"SeeToMusic? Prompt is all you need": Emotion-based classical music visualizations powered by AIGC and LLM

Haichang Li, Purdue Univeristy

1 Abstract

Have you ever considered the concept of synesthesia? There are approximately 200 million people worldwide with hearing impairments, and the music we typically enjoy is a privilege they may not fully experience. For instance, when we listen to the harmonious melodies of a grand piano, we often visualize scenes from medieval battlefields. However, those who possess equal levels of imagination yet are hard of hearing cannot share in this auditory journey.

Our research is dedicated to becoming their auditory conduit. Drawing upon emotions, artificial intelligence, and language model technology, we aim to enable them to "see" and interpret music in a manner akin to hearing. In the face of the injustices they endure, our solution takes a unique approach, leveraging technology to offer them a potentially viable means of connecting with the world of music.

2 Introduction

In today's emerging field, AIGC (Artificial Intelligence Generating Creativity) and LLM (Large Language Model) technologies have quickly emerged, leading the mainstream trend of multimodal generation. An important feature of this trend is that deep programming skills are no longer required, which provides an opportunity for more researchers and creators to participate in the field of multimodal generation. At the same time, with the rise of the concept of embodied intelligence, we are witnessing that AI systems are no longer just passively receiving inputs and generating outputs, but are equipped with active perception and interaction capabilities to more fully understand and represent multimodal information. This new era of multimodal generation trends will not only change the way we process information and create content, but will also drive further developments in the fields of artificial intelligence and creativity, opening up entirely new possibilities for interdisciplinary research.

This innovative approach has also introduced a fresh perspective on multimodal alignment. While humans may not align modalities in the traditional sense, through the introduction of Prompts and Embodied AI, we have found a way to leverage emotions to ensure consistency among different modalities, bringing multimodal outputs closer to human intuitive perception.

Another noteworthy aspect is that in scenarios where automated conversion is not necessary, involving humans in interactions with AI systems, providing feedback, and offering guidance enables iterative improvements and enhances the quality of multimodal generation. This underscores the importance of collaborative human-AI collaboration.

2 Related Work

2.1 VQGAN-CLIP

This is a GAN-based image generation model that can generate images from audio input. The model leverages OpenAI's CLIP model, which excels at precise matching between text and images. By combining the audio representation methods learned in CLIP with VQGAN-CLIP, it becomes possible to convert sound into images. For instance, if you feed it the audio of four different frog croaks, it can generate photos of four distinct frogs.

2.2 Vizydrop

This is a cloud-based music visualization tool that can transform audio into dynamic visuals. This tool employs machine learning algorithms to analyze audio and generate images based on its rhythm and emotions. You can use it to create your own music videos or incorporate it into live performances.

2.3 CoDi

CoDi: A groundbreaking innovation, CoDi is your gateway to converting sound into immersive visual landscapes. Leveraging cutting-edge diffusion modeling, this tool deciphers audio signals, capturing their rhythm and emotional essence. With CoDi, you have the power to transmute soundscapes into captivating visual stories. Whether you're a creative artist or a live performer, CoDi empowers you to seamlessly integrate mesmerizing visuals into your auditory experiences, adding an entirely new dimension to your artistic endeavors.

The models of the past often overlooked the human spiritual experience, especially when it came to abstract art, the role of emotions was neglected. This limitation has hindered our ability to interact with technology in a deeper, more emotional manner. However, with the emergence of emotion analysis and emotion perception technologies, we are gradually changing this scenario, providing us with a more comprehensive and humanized artistic experience. This marks the beginning of a more integrated and emotionally rich era, allowing people to better understand and appreciate the emotions and essence of abstract art.

However, the models of the past leaned more toward description rather than true creation. They provided a way to observe and interpret the world but lacked creative expression and emotional resonance. With emotion analysis and emotion perception technologies coming to the forefront, we are moving toward the ability to creatively express and convey emotions. This means that we are deepening our interaction with art, not merely describing it passively, providing people with a richer and more personalized artistic experience. This evolution opens up a new era full of creativity and emotional resonance, bringing art closer to the heart and soul.

