Visualize Music Using Generative Arts

Brian Man-Kit Ng, Samantha Rose Sudhoff, Haichang Li, Joshua Kamphuis, Tim Nadolsky,
Yingjie Chen, Kristen Yeon-Ji Yun, and Yung-Hsiang Lu
Purdue University, West Lafayette, Indiana, USA.
{ng118, ssudhoff, li4560, jpkamphu, tnadolsk, victorchen, yun98, yunglu}@purdue.edu

***Abstract*—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

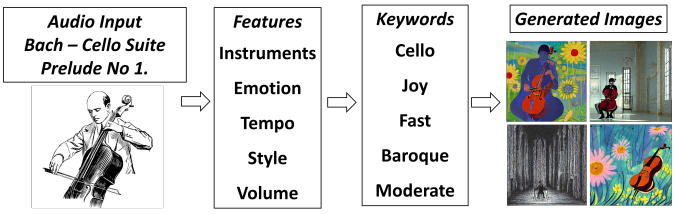


Fig. 1: The proposed method has three steps: Music Analysis, Prompt Generation, and Image Generation. The images change as the music is played. Non-AI Image Source: [1]

Music is a multifaceted form of expressions and can be felt through humans' multiple senses. Listening to music is heavily involved with visual senses: The color spectrum is related to music [2]. 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 [3]. Pairing music with the right visuals and vice versa can often lead to a more holistic entertainment experience [4], as done very often with various forms of media (live performing, Karaoke, music TV, cinema, live orchestras, etc.).

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 that better match the mood of the performer and theme of the music. Second, generative arts can be produced quickly and inexpensively. As a result, this can potentially give musicians a more flexible way to design the performing stage, and give audiences a richer experience while enjoying the show.

The original contributions of this study are the following: (a) The creation of a software system to autonomously generate representative images from music audio using generative artificial intelligence (GAI) methods. (b) A comprehensive user evaluation of the generated images comparing to human-chosen images. We convert music to visuals 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 using Spotify's music classifier Basic Pitch [5], and OpenSmile's audEERING feature extraction[6]. (c) Generative arts based on the textual descriptions using pre-trained diffusion models [7]. The visual representations can be updated in real-time 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 express various emotions and scenes. The generated images should reflect these dynamic changes in the music.

We used human subjects to evaluate the effectiveness of our system and examine two aspects: (1) Do these generated images reflect the music? (2) Do users prefer the images generated by our system? We generated both abstract images and realistic images, and compared these generated images with manually selected images. We used an online survey to examine whether the users prefer the system-generated images or the manually selected images. 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. The survey is available at https://ai4musicians.org/visualize.html.

II. Related Work

A. Generative Artificial Intelligence

Diffusion models have made recent developments into the field of computer vision [8]; image generation is one of the most common applications. Stable Diffusion [9] has been widely used for AI generated images. The model is primarily based on using prompts as inputs; these prompts allow images to be retroactively adjusted [10].

The visual notion of music has been investigated in several studies. Braganca et al. [11] evaluate the cross-modal association of sensations and their relationship to musical perception with a focus on synesthesia. Actis-Grosso et al. [3] explore similarities between music and visual arts. Modem Works [12] utilizes Stable Diffusion and Teenage Engineering's OP-Z track sequencer and synthesizer to translate music into imagery. Cowles [13] experiments on pairing audio with visual stimuli; correlations were found between subjects choosing certain selected images and music. Gayen et al. [14] find common trends in painted depictions of music with contrasting emotional tones. Wehner [15] uses paintings and music from Paul Klee to test and evaluate the ability of people to correlate paintings with music. Inspired by such prior works that show 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 (Musical Instrument Digital Interface) files and signal processing techniques like Mel Spectrograms [16]. The former represents music as a digitized pattern of notes and the latter represents music as a non-linear transformation on the frequency scale of an audio file. MusicBert [17] uses MIDI to develop a "Symbolic Music Representation" to analyze music through patterns of notes. Riffusion [18] (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 digitized 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 [11]. The subjective nature of synesthetic perceptions acts as an abstract association in achieving seamless audio-to-image generation.

