Music to Visual Cues: Exploring How AI Can Communicate A Song Visually

Yogie Permana
yogie@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Sabika Amalina
sabika@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Natalia Sempere
sempere@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

***Abstract* — In this study, we explore how artificial intelligence (AI) can generate visual representations of music by translating auditory features such as rhythm, melody and harmony into expressive visuals. We developed an AI framework that processes key audio features and derives mood-related descriptors, which are then used to generate visuals through a diffusion-based image generator. To assess the impact of these visualizations, we conducted a user study examining participants' emotional engagement and cognitive interpretation. Our results reveal varied preferences, with casual listeners favoring structured visuals and individuals with artistic backgrounds preferring abstract designs. These findings demonstrate both the potential and limitations of AI in capturing the affective and structural qualities of music. We contribute to ongoing discussions on AI-driven creative processes and highlight applications in accessibility and human-computer interaction. In future work, we aim to refine generative techniques, enhance real-time visualization, and develop interactive systems for engaging with AI-generated music visuals.**

Keywords — Generative AI, Multimodal AI, AI Music Visualization, Emotional Representation, Image Generation from Text

I. INTRODUCTION

The convergence of artificial intelligence (AI) and creative expression has introduced new possibilities for cross-modal translation of sensory experiences, enabling innovative ways to interpret and represent artistic and perceptual phenomena. Music, as a deeply expressive and emotionally resonant art form, has been extensively studied for its influence on human perception, cognition, and affective states. Historically, researchers have explored the relationship between auditory stimuli and their psychological and physiological effects, demonstrating that music can evoke strong emotional responses and shape individual experiences. With recent advancements in multimodal AI, an emerging area of interest involves the automated transformation of auditory information—including rhythm, melody, harmony, and lyrical content—into meaningful visual representations. Such capabilities not only extend the possibilities of artistic expression but also hold significant implications for accessibility, sensory augmentation, and interdisciplinary research in human-computer interaction (HCI) and computational creativity.

This study investigates the potential of AI-driven methodologies to generate visual representations of music by leveraging state-of-the-art generative models. Specifically, it seeks to address two fundamental research questions:

- **RQ1:** To what extent can AI effectively translate core auditory features—such as rhythm, melody and harmony—into visual representations that encapsulate the expressive and structural characteristics of music?
- **RQ2:** How do these multimodal translations influence users' auditory experience, particularly in terms of emotional engagement, cognitive interpretation, and accessibility?

By examining these questions, this research contributes to ongoing discussions on AI's role in creative synthesis, as well as its potential applications in music visualization, digital art, and assistive technologies for individuals with hearing impairments.

To systematically explore these objectives, we designed an AI-driven framework that processes key audio features—including tempo, zero-crossing rate, and Mel-Frequency Cepstral Coefficients (MFCCs)—to extract mood-related descriptors via a large language model (LLM). These descriptors were subsequently used to generate corresponding visual representations through a diffusion-based AI image generator. The methodological approach encompassed both fully automated and semi-automated techniques, incorporating user feedback in the latter to iteratively refine the visual outputs. By integrating computational models of music perception with generative AI techniques, this study aims to bridge the gap between auditory and visual modalities, contributing to the growing body of research on multimodal machine learning and AI-assisted creativity.

To evaluate the effectiveness and subjective perception of AI-generated music visuals, a user study was conducted, wherein participants were exposed to a range of visual outputs corresponding to diverse musical genres and styles. These visualizations encompassed abstract art, structured album-cover aesthetics, and conceptual landscapes, allowing for an analysis of user preferences and interpretative alignment with the corresponding music. The findings revealed notable variations in preference across different participant demographics, with structured visuals (e.g., album cover-style representations) being favored by casual listeners, while abstract and conceptual designs were more frequently appreciated by individuals with artistic or design-related backgrounds.

II. BACKGROUND

Recent advancements in generative AI have enabled the translation of auditory experiences into visual representations, forming the foundation for multimodal AI research in music perception and accessibility. Several studies have explored different approaches to

this translation, leveraging deep learning models, music emotion recognition (MER), and human-centered design.

A. Music Perception and Multisensory Experience

Music perception is inherently multisensory, engaging auditory, cognitive, and emotional faculties. Research in cognitive neuroscience [1] reveals that music activates brain regions related to emotion, motor control, and even visual imagery. This multisensory interaction is notably observed in synesthetic experiences, where individuals may perceive music as colors or shapes. Additionally, studies in music cognition emphasize the significant role of rhythm, melody, and harmony in triggering both emotional and physiological responses, suggesting that these musical components may be mapped onto other sensory experiences.

Research on synesthesia [2], where individuals experience cross-modal sensory perceptions (e.g., seeing colors when hearing music), further highlights the inherent link between auditory and visual processing. This suggests that translating music into visual forms could tap into pre-existing cognitive mechanisms and create a richer, more holistic musical experience.

B. Multimodal Generative AI

Over the past few years, significant strides in multimodal AI have greatly enhanced machines' capability to process and integrate information from diverse modalities, including text, images, and audio. Two of the most advanced multimodal AI models are ChatGPT and Google Gemini, which exemplify the evolution of AI from unimodal to fully multimodal systems.

Previous research finds that the performance of Gemini Pro outperforms inference-optimized models such as GPT-3.5 [3] and performs comparably with several of the most capable models available. Among the most advanced multimodal AI models, Google Gemini takes multimodal AI a step further by seamlessly integrating text, audio, code, image, and video processing [4]. This comprehensive approach enables it to function as a highly versatile AI system capable of engaging with complex. For the specific research task of generating mood or emotion descriptions from music genres, Gemini Pro was selected due to its multimodal capabilities rather than its overall performance superiority.