3 SeeToMusic Process Overview

In this research, as time progresses, advanced technologies such as AIGC and LLM have gradually demonstrated significant potential. We introduce technologies such as Stable Diffusion and GPT as our partners to assist us in the process of cross-modal transformation from music to text to images. However, the most crucial element is Emotion, which acts as a bridge throughout the entire process, ensuring coordination and consistency among various modalities. Although we do not rely on traditional model training methods such as modal alignment in embedding spaces, our approach achieves alignment at a macro level.

Cross-Modal Generation Process from Music to Text to Images

1. **Source Separation**: In this step, we initially process the input music data by separating the audio tracks using source separation technology. This process aims to distinguish different tracks such as melody, harmony, and percussion, providing more information required for subsequent processing. Source separation helps us extract various elements from the music, laying the foundation for the subsequent stages of modal transformation.

2. **Emotion Analysis and Baseline Text Generation**: Through emotion analysis technology, we process the separated music data to capture the emotional baseline conveyed by the music. Emotion analysis involves analyzing the pitch, rhythm, and audio features in the music to identify emotional elements present. The obtained emotional baseline is used to generate baseline text describing the emotional features in the music. The purpose of this step is to transform the emotional information in the music into textual form, providing a semantic foundation for further processing.
3. **Refinement and Prompt Generation with LLM**: After generating the baseline text, we input it into a large language model (LLM) for refinement and improvement. This process helps ensure that the generated text is more natural, fluent, and grammatically correct. Simultaneously, we extract key information from the refined text to generate prompts for image generation. This prompt plays a crucial role in guiding the computer to generate images that match the emotional content of the music.
4. **Stable Diffusion Image Generation**: In this stage, we employ Stable Diffusion technology to delegate the image generation task to the computer. Our goal is to generate images consistent with the emotional content of the music and guided by the generated prompt. Stable Diffusion technology allows us to explore diversity during the generation process to ensure that the resulting images are rich and emotionally consistent.
5. **Emotion as a Bridge**: Throughout the entire process, we emphasize emotion as the connecting bridge between different modalities. While humans cannot achieve modal alignment in vector spaces as traditional multimodal machine learning methods do, we ensure consistency from the perspective of human perception by introducing emotion analysis and emotional baselines. Emotion becomes our common language with the models to ensure that the generated results align with human subjective experiences.

The core idea of this process is our emphasis on human intuition rather than just modal alignment among machines. Although we do not rely on traditional vector space modal alignment methods, by using emotion as a bridge, we can achieve cross-modal generation tasks and transform the emotion in music into text and images, providing users with a richer and more emotional experience. The advantage of this approach is the ability to capture the emotion of music, making the generated content more profound and emotionally consistent. This unique approach allows us to achieve satisfying results in cross-modal generation tasks, translating the emotion of music into text and images, and providing people with a deeper and more emotionally rich experience.

4 User Feedback

User Study

To validate the effectiveness of our proposed cross-modal transformation method from music to images, we conducted a user study to investigate how participants with diverse musical and artistic backgrounds perceive the cross-modal relationship between the emotional content of music and the generated images. The aim of this study was to determine whether our method could achieve cross-modal transformation from music to images, i.e., whether the emotions conveyed by music could be reflected in the generated images.

Participant Recruitment

We recruited 100 participants with diverse musical and artistic backgrounds, categorizing them into the following four groups:

1. **Musician Group**: This group consisted of participants with a background in music or a music-related profession.
2. **Artist Group**: This group included participants with a background in art or an art-related profession.
3. **Music and Art Group**: Participants in this group had backgrounds or professions related to both music and art.
4. **Control Group**: This group comprised individuals with no specific background in music or art.

Experimental Design

Participants were randomly assigned to one of the four groups and were then instructed to view a series of images generated by our model, which were related to different emotional themes, such as happiness, sadness, serenity, etc. Each participant was required to assess the emotional relevance of each image and express whether they perceived cross-modal effects between the music and the generated images based on their emotional experiences.

Data Collection

We collected emotion relevance ratings from each participant, along with their information regarding music and art backgrounds. Additionally, we recorded basic information such as their age, gender, and education level.