C. Comparisons

Several methods have used AI models to generate images from music. Modem's OP-Z/Stable Diffusion [12] utilizes prompt engineering to provide imagery from solely MIDI inputs. Using MIDI considers basic music elements but lacks over-encompassing details such as genre, instrumentation, or contextual clues from chord progressions. As such, the results are mostly abstract images that lack contextual connection with the music. Liu et al. [19] create "Generative Disco" using human-chosen prompts to generate images. This method takes a text-to-image approach rather than music-to-image, and focuses on utilizing user-inputs and lyrics as a medium for determining prompts in generating images. It is labour intensive and will be hard to create images in real time. Betin [20] stylizes existing images based on an audio input in real-time. The method serves primarily as an abstract image adjustment based on existing image's structure and changes the color styling based on the physical elements of a Mel Spectrogram. Hence, the result is not full image generation, but rather image alteration. 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 [12]	Prompt Generation	MIDI Generated Images
Liu [19]	Prompt Utilization (lyrics)	Specialized Text-to-Image
Betin [20]	Signal Processing	Image Alteration
This paper	Prompt Generation	Real-Time Music-to-Image

III. Visualize Music by Generative Artificial Intelligence

Our approach entails interpreting musical 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 two mediums (sound and visual). The overall flow can be seen in Figure 2 and will be discussed in the following subsections.

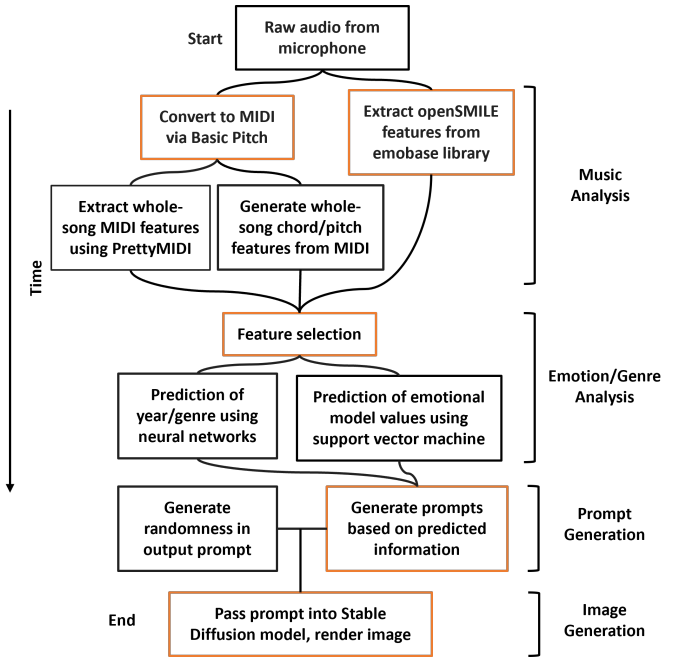


Fig. 2: The process of generating image from music. It starts with music analysis. A neural network predicts music genre, tempo, and emotional values. An prompts was generated from the prediction, and passed into Stable Diffusion for image generation.

A. Music Analysis

We start with analyzing several metrics form the music's audio recording and MIDI file. We calculate both temporal and physical statistics about the audio using spectrogram analysis such as root mean square (RMS) amplitude, spectral width and centroid, etc., as well as musical data such as pitch, overall chord patterns and tempo. We used Spotify's Basic Pitch [5] to extract MIDI features through chords and pitch, and OpenSMILE[6] to extract audio features.

B. Emotion/Genre Analysis

We then feed these calculated metrics into a fully connected neural network. We use feed-forward neural networks to estimate the genre of the music piece and valence-arousal emotion values. 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 [21]. This can be visualized as positive and negative values on a coordinate graph.

C. Prompt Generation

Based on these estimates, we use k-nearest neighbors to assign a set of prompt words to the music (such as genre, emotional words, colors, etc), where k is 1 as prompt features are relatively distinct. We would like these initial prompts to relate to the lighting and colors in the generated artwork. For example, when an emotion like "anger" is detected (one with a high positive valence and arousal), the generated image should 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. As in the case of Figure 3, the first passages of Beethoven Symphony No. 5 is classified with the emotional prompts of "angry", "aggressive", and "violent". This results in the images having a theme of either red or black hues. Additional analysis on the MIDI chords and MeL spectrograms defines the genre as a classical work, which contributes to the painted texture of the images. Further adjustment of the prompts through "prompt modifiers" [10] can help generate specific details and variations in the images. We 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.