Multimodal AI spans a range of applications, including image captioning, speech recognition, video summarization, text-to-image generation, sound-to-text transcription, and visual search [5]. Advances in this domain have far-reaching implications for music analysis, accessibility, and even the development of novel forms of musical expression. For accessibility in music, multimodal AI is transforming the landscape by creating more inclusive experiences. For instance, it enables the development of haptic interfaces that allow d/Deaf or Hard of Hearing (DHH) individuals to experience music through tactile sensations [6].

Models like stable diffusion have demonstrated remarkable capabilities in learning relationships between various modalities. It enables tasks such as image captioning, text-to-image synthesis, and audio-visual retrieval [7]. These models leverage large datasets to learn joint representations that capture the underlying semantic connections between different data types. The success of these models provides a strong foundation for exploring the translation of auditory musical information into the visual domain. Furthermore, research in multimodal learning has explored various architectures and training strategies for effectively combining information from different sensory inputs, offering potential avenues for optimizing the mapping of musical features to visual representations [8].

C. AI in Emotional Perception

The ability of AI to recognize and interpret emotions has seen significant progress. Emotion recognition in AI has been extensively studied in speech processing, where models analyze vocal tone, pitch, and rhythm to determine emotional states [9]. Similarly, in natural language processing (NLP), sentiment analysis techniques extract affective information from textual content, enabling AI to assess emotions based on linguistic patterns [10]. Additionally, computer vision approaches can detect emotional expressions in images and videos, further broadening AI's capacity for emotion recognition [11].

In the context of music, AI models can analyze acoustic features to identify and classify the emotional content of musical pieces [12]. This capability is crucial for the proposed research, as it allows for the translation of the emotional essence of music into corresponding visual representations. By leveraging AI's ability to understand musical emotions, the generated visuals can effectively capture and communicate the affective dimension of the music.

D. AI-Generated Art and Music Visualization

AI-powered tools present artists with innovative and efficient means of creating visually compelling representations of music. The advancement of machine learning algorithms and neural networks has enabled the generation of artwork that is not only aesthetically striking but also deeply connected to the thematic, emotional, and conceptual dimensions of music. Through the use of AI-driven platforms, artists can produce high-quality visuals with enhanced precision and reduced manual labor, fostering a more dynamic approach to artistic creation. These tools allow for the exploration of novel artistic styles and the generation of visuals that might be challenging to achieve through conventional methods, thereby expanding the boundaries of creative expression[13][14].

The integration of visual art with music has long been an essential aspect of music releases, functioning as a visual manifestation of the artist's creative vision[15]. Historically, album covers, music videos, and promotional visuals have served not only as identifiers of musical works but also as conduits for conveying deeper meanings and enhancing the audience's connection to the music. The artwork associated with a music release often encapsulates the emotional core of the composition, offering a visual counterpart to the auditory experience. In this context, AI integration represents a significant shift, offering artists the potential to redefine how visual art interacts with music, while simultaneously maintaining the critical role that artwork plays in the broader music industry.

The adoption of AI technologies in the artistic process allows for the exploration of new dimensions of artistic expression and operational efficiency. Rather than functioning as a mere replacement for human creativity, AI serves as an augmentative tool that facilitates the expansion of artistic possibilities. Technologies such as Generative Adversarial Networks (GANs)[13] and deep neural networks can create complex and intricate visual outputs inspired by diverse artistic movements and styles[16]. By enabling artists

to experiment with various creative approaches, AI contributes to the transformation of the creative process, allowing for a broader range of visual representations that align with the emotional and thematic depth of music.

E. Evaluation of AI-Generated Visual Representations

Evaluating the effectiveness of AI-generated visual representations is essential for assessing their alignment with human perception and preferences. Previous research has explored various aspects of generative AI in music visualization, focusing on user preference, photorealism, and perceptual quality assessments. These studies provide valuable insights into how AI-generated images are perceived, their realism, and their ability to enhance user engagement.

A key consideration in evaluating AI-generated visuals is whether users can distinguish and prefer images that accurately reflect music. One study investigated whether users could recognize images that effectively represent music and whether they preferred AI-generated visuals over manually selected ones[12]. The findings suggest that users tended to favor AI-generated images, indicating that the system successfully captured the emotional and aesthetic essence of the music. The study also revealed that certain image styles influenced user preferences, suggesting that the artistic characteristics of AI-generated visuals play a role in shaping perception. Additionally, while AI-generated images were generally well-received, some biases related to specific styles were observed, highlighting the importance of refining generation methods to better align with diverse user expectations.

Another study focused on the realism of AI-generated images by assessing human perceptions of their authenticity, texture, and lighting compared to real-world images[17]. Participants evaluated images from various AI models alongside actual photographs, considering whether the AI-generated visuals could plausibly exist in reality. The results suggest that while AI models can produce highly detailed and visually convincing images, differences in lighting, texture consistency, and natural imperfections still distinguish them from real photographs.

Researchers also have developed AGIQA-3K[18], a large-scale dataset designed to evaluate the perceptual quality and text-to-image alignment of AI-generated images. This study examined the quality variance among AI-generated images by assessing their alignment with human subjective ratings. Researchers built a comprehensive dataset of AI-generated images and systematically analyzed how well existing AI evaluation models correspond to human perceptions of image quality and meaning. The study highlights that while AI-generated images can achieve impressive visual fidelity, inconsistencies in how models interpret text prompts can lead to variations in image accuracy.

