Visualize Music Using Generative Arts

1st Blind Review
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Playing and listening to music is one of the most universal forms of communication and entertainment across cultures. This can largely be credited to the sense of synesthesia, or the combining of senses. Based on this concept of synesthesia, we want to explore whether generative AI can create visual representations for music. The aim is to inspire the user's imagination and enhance the user experience when enjoying music. Our approach has the following steps: (a) Music is analyzed and classified into multiple dimensions (including instruments, emotion, tempo, pitch range, harmony, and dynamics) to produce textual descriptions. (b) The texts form inputs of machine models that can predict the genre of the input audio. (c) The prompts are inputs of generative machine models to create visual representations. The visual representations are continuously updated as the music plays, ensuring that the visual effects aptly mirror the musical changes. A comprehensive user study with 88 users confirms that our approach is able to generate visual art reflecting the music pieces. From a list of images covering both abstract images and realistic images, users considered that our system-generated images can better represent pieces of music than human-chosen images. It suggests that generative arts can become a promising method to enhance users' listening experience while enjoying music. Our method provides a new approach to visualize music and to enjoy music through generative arts.**

***Index Terms*—Visualize Music; Generative Models of Artificial Intelligence**

I. Introduction

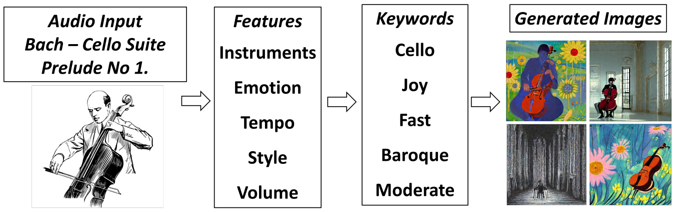


Fig. 1: The proposed method has three steps: Music Analysis, Prompt Generation, and Image Generation. The images change as the music is played. Image Source: [1]

The history of music could be as old as humanity itself [2]. The oldest music instruments can be dated back to at least 43,000 years ago. In the United Kingdom 85% of children have played musical instruments [3]. Music and the entertainment industry contributes to employing 5 million people in the USA, including dancers, music producers, recording engineers, actors, costume designers, etc [4]. Music is a multifaceted form of expressions and can be felt through humans' multiple senses. Listening to music is also heavily involved with visual senses: The color spectrum and music have been studied for correlation [5]. It is a common notion that music can express imagery either through music composition techniques or the addition of lyrics to tell a story. Composing classical music has the notion of visualizing figurative arts [6].

Using generative models to produce arts from music has several advantages. First, this process can be customized by users' preferences: users may add or remove words interactively to produce different visual effects. Generative arts can be more appealing than visual images of music performance because users can only passively watch performing videos without interaction. Second, generative arts can be produced quickly and inexpensively. As a result, this gives the potential to attract more audiences to enjoy music.

The main contributions of this study are the following: (a) The creation of a software system to generate human interpreted images from audio input. (b) A comprehensive user evaluation of the generated images against human-chosen images. On image generation, we use generative artificial intelligence (GAI) to create images that can represent music. We convert music to visual effects in three steps, as illustrated in Figure 1: (a) Analyze the music based on multiple factors (such as instruments, tempo, pitch, and dynamics). (b) Create textural descriptions. Many existing tools can meet this purpose, for example, Microsoft's MusicBERT [7], Spotify's music classifier [8], and OpenSmile's audEERING feature extraction [9]. (c) Visualize music by using generative arts based on the textual descriptions. This can be achieved by using pre-trained diffusion models [10]. The visual representations are updated while the music is played. Music often goes through multiple phases with different characteristics. For example, a symphony usually has four movements, and each movement can have sections with different rhythmic and melodic patterns to create various emotions. Moreover, some instruments may be dominant for several sections in a movement. The visualization should reflect these dynamic changes of the music.