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 [22] LLM will be introduced to assist in prompt engineering for more detailed image generation. Additionally, throughout this process, the LLM is emphasized to consistently maintain the alignment of emotions conveyed by both pictures and music. After 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: Examples of generated images from the system. Beethoven Symphony No. 5 is depicted with imagery of a thunder storm or a bird on fire, while the more mellow Mozart Violin Sonata No. 21 both indicate the violin instrumentation and also an overall brighter color palate.

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?", and "Do audiences like the generated visual?". In the study, we evaluated the visual arts generated from different pieces of music. After hearing a piece of music, a user selects 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 emails to students and faculty at Purdue 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 with each 10 seconds long. The pieces are chosen from 5 major classical music genres: choir, opera and ballet, chamber music, solo performance, and larger group of ensemble (orchestra or band). Three pieces per genre. These pieces are well-known and representative for its category i.e. Beethoven's 9th Symphony (Choir) and Bach Cello Suite No. 1 Prelude (Solo). When selecting the pieces, we considered a diverse set of musical features such that our system can be generalized broadly.

C. Visual Representations of Music

For each music piece, our system generates six images (per trial). For comparison, musicians in our team select

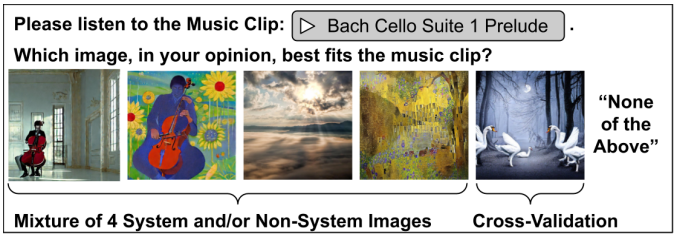


Fig. 4: A sample question. The user was asked to choose a image best fits the music. Non-AI Image sources: [23], [24].

six images manually from three online image repositories: Pexels, Pixabay, and Unsplash. These images also reflect the music pieces based on the musicians' judgement. 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 can generate images that are closer to the music 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 depict the subject matter with a high degree of 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. The survey includes 82 photos or realistic images and 113 abstract images, total 195 images.

D. Questionnaire

We designed 15 questions. During survey, a user receives 10 random questions plus one additional question measures users' preferences of subjectivity (toal 11 questions). Figure 4 is an example of a question. Each question includes a 10-second music clip. The user clicks the button to play the music. The system 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".

E. Result and Analysis

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. If users had selected images randomly, the expected numbers of system-generated images and non-system-generated images chosen would have followed the percentage

TABLE II: Proportion of Images Chosen & Expected Values.

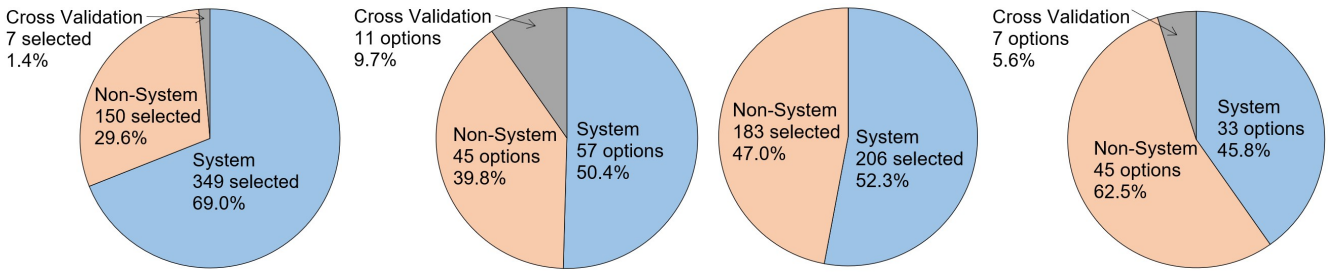
Subjectivity Level:	Realistic	Abstract
System Expected %	40.2%	50.4%
System User Chosen %	53.0%	69.0%
Non-System Expected %	54.9%	39.8%
Non-System User Chosen %	47.0%	29.6%
Distraction Expected %	4.9%	9.7%
Distraction User Chosen %	0.0%	1.4%
P-Value	< 0.01	< 0.01