III. METHOD

A. System Design

To address the research questions, we designed a system that translates snippets of audio into images representing the mood of a song. Two approaches to image generation have been considered: automatic and semi-automatic.

3.1.1 Automatic approach: The automatic generation of images relies entirely on the system's pre-programmed flow, with the only user input being the selection of the song. In this approach, the architecture first asks the user to upload a song and subsequently processes the audio to extract relevant features. The selected audio features are: tempo, zero-crossing rate, and Mel-Frequency Cepstral Coefficient (MFCC) values, which are explained in detail in the following subsection. Once the audio features are extracted, the song is represented as an array of numerical values, enabling its interpretation by a large language model (LLM). Previous research has demonstrated that contemporary LLMs possess the capability to interpret music in ABC notation and that models such as GPT-4 can accurately extract features from text-based music [14]. Similarly, we anticipate that current LLMs can interpret numerical audio features with sufficient accuracy to generate descriptions of a song's mood.

To this end, we selected Gemini Pro and conducted prompt engineering to formulate a prompt that queries a description of the emotion represented in the extracted audio features. The constructed prompt was as follows:

"What emotion can you infer from this musical piece with the following audio features? Tempo: tempo BPM. Sum of Zero Crossing rate: zcr. MFCC values: mfcc values. Provide the answer in four nouns or adjectives."

Once the prompt is sent to Gemini Pro, the response is received as a concise description of the song's mood. This description is then processed and transformed into a prompt for Stable Diffusion, which has also demonstrated effectiveness in image generation within musical contexts, as indicated by Yang (2024). The Stable Diffusion prompt includes an additional variable: the style of the generated image. Previous studies have shown that abstract styles are particularly effective for musical image generation [12]. Therefore, one of the initial styles tested was Abstract. The corresponding Stable Diffusion prompt for this style was:

"Abstract geometric art with the mood response Gemini"

However, Stable Diffusion supports a diverse range of artistic styles, which can be leveraged to enhance the generation of images based on the emotion elicited by the song. Examples of such styles include Imaginary Places, Clouds, and Album Covers, which are evaluated through user testing in Section 4. The Figure 1 provides a visual summary of the automatic image generation approach.

3.1.2 Semiautomatic approach: The automatic approach can be refined by incorporating human input at a critical point of the process, specifically, before generating the Stable Diffusion prompt. We refer to this as the semi-automatic approach, as the system does not generate images fully autonomously but instead requires user input to refine its results.

This approach enters a refinement loop immediately after obtaining the first response from Gemini, as depicted in Figure 2. In this loop, the user is asked to decide whether the given description of the song, which consists of four nouns or adjectives, accurately reflects the mood elicited by the audio. If the user responds affirmatively, the system proceeds in exactly the same way as the automatic approach. If the user responds negatively, they are prompted to suggest a modification to the description, and a new response is generated. This loop repeats until the user agrees with the description. As a result, a refined prompt is produced and sent to Stable Diffusion for image generation. The flow of the system is illustrated in the following example:

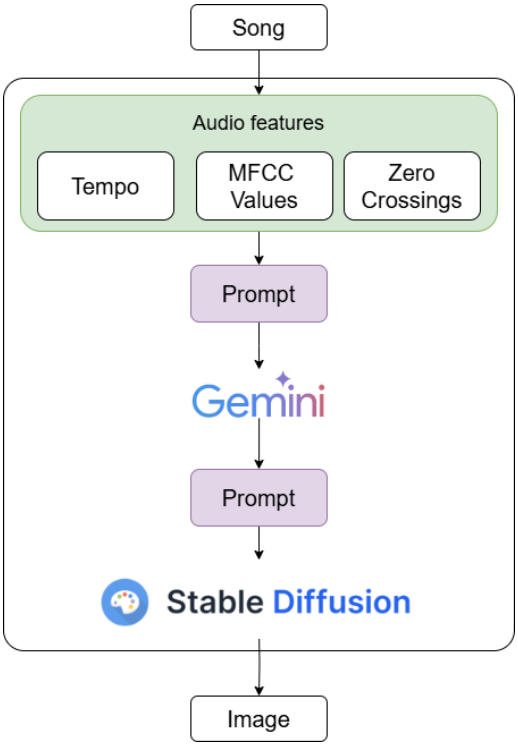


Figure 1: Translation of song to image with the automatic approach.

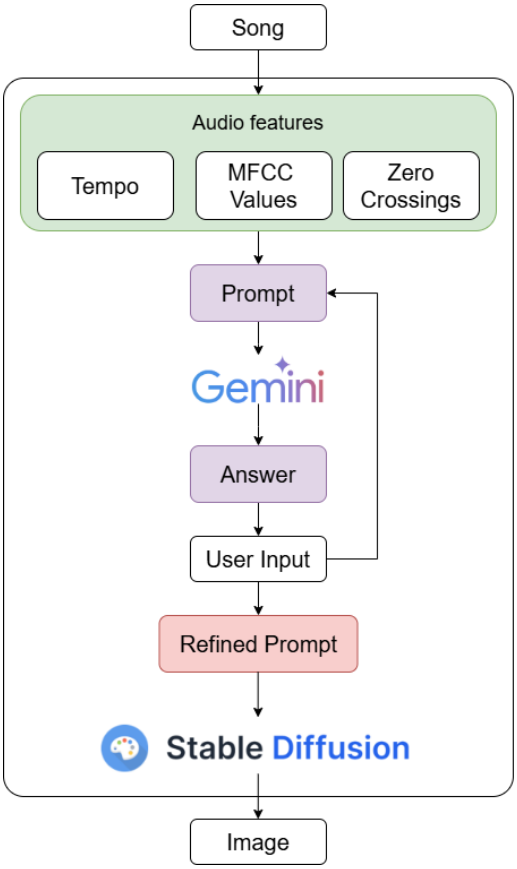


Figure 2: Translation of song to image with the semiautomatic approach.