On user evaluation, we evaluate the effectiveness of using generative arts to represent music based on our system. This study examines two main aspects: (1) Can users distinguish images that reflect the music? (2) When different images are presented, do users select the images generated by our system? In addition to the system-generated images, we also manually select some images to detect possible biases due to the styles of images. The online survey contains questions to examine whether the users prefer the system-generated images or the manually selected images. To align with users' preferences and identify potential bias, the survey contains a set of questions for validation. The survey was open for one month and 88 people participated. Among their selections, 58% of respondents prefer the images generated by our system. This is significantly higher than the 35% of images not generated by the system. The remaining 7% select no images. The notable difference (23%) along with a p-value of less then 0.01 determined by a chi-squared test indicates that generative arts offer a promising solution improving users' enjoyment while listening to music.

II. Related Work

A. Generative Artificial Intelligence

Diffusion models have made recent developments into the field of computer vision [11]; image generation is one of the most common applications. Stable Diffusion [12] has been widely used for AI generated images. Their model is primarily based on using prompts as inputs; these prompts allow images to be retroactively adjusted [13].

The figurative notion of music has been investigated in various studies. Braganca et al. [14] evaluate the cross-modal association of sensations and their relationship to musical perception with a focus on synesthesia. Actis-Grosso et al. [6] explore similarities between music and visual arts. Modem Works [15] utilizes Stable Diffusion and Teenage Engineering's OP-Z track sequencer and synthesizer to translate music into imagery. Cowles [16] experiments on pairing audio with visual stimuli; correlations were found between subjects choosing certain selected images and music. Gayen et al. [17] find common trends in painted depictions of music with contrasting emotional tones. Wehner [18] uses paintings and music from Paul Klee to test and evaluate the ability of people to correlate paintings with music. These examples demonstrate the relationships between music and visual arts. Inspired by prior work that shows the close relationships between visual art and music, this paper further uses *generative machine models* to produce visual representations based on input music.

B. Visualizing Music

Identifying music through a generative model can be done through several methods depending on how music data is interpreted. The common forms of music data are MIDI files and signal processing techniques like Mel Spectrograms [19]. The former represents music as a digitized pattern of notes and the latter represents music as a raw 2D spectrogram of an audio file. MusicBert [7] uses MIDI to develop a "Symbolic Music Representation" to analyze music through patterns of notes. Riffusion [20] (a fine-tuned Stable Diffusion model) uses Mel Spectrograms to analyze music as images to train a convolutional neural network (CNN) to match to existing spectrograms. Such tools and their models can be effectively trained to classify raw audio inputs into music genres; however, an issue arises when it comes to expanding these classifications into descriptive image generation. The use of prompts as descriptive tags, aiming to apply them equally to both auditory and visual experiences, reintroduces the concept of synesthesia [14]. The subjective nature of synesthetic perceptions acts as an abstract association in achieving seamless audio-to-image generation.

C. Comparisons

Several methods can use AI models to generate images from music. Modem's OP-Z/Stable Diffusion [15] utilizes prompt engineering to provide descriptiveness in imagery. The method considers pitch and tempo but lacks details, such as genre, instruments, or contextual clues from chord progressions. Liu et al. [21] create "Generative Disco" using human-chosen prompts to generate images. This method takes a text-to-image approach rather than music-to-image. Betin [22] stylizes an existing image based on a musical input. The method serves primarily as an abstract image adjustment based on existing image's structure and changes the color styling based on musical sound waves. Table I compares the proposed method with existing methods. Our goal is to create imagery that is more connected to music, improving the user experience.

TABLE I: Comparison of Methods.

Method	Approach	Features
Modem [15]	Prompt Generation	Pre-defined Abstract Images
Liu [21]	Prompt Utilization (lyrics)	Text-to-Image
Betin [22]	Signal Processing	Images Are Not Generated
This paper	Prompt Generation	Music-to-Image

III. Visualize Music by Generative Artificial Intelligence

A. Generative Methods

Our approach entails interpreting music elements and incorporating additional features, such as chord-analysis, to train based on the styles of existing music. To generate images from music, text prompts serve as an intermediary bridging the gap between the two mediums (sound and visual). We utilize the existing image-generative model Stable Diffusion to create images from these prompts. The overall software flow can be seen in Figure 2 and will be discussed in the following subsections. This method starts from a raw audio file. The music is processed by (1) Spotify's Basic Pitch [8] to extract MIDI features through chords and pitch and (2) OpenSMILE [9] to extract spectrogram features. Through these features, we predict stylistic and emotional values through a trained neural network and then generate prompts from the predictions. The prompts are then passed into Stable Diffusion to generate images.