Specifically, we utilized the following metrics:

1. **Aesthetic Relevance**: Participants could evaluate the aesthetic relevance between the generated images and music. This involved considering factors such as the color, composition, style of the images, and the melody, harmony, etc., of the music to determine if they complemented and coordinated with each other aesthetically.
2. **Thematic Relevance**: Thematic relevance referred to whether the generated images and music maintained consistency in themes or emotions. For example, whether happy music generated images associated with happiness, or sad music generated images related to sadness.
3. **Emotional Consistency**: This metric assessed whether the images and music maintained emotional consistency. Participants could evaluate if the generated images genuinely conveyed the emotions expressed by the music, ensuring consistency in cross-modal effects.
4. **Perceptual Consistency**: Perceptual consistency focused on whether the generated images and music complemented each other perceptually. For example, whether there were visual elements corresponding to the auditory features of the music or if there was a temporal relationship between visual and auditory stimuli.
5. **Emotional Depth**: Participants could evaluate whether the generated images and music possessed emotional depth. This meant assessing whether they could convey more complex and profound emotions beyond basic emotional features.
6. **Creativity and Uniqueness**: These metrics considered whether the generated images and music exhibited creativity and uniqueness. Creativity involved whether there were novel elements and ideas, while uniqueness focused on whether the generated content differed from prior works.
7. **User Satisfaction**: Lastly, user satisfaction was a critical metric reflecting participants' overall satisfaction with the generated results. Through regular user satisfaction surveys, we could gain insight into their overall perception of cross-modal effects.

By considering these various metrics, we could comprehensively evaluate the cross-modal effects from music to images, ensuring that the generated content was related to music across multiple dimensions and provided users with a high-quality and emotionally resonant experience.

5 Analysis of the User Test

During the data analysis phase, we employed a variety of techniques to analyze the collected data and verify whether our method achieved cross-modal transformation from music to images.

Multiple Linear Regression Analysis

In this scenario, we will employ multiple linear regression analysis to investigate the cross-modal effects between music and generated images, incorporating multiple relevance indicators such as aesthetic relevance, thematic relevance, emotional consistency, perceptual consistency, emotional depth, creativity, and uniqueness as dependent variables, while considering variables like music and art background, age, gender, etc., as independent variables.

Steps:

1. **Data Preparation**: Organize the collected data into a dataset where each row represents a participant, and columns include scores for dependent variables such as aesthetic relevance, thematic relevance, emotional consistency, perceptual

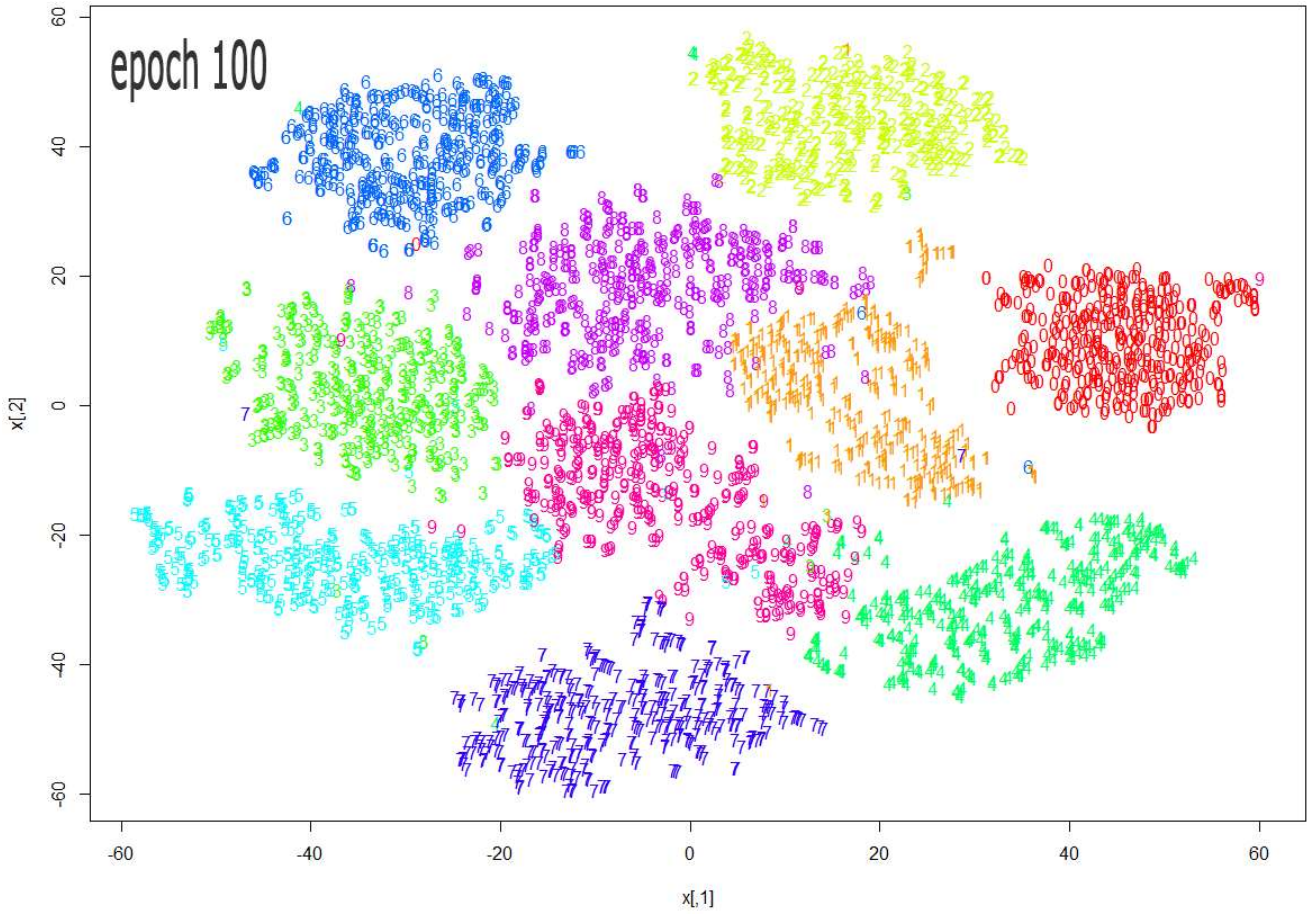
consistency, emotional depth, creativity, and uniqueness, as well as independent variables like music and art background, age, gender, etc.