(1) The user uploads a classical song.
(2) Gemini's response is: "Energetic, Tense, Bright, Driving."
(3) The user requests: "More vibes of waltz dancing."
(4) Gemini's revised response is: "Elegant, Romantic, Nostalgic, Graceful."
(5) The user agrees.

B. *Models and Features Selection*

During the system development, three key choices were made: the selection of audio features to describe the song, the choice of the LLM to generate the text description, and the selection of the image generation model.

3.2.1 Audio Features: The selected audio features to describe the songs were tempo, MFCC values, and the zero-crossing rate. Tempo refers to the beats per minute (BPM) of a song and how frequently these beats occur. Songs with a higher tempo can convey a sense of excitement or happiness, while songs with a lower tempo are often associated with sadness or melancholy. However, previous research has demonstrated that happy/sad emotions cannot be explained exclusively by tempo [19]; therefore, additional audio features are required. MFCCs (Mel-frequency cepstral coefficients) are a set of numerical values that represent the spectrum of a sound. For song description, 13 MFCC values were selected. The third selected audio feature was the zero-crossing rate, which indicates how frequently the signal changes from positive to negative or vice versa. A frequent zero-crossing rate can indicate an upbeat or energetic song, while a lower rate is typically associated with calm and relaxed melodies.

3.2.2 Models: The selected models were Gemini Pro and Stable Diffusion, chosen due to their success in previous research studies [14] and their publicly available and free-to-use API.

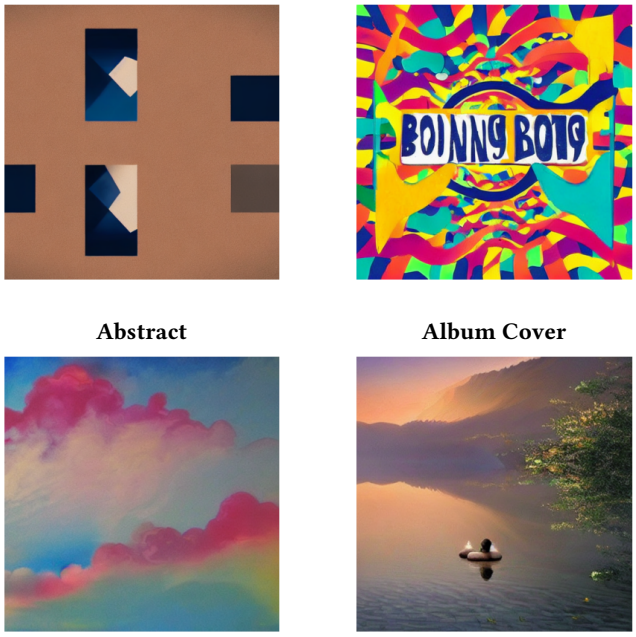
C. *Implementation Process*

The system was created and tested fully in a Google Collab environment with a RAM of 12.7 GB and a GPU of 15 GB in Python 3. For the audio feature extraction Librosa was employed, and for the testing snippets of songs of less than 20 s were used. The songs were selected from the DEAM dataset [20], which contains 1802 pieces of songs with annotations of their valence and arousal. The valence and arousal of this dataset were not used because the model is required to work with any song and the audio features are obtained in real time instead.

IV. USER TESTING

A. *Introduction*

To assess the effectiveness of the visualizations generated by our AI framework, we conducted a user study to examine the emotional



Abstract **Album Cover**

Diffuse Clouds **Imaginary Place**

Figure 3: Four distinct visual styles generated by Stable Diffusion

involvement and cognitive interpretation of participants. The experiment aimed to assess four distinct visual styles: Abstract Style, Imaginary Place Style, Diffuse Clouds Style, and Album Cover Style. Each style was designed to capture different aesthetic and conceptual interpretations of music. The primary objectives of the study were to identify which visual style best aligns with user perception, analyze variations in style preferences across different user groups, and explore potential alternative styles beyond the predefined categories.

B. *Participants*

A total of 23 participants were recruited for this study, representing diverse backgrounds in music and visual arts. The cohort included casual music listeners, avid music listeners, musicians, and visual artists or designers. Participants self-reported their level of engagement with music and visual arts, which allowed the segmentation of responses for comparative analysis.

C. *Procedure*

The study was conducted as an unmoderated test using Google Forms. Each participant was presented with a sequence of tasks to evaluate AI-generated visual representations of music. Participants listened to a music clip provided via a Google Drive link and subsequently viewed four AI-generated images corresponding to the song. They were asked to select the image that they felt best represented the essence of the song. In addition, participants were given the opportunity to justify their selection through an optional written response. This process was repeated for ten different songs to ensure consistency in style preferences. Finally, participants responded to a concluding question about the suitability of the styles provided and potential suggestions for alternative styles.

V. RESULTS

A. *Preliminary Results*

Before the user test was performed, the system was tested with two songs from the DEAM dataset, corresponding to the Blues and Pop genres.