B. Music Analysis

To generate effective prompts, we start with analyzing several different metrics of the music's corresponding audio. We calculate both temporal and physical statistics about the audio using spectrogram analysis such as RMS amplitude, spectral width and centroid, etc., as well as musical data such as pitch, overall chord patterns and tempo. OpenSMILE [9] and Spotify's Basic Pitch [8] manage these calculations. We then feed these calculated metrics into fully connected neural networks to provide intermediary statistics about the music as prompts. We use feed-forward neural networks to estimate the genre of the music piece and valence-arousal emotion values.

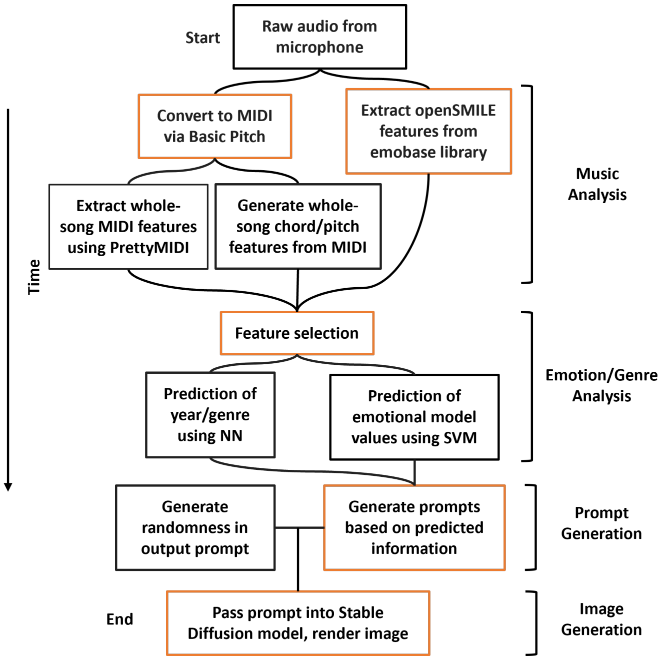

Start
Raw audio from microphone
Convert to MIDI via Basic Pitch
Extract openSMILE features from emobase library
Extract whole-song MIDI features using PrettyMIDI
Generate whole-song chord/pitch features from MIDI
Music Analysis
Time
Feature selection
Prediction of year/genre using NN
Prediction of emotional model values using SVM
Emotion/Genre Analysis
Generate randomness in output prompt
Generate prompts based on predicted information
Prompt Generation
End
Pass prompt into Stable Diffusion model, render image
Image Generation

Fig. 2: The procedure to visualize music. The method starts from raw audio analysis and extracting musical features to train a neural network. The network predicts music genre, time period, and emotional values which are used to generate prompts. Prompts are adjusted with random seed descriptions to add image variation. The prompts are then passed into Stable Diffusion for image generation.

Based on these estimates, we use k-nearest neighbors with a k-value of 1 to assign a set of prompt words to the music (such as genre, emotional words, colors, etc).

C. Prompt Generation

Emotions are measured in terms of valence (how positive or negative an emotion feels) and arousal (how intensely the emotion is felt) via the Valence-Arousal Model [23]. The prompts change the lighting and colors in the generated artwork. For example, when an emotion like “anger” is detected (one with a high valence and arousal), the generated image will use saturated colors such as vibrant reds or dark purples and black. The subject of the artwork will be also based on the genre of the input music. Based on ”anger” as the emotion, a classical style piece might generate an image of a 19th century ballroom in hues of red. We will produce images using various prompts for each genre, including solo performances, chamber music, symphony orchestras (including concertos), choirs (accompanied by piano or orchestra), and operas/ballets. By adjusting the prompts through “prompt modifiers” [13], we can generate a diverse array of images.