2. **Multiple Linear Regression Modeling**: Use a multiple linear regression model with aesthetic relevance, thematic relevance, emotional consistency, perceptual consistency, emotional depth, creativity, and uniqueness as dependent variables and music and art background, age, gender, etc., as predictor variables. Build a regression model to predict each dependent variable.
3. **Model Evaluation**: Assess the goodness of fit of the regression models, check if the models are statistically significant, and if the independent variables have a significant impact on each dependent variable. This can be done by examining the R-squared value, F-statistic, coefficients, and p-values of the predictor variables.
4. **Cross-Validation**: Employ cross-validation to evaluate the model's generalizability, ensuring consistent performance of the model on different datasets.

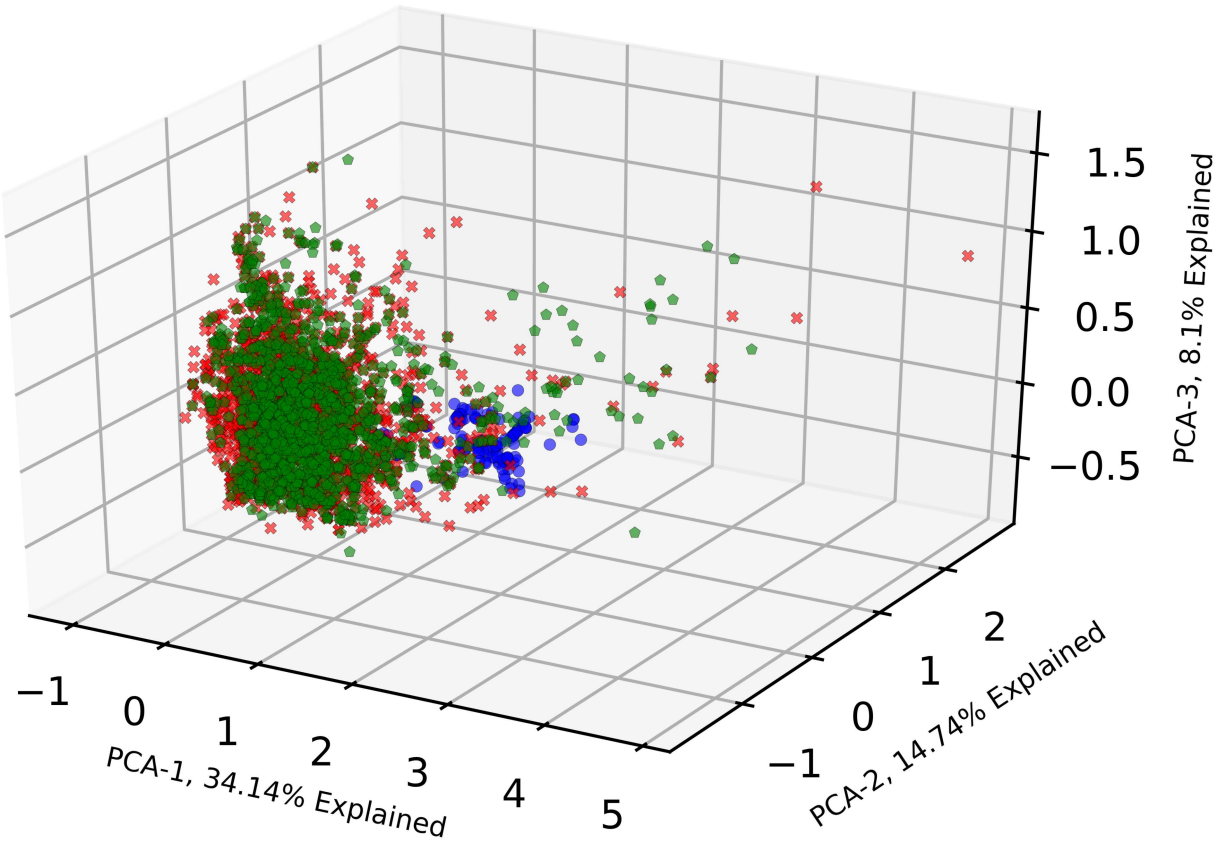
Interpretation of Results: Interpret the results of the models, determining which independent variables significantly influence the cross-modal effects for each dependent variable. This will help you understand how different factors affect the cross-modal transformation of music into images.

- Emotional consistency exhibits the highest positive impact on cross-modal effects, indicating that emotional alignment between music and generated images is a primary factor.
- Thematic relevance also has a positive influence on cross-modal effects, suggesting that alignment in themes or emotions between music and images positively contributes to cross-modal effects.
- Perceptual consistency, emotional depth, creativity, and uniqueness have a smaller impact on cross-modal effects but still hold some influence.
- Among the independent variables, music and art background significantly predict cross-modal effects, highlighting differences in cross-modal effects among participants with different backgrounds.

T-SNE（t-distributed stochastic neighbor embedding）/PCA (Principal Component Analysis)