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 the survey. Figure 5 (a) and (b) show the percentages of selections and options of abstract images. The images generated 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). Chi-square analysis (table II) shows that there is a statistically significant preference for trial images found for both the realistic and abstract images. The p-values for both realistic and abstract images are less than 0.01. Consequently, this suggests that *users perceive the images generated by our system as better representations of the music than human-chosen images*

For triangulation, we also examined if users are able to identify images that do not reflect the music. In each question, there is one distraction image out of 5 possible images. If users randomly choose an 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%, signifying that users are able to tell which images do not reflect the music.

Overall, the total percentage of system images chosen in the survey is 58%, the percentage of human-chosen 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 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 across both subjectivity levels is less than 0.01. This signifies that *our system creates effective visual representations of music that are more preferred by users*. Additionally, our distraction images test shows that users are able to tell which images are not correspond to the musical clips. This suggests that the *System-generated images are preferred over human-selected images not because of their type, but due to their meaningful representation of the music.*.

We further examined all the 15 music pieces used in this survey. 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 (best system performance) of system-generated images is Albeniz's Asturias,



(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.

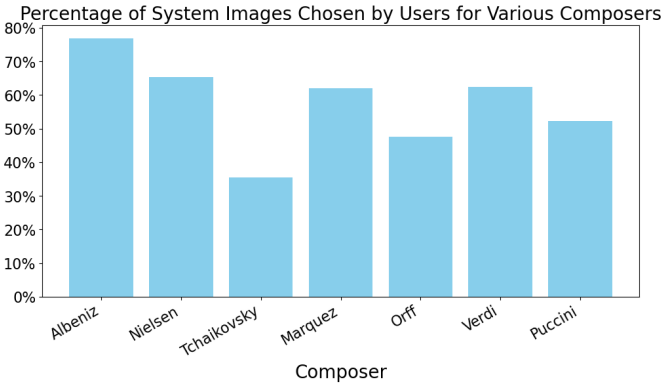


Fig. 6: Percentages of system-generated chosen by users for different composers. The figure shows 7 of the 15 composers in our survey.

where $\frac{50}{65} = 76.9\%$ of the images selected by users are system-generated. The piece with the lowest proportion (worst system performance) 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. There is a large difference between the largest and smallest percentage of system-generated images chosen between pieces, suggesting that our system may not able to equally visualize different types of musics.

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. Also, it seems our system's performance varies when dealing with different music. Is there a systematic difference (i.e. always perform worse on certain types of music), or just random error, still needs more investigation.

The majority of our users fall into the age range of 18-25 (84.1%) because the place (university) of this study. 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 along with defining a concrete qualitative evaluation of results with a more diverse study group. This study considers only classical music. A future study should consider other types of music, such as jazz, rock, and pop.

Fig. 7: Our system in live Cello performing

B. Applications

There lies a 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. Musicians can efficiently provide a visual experience to the 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" [19] or Betin's "Visualizing Sound with AI" [20]. Our method can provide human-interpreted image quality in these applications. Recently we have put our system in a live performing event (Fig. 7 https://www.youtube.com/

watch?v=LF172wWu2jU). The system runs smoothly. It saved a lot of effort from the performer in choosing images for the background visual effect of the music. The performer and audience feel the generated image at the background largely reflect the nature and characteristics of the music.

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, and generates text prompts. The prompts are then input to 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. This work is supported in part by NSF IIS-2326198 and by the CREATE program of Purdue. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

References

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

[2] 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.

[3] 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.

[4] Mats B Küssner and Tuomas Eerola. The content and functions of vivid and soothing visual imagery during music listening: Findings from a survey study. *Psychomusicology: Music, Mind, and Brain*, 29:90, 2019.

[5] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022.

[6] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, page 1459–1462, 2010.

[7] 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].

[8] 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.

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

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

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

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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].

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

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

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

[21] 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 *International Conference on Image Information Processing*, pages 50–55, 2015.

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

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

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