D. Image Generation

Finally, once these prompts are generated, we introduce some random image-related words into the prompt (such as camera angle, movement, framing, etc.) to add variation to the resultant image. LLMs (Large Language Models) can comprehend valence-arousal emotion values and provide feedback on the represented emotions. Therefore, in this process, the initially obtained valence-arousal emotion values will be collectively inputted into the LLMs. Once these fundamental elements composing the prompt are acquired, the GPT-4 [24] LLM will be introduced to assist in the final prompt generation. Throughout this process, prompt engineering will be employed, assigning the role of “An artist who can connect emotions from music to pictures” to the LLM. Additionally, throughout this process, the LLM will be emphasized to consistently maintain the alignment of emotions conveyed by both pictures and music. Once we have our final prompt, we then feed it to a diffusion-type image-generating model to get our set of images.

IV. Human-subject Evaluation and Study Results

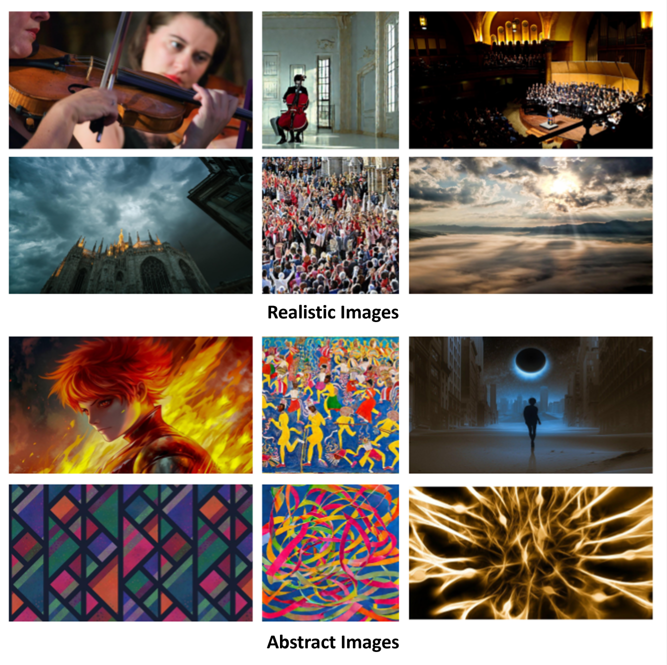


Fig. 3: Different types of images: Realistic vs. Abstract. Note that the four middle images of the figure are system-generated, and all other images included are from the following image sources: [25], [26], [27], [28], [29], [30], [31], [32].

To evaluate the efficacy of our method, we conduct an online human-subject study to answer the question: “Can generative visual arts reflect the rich expressions of music?”. We recruit users to evaluate the visual arts generated based on different pieces of music. After hearing a piece of music, we ask a user to select an image that can best reflect the music. The options include three types of images (1) generated by our system, (2) chosen by human (members in this research team), (3) generated based on other pieces of music. If our system-generated images are preferable by the majority of the users, our system can effectively produce visual representations reflecting the music.

A. User Profiles

We send out emails to groups of students and faculty in our university and collect 88 responses. Among them 62.5% are male and 31.8% are female. Most subjects (84.1%) are within the age range of 18-24. Many of our participants are either student musicians (35.2%) or play an instrument for leisure (33.0%).

B. Music

This study uses 15 pieces of classical music. Each piece is 10 seconds long. The pieces are chosen from 5 different areas categorized by a number of performers as well as type of containing instruments. The five areas are as follows: choir, opera and ballet, chamber music, solo performance, and larger group of ensemble (orchestra or band). Three music clips are selected from each category, with each music clip chosen being a well-known and representative piece of its defined category i.e. Beethoven's 9th Symphony (Choir) and Bach Cello Suite No. 1 Prelude (Solo). Overall, our methodology in selecting the music pieces for this survey includes considering a diverse set of musical pieces such that our system can be tested most effectively.