5.1.1 Audio Features: The tempo of the Blues song was 89 BPM, while the tempo of the Pop song was 172 BPM. Figure 4 shows a graph of the mean MFCC values across both of the songs. The sum of the zero-crossing rate for the Blues song was 19,900 crossings, while for the Pop song, it was 43,200 crossings.

5.1.2 Text Description: For the automatic approach, based on the extracted audio features, the song description for the Blues song was *"Calm, Relaxed, Neutral, Subdued"*, while for the Pop song, it was *"Energetic, Excited, Upbeat, Driving"*. In contrast, for the semiautomatic approach, the text description of the song depends on the user's input.

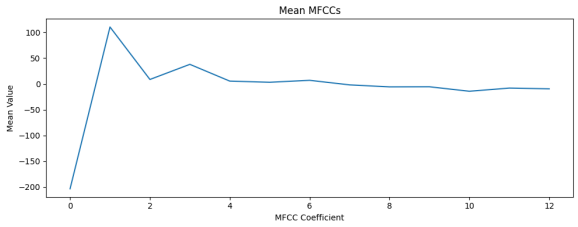
5.1.3 Image Generation: For the automatic approach, examples of the generated images for the Blues and Pop songs in abstract style are shown in Figure 5.

For the semiautomatic approach, a different genre was selected: Classical. This genre was found to struggle in maintaining a consistent text description, and as a result, the provided images exhibited the most variability compared to other genres. The text description of the Classical song was often closer to that of a Pop song, making it highly benefited from the human refinement involved in the semiautomatic approach. Figure 6 represents the comparison of image generation with and without human refinement for the Classical song in abstract style.

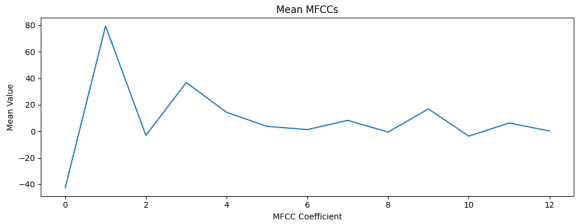
B. *User Testing Results*

5.2.1 Dominant Style Preferences per Song Responses were compiled and analyzed using Google Sheets, employing pivot tables to determine the style most frequently selected for each song. The results indicate that the Album Cover style was the most frequently chosen overall, particularly in Rap, Pop, Blues, and Etno Spanish, suggesting a strong association between structured, thematic visuals, and specific music genres. The Imaginary Place style was preferred for Jazz, Pop and Etno Senegal, implying that these genres evoke more conceptual and immersive imagery. Abstract style was dominant only in classical music, indicating that listeners of this genre may perceive music more fluidly and conceptually. Notably, no genre had Diffuse Clouds as the most preferred style, suggesting that this style did not strongly resonate with participants. See **Figure 7** for a visual representation of the results.

5.2.2 Style Preferences Across User Groups A comparative analysis was conducted to examine differences in style preferences across various user groups, including casual listeners, avid listeners, musicians, and visual artists. The findings suggest that Album Cover was the predominant choice for casual listeners and musicians, while visual artists and art hobbyists exhibited a stronger preference for

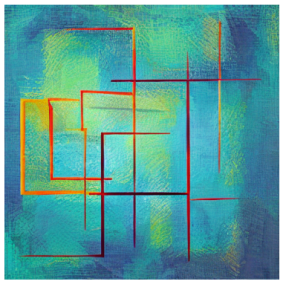


Blues song



Pop song

Figure 4: Mean of the MFCC values across the song



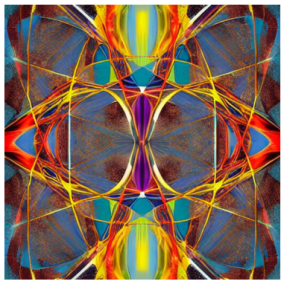
Blues song **Pop song**

Figure 5: Generated images in abstract style.

Abstract and Imaginary Place styles. Casual listeners tended to favor structured imagery, whereas artistically inclined participants were more receptive to conceptual visuals. Avid listeners showed a more mixed preference, with some leaning towards more abstract and conceptual styles. See **Figure 8** for a breakdown of user group preferences.

5.2.3 Evaluation of Alternative Style Preferences Participants provided open-ended feedback regarding potential alternative styles that could enhance the AI-generated representations of music. Several key themes emerged from the responses:

(1) Preference for Conceptual and Suggestive Imagery: Many participants expressed a desire for visuals featuring people, landscapes, or recognizable elements, as opposed to purely abstract visuals.
(2) Interest in Thematic and Symbolic Representations**:** Some suggested metaphor-based styles, such as food-related themes (e.g., coffee vs. cocktails), geometric patterns, or Bouba-Kiki-inspired visuals.
(3) Desire for Digital and Retro-Futuristic Aesthetics**:** Recommendations included incorporating Vaporwave, Cyberpunk, and 80s-90s aesthetics, especially for electronic or digital genres.
(4) Exploration of Classical Art Styles: Participants proposed integrating Impressionism, Baroque, and Chiaroscuro-inspired visuals.
(5) Abstract vs. Structured Debate**:** While some found abstract imagery ineffective in conveying musical emotions, others advocated for further refinement of abstract visuals rather than introducing entirely new styles.



Automatic **Refined**

Figure 6: Generated images without and with human refinement of the text description for the Classical genre in abstract style.

