien Koelemeijer. *Dorien Koelemeijer: Audio visualizer*. Mar. 2014. URL: http://cantseemyownface.blogspot.com/2014/03/audio-visualizer.html.

[23] Taemin Lee, Hyunki Lim, Dae-Won Kim, Sunkyu Hwang, and Kyunghyun Yoon. "System for matching paintings with music based on emotions". In: *SIGGRAPH ASIA 2016 Technical Briefs*. 2016, pp. 1–4.

[24] Cheng-Te Li and Man-Kwan Shan. "Emotion-based impressionism slideshow with automatic music accompaniment". In: *Proceedings of the 15th ACM international conference on Multimedia*. 2007, pp. 839–842.

[25] Baixi Xing, Jian Dou, Qing Huang, and Huahao Si. "Stylized Image Generation based on Music-image Synesthesia Emotional Style Transfer using CNN Network." In: *KSII Transactions on Internet & Information Systems* 15.4 (2021).

[26] en. URL: https://platform.openai.com.

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, and Scott Gray. *DALL·E: Creating images from text*. en-US. URL: https://openai.com/research/dall-e.

[28] Sergio Oramas, Mohamed Sordo, Luis Espinosa-Anke, and Xavier Serra. "A semantic-based approach for artist similarity". In: *Müller M, Wiering F, editors. Proceedings of the 16th International Society for Music Information Retrieval (ISMIR) Conference; 2015 Oct 26-Oct 30; Malaga, Spain.[Sl]: International Society for Music Information Retrieval; 2015. p. 100-6.* International Society for Music Information Retrieval (ISMIR). 2015.

[29] Keunwoo Choi, George Fazekas, Brian McFee, Kyunghyun Cho, and Mark Sandler. "Towards music captioning: Generating music playlist descriptions". In: *arXiv preprint arXiv:1608.04868* (2016).

[30] Sergio Oramas, Luis Espinosa-Anke, Mohamed Sordo, Horacio Saggion, and Xavier Serra. "Information extraction for knowledge base construction in the music domain". In: *Data & Knowledge Engineering* 106 (2016), pp. 70–83.

[31] Jordi Pons and Xavier Serra. "musicnn: Pre-trained convolutional neural networks for music audio tagging". In: *arXiv preprint arXiv:1909.06654* (2019).

[32] Minz Won, Sanghyuk Chun, Oriol Nieto Caballero, and Xavier Serra. "Automatic music tagging with harmonic cnn". In: (2019).

[33] Byungsoo Jeon, Chanju Kim, Adrian Kim, Dongwon Kim, Jangyeon Park, and Jung-Woo Ha. "Music Emotion Recognition via End-to-End Multimodal Neural Networks." In: *RecSys Posters*. 2017.

[34] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. "Music mood detection based on audio and lyrics with deep neural net". In: *arXiv preprint arXiv:1809.07276* (2018).

[35] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. "Music mood detection based on audio and lyrics with deep neural net". In: *arXiv preprint arXiv:1809.07276* (2018).

[36] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. "Contrastive audio-language learning for music". In: *arXiv preprint arXiv:2208.12208* (2022).

[37] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. "Generated knowledge prompting for commonsense reasoning". In: *arXiv preprint arXiv:2110.08387* (2021).

[38] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021.

[39] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.