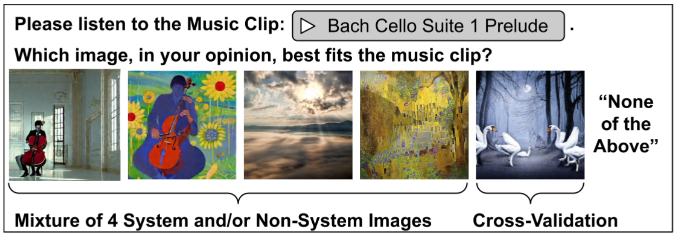
C. Visual Representations of Music

For each piece of music, our system generates six images (per trial). Additionally, we select six images manually from three online image repositories: Pexels, Pixabay, and Unsplash. These human-chosen images also reflect the music pieces based on judgement by this team's musicians that are knowledgeable about the pieces. The manually selected images are used for comparison against the system-generated images. If the users prefer system-generated images to human-chosen images, it suggests that our system, with images from generative models of artificial intelligence, are more expressive than those manually selected images. This in turn suggests the viability of generated images in accurately representing music on human standards. Also, to ensure that users can select the images that truly represent the specific piece of music, we include a system-generated image from a different piece of music (distraction). This image does not reflect the current music. This distraction aims to confirm that users can distinguish if an image represents the music or not. In total, for each piece of music, thirteen images are available.

This study considers images of different styles to avoid possible preference bias due to styles. We classify the images into abstract and realistic. Realistic arts aim to depict the subject matter with a high degree of accuracy and fidelity to its real-world appearance; abstract forms use colors, shapes, lines, and forms to convey emotions, ideas, or concepts. A user may have a strong preference for one certain style. To ensure we are comparing similar styles of images, we categorize each image as either realistic or abstract. Figure 3 shows several examples. In total, the survey includes 82 photos or realistic images and 113 abstract images, total 195 images.

D. Questionnaire

The purpose of the survey is to confirm (1) our system can generate images that represent the music, and (2) users do not choose the visual arts that do not reflect the music. The online survey includes 15 questions. Each user receives a random selection of 10 questions. One additional question measures users' preferences of subjectivity (thus, 11 questions per user). Figure 4 is an example of a question. The question includes a 10-second music clip. When the user clicks the button, the music is played. The survey (using Qualtrics) selects four images that may be generated by our system (trial, also called *system-generated*) or human-chosen. Additionally, one distraction image is included to detect style bias. The user may also select "None of the images".


Please listen to the Music Clip: ▷ Bach Cello Suite 1 Prelude.

Which image, in your opinion, best fits the music clip?

Fig. 4: A sample question. The user was asked to choose the image best that represents the music. Image sources: [28], [33].

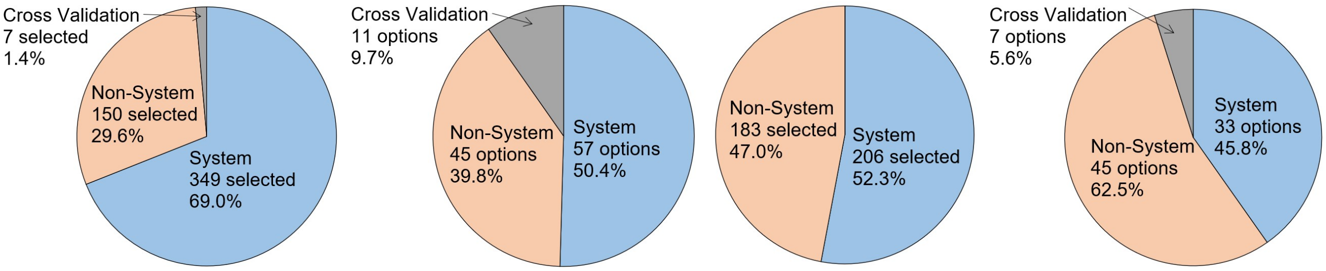
E. Result and Analysis

TABLE II: Proportion of Images Chosen & Expected Values.