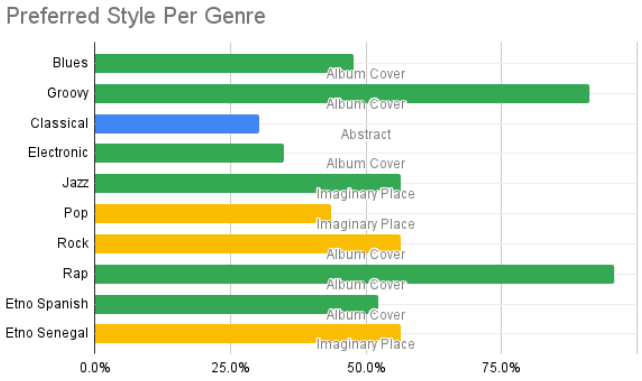


Figure 7: Preferred Style Per Genre

VI. DISCUSSION

A. System Performance

6.1.1 Audio Features: The preliminary results for the audio features of the Blues song and the Pop song revealed measurable differences between the two genres. The tempo and the sum of the zero-crossing rate were higher for the Pop song than for the Blues

Song	Casual Listeners	Avid Listeners	Visual Artists	No Background in Art	Art Hobbyist and Students
Blues	Album Cover	Album Cover	Album Cover	Album Cover	Diffuse Clouds
Groovy	Imaginary Place	Album Cover	Album Cover	Album Cover	Album Cover
Classical	Imaginary Place	Abstract	Imaginary Place	N/A	Abstract
Electronic	N/A	Album Cover	N/A	Abstract	Diffuse Clouds
Jazz	Imaginary Place	Imaginary Place	Abstract	Imaginary Place	Imaginary Place
Pop	Album Cover	Imaginary Place	Album Cover	Album Cover	Imaginary Place
Rock	Diffuse Clouds	Album Cover	Abstract	Album Cover	Album Cover
Rap	Imaginary Place	Album Cover	Album Cover	Album Cover	Album Cover
Etno Spanish	Album Cover	Album Cover	Album Cover	Album Cover	Album Cover
Etno Senegal	Imaginary Place	Imaginary Place	Imaginary Place	Imaginary Place	Diffuse Clouds

Figure 8: Image Style Preferences Across User Groups

song, enabling the LLM to detect a difference in mood. Furthermore, the graph of the mean MFCC values (Figure 4) for the Pop song shows more dynamic variation between coefficients, whereas the Blues song presents a smoother envelope. These subtle differences allow the LLM to capture the rhythm, variations, and overall emotion of the song.

6.1.2 Text to Image Generation: For the automatic approach, the four words suggested by the model accurately reflected the emotion conveyed by the song, except in one case: the Classical song. In this instance, the model failed to capture the calm essence of the song. This may be attributed to two main factors. The first possible explanation is the selection of the song itself; as only one song was tested for each of the selected genres. It is possible that this particular instance was more challenging for the system to describe. The second possible explanation is that the song was a waltz with a highly marked tempo and rhythm, and thus its audio features may be similar those of the Pop genre, which lead to an inaccurate description. However, this issue could be solved by incorporating the second approach, which includes human input. In this approach, the four initial words suggested by the LLM serve merely as a recommendation, allowing the user the freedom to further refine the description before the image is generated.

B. User Testing

The findings from the user study confirm that the Album Cover style remains the most widely accepted across multiple user groups, while Imaginary Place and Abstract styles appeal more to participants with artistic backgrounds. In addition, feedback indicated a strong interest in exploring alternative visual approaches, including figurative, symbolic, and digital aesthetics.

Future research should focus on refining AI-generated styles by incorporating more figurative and suggestive imagery. In addition, integrating new styles such as Vaporwave, Cyberpunk, and classical art influences could enhance engagement. Further A/B testing is recommended to compare structured versus abstract representations and their impact on user experience.

The insights gained from this study contribute to the advancement of AI-driven visualizations, fostering a deeper connection between music perception and digital art generation.

C. Strengths and Limitations

This study contributes to AI-generated music visualization by integrating multimodal AI models, specifically Gemini Pro for text generation and Stable Diffusion for image synthesis. The dual approach, combining automatic and semi-automatic methods, enhances the accuracy of visual outputs by allowing human refinement of AI-generated descriptions. Additionally, the selection of quantifiable audio features—tempo, MFCC values, and zero-crossing rate—ensures that the system captures the emotional essence of music. The study also highlights potential applications in music accessibility, particularly for deaf and hard of hearing individuals, and explores the intersection of AI and creative expression.

However, several limitations must be acknowledged. The study's small dataset of musical samples limits the generalizability of findings, and the lack of real-time visualization restricts its applicability in interactive or live music settings. Emotional perception remains subjective, posing challenges in evaluating the accuracy of AI-generated visuals. Additionally, the participant pool is relatively small, reducing the reliability of user preference insights. The study also primarily focuses on Gemini Pro and Stable Diffusion, without extensive comparisons to alternative generative AI models such as DALL·E or Midjourney.

D. Implications

This study advances research on AI-driven music visualization by addressing two key limitations in prior work: the constrained selection of musical genres and the lack of detailed classification of AI-generated visual styles. While previous studies have explored AI's ability to generate visual representations of music, they have often been limited in scope, either focusing on a narrow range of musical genres or employing broad, dichotomous classifications of visual output. By expanding both the diversity of musical genres analyzed and the range of visual styles considered, this study contributes to a more comprehensive understanding of how AI can effectively translate auditory experiences into meaningful and interpretable visual representations.

First, previous studies have primarily focused on a limited set of musical genres, such as Classical, Pop, and Jazz, when evaluating AI models for emotion recognition in music [14]. In contrast, this study incorporates a more extensive dataset comprising ten distinct genres, enabling a more comprehensive assessment of AI's ability to infer and represent musical emotions across diverse musical styles.