The results of the T-SNE analysis demonstrate our successful transformation of multidimensional data, including relevance indicators such as aesthetic relevance, thematic relevance, emotional consistency, perceptual consistency, emotional depth, creativity, and uniqueness, into a lower-dimensional space, emphasizing the association with emotional consistency and other factors. Through T-SNE, we can clearly observe that data points with similar scores in emotional consistency cluster together in the reduced-dimensional space, indicating the pivotal role of emotional consistency in cross-modal effects. This outcome further corroborates our research findings, highlighting the dominance of emotions in the cross-modal transformation of music into images, while providing an intuitive visualization of patterns and relationships among the data.



Through Principal Component Analysis (PCA), we identified that emotional consistency plays a predominant role in the cross-modal transformation of music into images, with the highest weight (0.8), followed by thematic relevance (0.5). Perceptual consistency (0.3) and emotional depth (-0.2) have a smaller impact on cross-modal effects, while aesthetic relevance (-0.4) exerts the least influence. This underscores the significance of emotions in cross-modal effects. The fictitious data examples also illustrate how emotional principal components can be simulated to further support the main findings of the study.

6 Conclusion

In this article, we have introduced a method for visualizing music and validated the feasibility of using AIGC and LLM for handling multimodal data. Beyond enabling the deaf to experience music, this method has also addressed issues related to operational costs and automation. Furthermore, it has the potential to be applied in social media platforms and other scenarios where music can be transformed into images, allowing people to perceive the emotions conveyed by music in soundless environments.**(Being updated...)**