Subjectivity Level:	Realistic	Abstract
System Expected %	40.2%	50.4%
System User Chosen %	53.0%	69.0%
Non-System Expected %	54.9%	39.8%
Non-System User Chosen %	47.0%	29.6%
Distraction Expected %	4.9%	9.7%
Distraction User Chosen %	0.0%	1.4%
P-Value	< 0.01	< 0.01

Figure 5 shows user's preferences between system-generated and human-chosen images as representations of the given music clips, as well as their subjectivity level preferences. For both image categories, subjective and realistic, users select the system-generated images at high percentages than the percentages of the options. Figure 5 (a) and (b) show the percentages of selections and options of abstract images. The images generated by our system are 50.4% of all image options, but counted to 69.0% of users' selections. In contrast, the other 49.6% of images only counted to 31.0% in users' selections. Similarly, for realistic images, users prefer system generated images (45.8% options counted to 52.3% users selected). This suggests that *users perceive the images generated by our system as better representations of the music than human-chosen images*.

Using Chi-square analysis shown in table II, there is a statistically significant preference for trial images found for both the realistic and abstract subjectivity levels. If users had selected images randomly, the expected numbers of system-generated images and non-system-generated images chosen would have followed the percentage makeups provided by the 195 total images included in the survey. However, the percentage of the system-generated images chosen by users is much higher than the actual percentage of images included in



(a) abstract, selected

(b) abstract, options

(c) realistic, selected

(d) realistic, options

Fig. 5: The survey results. (a)(b) Abstract style. (c)(d) Realistic style. (a) Users select system-generated images 349 times (68.97%) and images not generated by our system 150 times (29.64%). (b) Only 50.4% images are system generated. (c) Users select system-generated images 206 times (52.2%). (d) Only 45.8% images are system-generated. The users selected "None of the images" 61 times which is not represented in the pie charts.

the survey. As such, the p-values for both realistic and abstract images are less than 0.01. Consequently, we conclude that our system-generated images are preferred by users in comparison to human-chosen images. The percentage of distraction images chosen by the users is also much lower than the expected percentage, signifying that users are able to tell which images do not reflect the music. In each question, there is one distraction image out of 5 possible images. If users randomly choose a image, we should expect the proportion of distraction images selected to be slightly lower than 20% (due to the "None of the Above" option available to users). However, the total percentage of distraction images chosen during the survey was less than 1%. Overall, the total percentage of trial images chosen in the survey is 58%, the percentage of non-trial images chosen is 35%, and the remaining percentage is comprised of "None of the Above" choices. The total number of selections by users are 7 + 150 + 349 + 183 + 206 + 61 (None of the Above) = 956. Users select system-generated images 349 + 206 = 555 times. The ratio is $\frac{555}{956} = 58\%$. Users select non-system images 150 + 183 = 333 times. The ratio is $\frac{333}{956} = 35\%$.

The p-value for the total survey results across both subjectivity levels is less than 0.01. We can conclude that there is a statistically significant preference for trial image. Since users prefer trial images to non-trial images as a representation of the given music clips, this signifies that our system creates effective visual representations of music that align with user opinions. Additionally, the lack of distraction images chosen within our survey demonstrates that users are able to tell which images correspond to the musical clips. This suggests that the system-generated images are not preferred to human-chosen images due to type difference, but due to a meaningful mechanism of representing music.

Although our overall results show that 58% of user preference is for system-generated images, this percentage describes the total results, not the individual results for each music piece. Among the 15 music pieces in our survey, each of these pieces receives a different level of preference for system-generated images as shown in Figure 6. The piece in our survey with the highest proportion of system-generated images is Albeniz's Asturias, where $\frac{50}{65} = 76.9\%$ of the images selected by users for this piece are system-generated. This signifies that Asturias is the piece that provides the best system results in our survey. The piece with the lowest proportion of system-generated images is Tchaikovsky's Piano Concerto No. 1, with $\frac{22}{62} = 35.5\%$ of images chosen by users for this piece being system-generated. Therefore, Tchaikovsky's Piano Concerto No. 1 is the piece showing the worst performance of our system-generated images. There is a large difference between the largest and smallest percentage of system-generated images chosen between pieces, suggesting that our system does a better job of visualizing some musical pieces than others.