Second, while previous research has largely categorized AI-generated images into broad stylistic classifications such as abstract and realistic[12], this study identifies four distinct visual styles—Abstract, Imaginary Place, Diffuse Clouds, and Album Cover. This more granular classification provides deeper insights into the interplay between musical perception and visual representation, allowing for a more refined analysis of how different styles resonate with listeners. Additionally, the inclusion of multiple visual styles

mitigates potential biases in user preferences and highlights the influence of artistic conventions in AI-generated music visualizations.

By addressing these gaps, this research enhances the understanding of AI-generated music visualization in multimodal learning and accessibility. The findings suggest that AI-generated imagery can be optimized for specific genres and user expectations, contributing to the development of more adaptable and contextually relevant AI systems.

VII. CONCLUSION

A. Key Findings

This study examined the extent to which artificial intelligence can effectively translate core auditory features into visual representations and explored the influence of these multimodal translations on auditory experience. The findings provide direct answers to the research questions, offering insights into both the potential and limitations of AI-driven music visualization.

Regarding the AI's ability to translate auditory features into visual representations (**RQ1**), the study demonstrates that AI can effectively process musical attributes such as rhythm, melody, and harmony to generate meaningful visual outputs. By extracting tempo, zero-crossing rate, and Mel-Frequency Cepstral Coefficients (MFCCs), the AI framework successfully derived mood-based descriptors, which were then used to generate visual representations via a diffusion-based model.

Our study builds upon prior research [12] that suggested abstract imagery is particularly effective in translating musical elements into visual form. The user study indicates a more nuanced understanding of how different styles are perceived by listeners. Examining the influence of multimodal translations on auditory experience (**RQ2**), the study reveals that AI-generated visuals impact cognitive interpretation on the generated image that represents music.

The findings indicate that while abstract visuals were initially considered a strong candidate for conveying the mood and structure of a song, the album cover style emerged as the most preferred style overall. This preference was particularly evident among casual listeners and those with limited artistic backgrounds, suggesting that structured and genre-specific imagery may provide a more intuitive representation of music.

However, the abstract style remained highly favored by certain demographic subsets, particularly visual artists and those with a background in creative fields. Participants in this group often suggested refinements to the abstract style, highlighting the potential for further development in enhancing its ability to capture musical expression. This feedback underscores the need for iterative improvements to the generative process, ensuring that abstract imagery can more effectively communicate auditory themes without becoming overly ambiguous.

Overall, these results suggest that a hybrid approach—integrating the clarity of structured visuals with the expressive potential of abstract imagery—may provide a more comprehensive method for AI-generated music visualization. In addition, the use of a semi-automatic approach for image generation including human input demonstrated to be an effective strategy for more personalized image generation.

B. Future Work

This study makes significant strides in AI-driven music visualization by incorporating a wider variety of genres and refining visual style classification. However, certain limitations persist, necessitating further exploration. Future research should focus on overcoming these challenges by broadening dataset diversity, optimizing AI models, and refining evaluation techniques. A more extensive dataset encompassing diverse musical styles, emotional expressions, and cultural contexts would enhance the precision and applicability of AI-generated visuals.

A key area for future work is the inclusion of a more diverse participant pool in user studies. While this study gathered insights from individuals with varying levels of musical and artistic expertise, it did not specifically investigate how AI-generated visuals could benefit individuals with hearing impairments. Future research should explore the potential of AI-driven music visualization as an accessibility tool for d/Deaf and Hard of Hearing (DHH) individuals, who primarily rely on visual and vibrational cues to perceive music. Investigating how different visual styles enhance the musical experience for this group could provide valuable insights into improving accessibility through AI-generated imagery. Incorporating participants with diverse sensory experiences would contribute to the development of AI models that better align with different perceptual needs and expand the inclusivity of AI-driven multimodal systems.

Furthermore, to establish a more rigorous framework for evaluating AI-generated visuals, future studies should also incorporate standardized assessment methods from computational creativity and visual perception research. Combining objective evaluation metrics with subjective user feedback would improve the reliability, validity, and replicability of findings.

Finally, future research should conduct user testing on the semi-automatic approach to further explore the possibilities in image generation.

References

[1] Maïté Castro et al. "Personal familiarity of music and its cerebral effect on subsequent speech processing". en. In: *Scientific Reports* 10.1 (Sept. 2020), p. 14854. ISSN: 2045-2322. DOI: 10.1038/s41598-020-71855-5. URL: https://www.nature.com/articles/s41598-020-71855-5 (visited on 02/27/2025).

[2] Jamie Ward et al. "Synaesthesia is linked to differences in music preference and musical sophistication and a distinctive pattern of sound-color associations". en. In: *Psychology of Music* (June 2024), p. 03057356241250020. ISSN: 0305-7356, 1741-3087. DOI: 10.1177/03057356241250020. URL: https://journals.sagepub.com/doi/10.1177/03057356241250020 (visited on 02/27/2025).

[3] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. Version Number: 4. 2023. DOI: 10.48550/ARXIV.2312.11805. URL: https://arxiv.org/abs/2312.11805 (visited on 02/27/2025).

[4] Ankit Pande et al. "Comprehensive Study of Google Gemini and Text Generating Models: Understanding Capabilities and Performance". In: *Grenze International Journal of Engineering and Technology* June (2024).