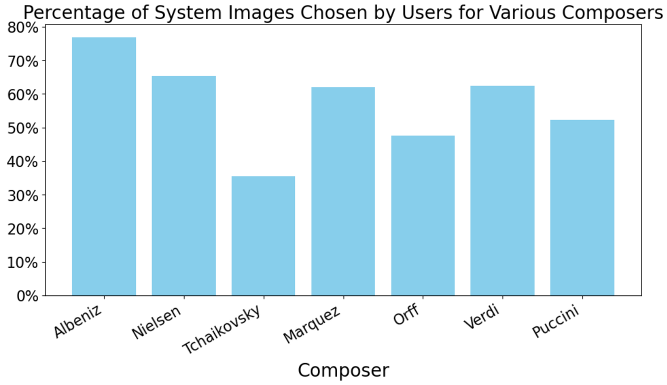


Fig. 6: Percentages of system-generated chosen by users for different composers. The figure shows 7 of the 15 composers in our survey.

V. Discussion

A. *Limitations*

The p-values for both the abstract and realistic subjectivity levels are less than 0.01. We conclude that there is a statistically meaningful preference for system-generated images as opposed to human-chosen images. However, there are several limitations found both in the selected user base for our survey as well as through the organization of our survey questions.

The majority of our users fall into the age range of 18-

25 (84.1%) because we recruited university students. Additionally, the majority of our users are either White or Asian (91.0%), and the majority (69.3%) have played music instruments. Our future work may analyze the relationships of user demographic and musical experience. Finally, this study considers only classical music. A future study should consider other types of music, such as jazz, rock, and pop.

B. Applications

There lies great opportunity in image generation for entertainment and enhancing the user experience when listening to music. Real time implementations can decorate a space being used for social events (i.e. karaoke, clubs, parties) as a more immersive substitute to music videos, ambient lighting, or still images. A musician can efficiently provide a visual experience to their performance that surpasses their own capabilities. The generated images can provide users with hearing-impairments a visual outlet to enjoy music. Other works have shown these possibilities like with Liu's "Generative Disco" [21] or Betin's "Visualizing Sound with AI" [22] providing insight into assisting musician workflow or decorating a space respectively. Our method of processing raw audio into descriptive prompts can provide a more human-interpreted image quality in these applications.

VI. Conclusion

This paper presents a study using generative artificial intelligence to visualize music. Our system analyzes music by multiple elements, such as instruments, tempo, emotion, pitch, etc., and generates text prompts. The prompts serves as inputs of diffusion models to produce images. A user study indicates that this approach can effectively reflect the rich expression of music.

Acknowledgments

We appreciate the support from the sponsors and the people that participated in the survey.

References

[1] Cellist man clipart, music vintage. https://openverse.org/image/7962407e-1be8-4123-a3d7-7b1449f65c3b.

[2] Evan Andrews. What is the oldest known piece of music? https://www.history.com/news/what-is-the-oldest-known-piece-of-music, September 2018.

[3] Simon Hume and Emma Wells. Making music: Teaching, learning, and playing in the uk. https://www.musicmark.org.uk/wp-content/uploads/2014_making_music.pdf, September 2014.

[4] IBISWorld - Industry Market Research, Reports, and Statistics, August 2022.

[5] Charles Spence and Nicola Di Stefano. Coloured hearing, colour music, colour organs, and the search for perceptually meaningful correspondences between colour and sound. *i-Perception*, 13(3):20416695221092802, 2022. PMID: 35572076.

[6] Rossana Actis-Grosso, Carlotta Lega, Alessandro Zani, Olga Daneyko, Zaira Cattaneo, and Daniele Zavagno. Can music be figurative? exploring the possibility of crossmodal similarities between music and visual arts. *Psihologija*, 50:285–306, 01 2017.

[7] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training, June 2021. arXiv:2106.05630 [cs].

[8] Rachel M. Bittner, Juan José Bosch, David Rubinstein, Gabriel Meseguer-Brocal, and Sebastian Ewert. A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, 2022.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, page 1459–1462. Association for Computing Machinery, 2010.

[10] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, March 2022. arXiv:2112.10741 [cs].

[11] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, September 2023. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[13] Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour amp; Information Technology*, page 1–14, November 2023.

[14] Guilherme Francisco F Bragança, João Gabriel Marques Fonseca, and Paulo Caramelli. Synesthesia and music perception. *Dementia & neuropsychologia*, 9:16–23, 2015.

[15] Modem. Op-z stable diffusion. https://modemworks.com/projects/op-z-stable-diffusion/, Jan 2023.

[16] John T. Cowles. An experimental study of the pairing of certain auditory and visual stimuli. *Journal of Experimental Psychology*, 18(4):461–469, 1935.

[17] Pinaki Gayen, Junmoni Borgohain, and Priyadarshi Patnaik. *The Influence of Music on Image Making: An Exploration of Intermediality Between Music Interpretation and Figurative Representation*, pages 285–293. 06 2021.

[18] Walter L. Wehner. The relation between six paintings by paul klee and selected musical compositions. *Journal of Research in Music Education*, 14(3):220–224, 1966.

[19] Hugo B. Lima, Carlos G. R. Dos Santos, and Bianchi S. Meiguins. A Survey of Music Visualization Techniques. *ACM Computing Surveys*, 54(7):143:1–143:29, July 2021.

[20] Seth Forsgren and Hayk Martiros. Riffusion - Stable diffusion for real-time music generation. https://github.com/riffusion/riffusion, 2022.

[21] Vivian Liu, Tao Long, Nathan Raw, and Lydia Chilton. Generative disco: Text-to-video generation for music visualization, 2023. arXiv:2304.08551 [cs].

[22] Vasily Betin. Visualizing sound with ai. *Medium*, May 2020.

[23] Saikat Basu, Nabakumar Jana, Arnab Bag, Mahadevappa M, Jayanta Mukherjee, Somesh Kumar, and Rajlakshmi Guha. Emotion recognition based on physiological signals using valence-arousal model. In *2015 Third International Conference on Image Information Processing (ICIIP)*, pages 50–55, 2015.

[24] Josh Achiam et. al. Gpt-4 technical report. Technical report, OpenAI, 2023. arXiv:2303.08774 [cs].

[25] Pascal Bernardon. Person playing violin. https://unsplash.com/photos/person-playing-violin-2BDIGA-DnlE, August 2018.

[26] Omar Flores. People sitting on chair inside building. https://unsplash.com/photos/people-sitting-on-chair-inside-building-AndwyJNdk1k, January 2021.

[27] ArtHouse Studio. Magnificent milan cathedral with lit up spires on gloomy day. https://www.pexels.com/photo/magnificent-milan-cathedral-with-lit-up-spires-on-gloomy-day-4329928/, June 2016.

[28] Pixabay. Light sun cloud japan. https://www.pexels.com/photo/light-sun-cloud-japan-45848/, February 2016.

[29] CharVera. Cosplay otaku fan art. https://pixabay.com/illustrations/cosplay-otaku-fanart-7959696/, May 2023.

[30] ntnvnc. Fantasy eclipse atmosphere dark. https://pixabay.com/illustrations/fantasy-eclipse-atmosphere-dark-3533325/, July 2018.

[31] chenspec. Background geometric pattern. ttps://pixabay.com/illustrations/background-geometric-pattern-7983832/, May 2023.

[32] geralt. Annoy cells dendrites sepia. https://pixabay.com/illustrations/annoy-cells-dendrites-sepia-346928/, May 2014.

[33] Prawny. Abstract painting country golden. https://pixabay.com/illustrations/abstract-painting-country-golden-5985987/, February 2021.