[5] Summaira Jabeen et al. “A Review on Methods and Applications in Multimodal Deep Learning”. en. In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 19.2s (June 2023), pp. 1–41. ISSN: 1551-6857, 1551-6865. DOI: 10.1145/3545572. URL: https://dl.acm.org/doi/10.1145/3545572 (visited on 02/27/2025).

[6] Doga Cavdir. “Development of embodied listening studies with multimodal and wearable haptic interfaces for hearing accessibility in music”. In: *Frontiers in Computer Science* 5 (Jan. 2024), p. 1162758. ISSN: 2624-9898. DOI: 10.3389/fcomp.2023.1162758. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1162758/full (visited on 02/27/2025).

[7] Zineng Tang et al. *Any-to-Any Generation via Composable Diffusion*. arXiv:2305.11846 [cs]. May 2023. DOI: 10.48550/arXiv.2305.11846. URL: http://arxiv.org/abs/2305.11846 (visited on 02/27/2025).

[8] Maria Lymperaiou and Giorgos Stamou. “A survey on knowledge-enhanced multimodal learning”. en. In: *Artificial Intelligence Review* 57.10 (Sept. 2024), p. 284. ISSN: 1573-7462. DOI: 10.1007/s10462-024-10825-z. URL: https://link.springer.com/10.1007/s10462-024-10825-z (visited on 02/27/2025).

[9] Shashidhar G. Koolagudi and K. Sreenivasa Rao. “Emotion recognition from speech: a review”. en. In: *International Journal of Speech Technology* 15.2 (June 2012), pp. 99–117. ISSN: 1381-2416, 1572-8110. DOI: 10.1007/s10772-011-9125-1. URL: http://link.springer.com/10.1007/s10772-011-9125-1 (visited on 02/27/2025).

[10] Shangyue Lin. “Text emotional analysis in Natural Language Processing”. In: *Applied and Computational Engineering* 36.1 (Feb. 2024), pp. 163–172. ISSN: 2755-2721, 2755-273X. DOI: 10.54254/2755-2721/36/20230440. URL: https://www.ewadirect.com/proceedings/ace/article/view/10083 (visited on 02/27/2025).

[11] Hanna-Sophia Widhoelzl and Ece Takmaz. *Decoding Emotions in Abstract Art: Cognitive Plausibility of CLIP in Recognizing Color-Emotion Associations*. arXiv:2405.06319 [cs]. May 2024. DOI: 10.48550/arXiv.2405.06319. URL: http://arxiv.org/abs/2405.06319 (visited on 02/27/2025).

[12] Brian Man-Kit Ng et al. “Visualize Music Using Generative Arts”. In: *2024 IEEE Conference on Artificial Intelligence (CAI)*. Singapore, Singapore: IEEE, June 2024, pp. 1516–1521. ISBN: 979-8-3503-5409-6. DOI: 10.1109/CAI59869.2024.00273. URL: https://ieeexplore.ieee.org/document/10605337/ (visited on 02/27/2025).

[13] Vikalp Thapliyal and Pranita Thapliyal. “AI and Creativity: Exploring the Intersection of Machine Learning and Artistic Creation”. In: *International Journal for Research Publication and Seminar* 15.1 (Mar. 2024), pp. 36–41. ISSN: 2278-6848. DOI: 10.36676/jrps.v15.i1.06. URL: https://jrps.shodhsagar.com/index.php/j/article/view/329 (visited on 03/11/2025).

[14] Meng Yang, Maria Teresa Llano, and Jon McCormack. *Exploring Real-Time Music-to-Image Systems for Creative Inspiration in Music Creation*. arXiv:2407.05584 [cs]. July 2024. DOI: 10.48550/arXiv.2407.05584. URL: http://arxiv.org/abs/2407.05584 (visited on 02/27/2025).

[15] *AI in artwork creation for music: the future of visual expression*. en. Oct. 2023. URL: https://kwettr.com/blog/ai-in-artwork-creation-for-music-the-future-of-visual-expression/ (visited on 03/01/2025).

[16] Haziq Yusoff Bin Abdul Wahab and Alexei Sourin. “Application of Generative Adversarial Networks and Latent Space Exploration in Music Visualisation”. In: *2021 International Conference on Cyberworlds (CW)*. Caen, France: IEEE, Sept. 2021, pp. 125–128. ISBN: 978-1-6654-4065-3. DOI: 10.1109/CW52790.2021.00027. URL: https://ieeexplore.ieee.org/document/9599437/ (visited on 03/11/2025).

[17] Memoona Aziz et al. *Global-Local Image Perceptual Score (GLIPS): Evaluating Photorealistic Quality of AI-Generated Images*. arXiv:2405.09426 [cs]. May 2024. DOI: 10.48550/arXiv.2405.09426. URL: http://arxiv.org/abs/2405.09426 (visited on 03/11/2025).

[18] Chunyi Li et al. *AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment*. arXiv:2306.04717 [cs]. June 2023. DOI: 10.48550/arXiv.2306.04717. URL: http://arxiv.org/abs/2306.04717 (visited on 03/11/2025).

[19] Stéphanie Khalfa et al. “Role of tempo entrainment in psychophysiological differentiation of happy and sad music?” en. In: *International Journal of Psychophysiology* 68.1 (Apr. 2008), pp. 17–26. ISSN: 01678760. DOI: 10.1016/j.ijpsycho.2007.12.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167876007002504 (visited on 03/10/2025).

[20] Mohammad Soleymani, Anna Aljanaki, and Yi-Hsuan Yang. “DEAM: MediaEval Database for Emotional Analysis in Music”. In: (). URL: https://cvml.unige.ch/databases/DEAM/manual.pdf.