Emotion-Driven Audio-Visual Experience: Enhancing Human-Computer Interaction through Real-Time Multimodal Feedback

Aditya Vikram Singh
Northeastern University
Boston, United States
singh.adityav@northeastern.edu

Ankit Gundewar
Northeastern University
Boston, United States
gundewar.a@northeastern.edu

Shishir Kallapur
Northeastern University
Boston, United States
kallapur.shi@northeastern.edu

Yifan Tang
Northeastern University
Boston, United States
tang.yifa@northeastern.edu

ABSTRACT

This project explores an AI-driven system for generating immersive and personalized audio-visual experiences by detecting user emotions and intent in real time. Our approach focuses on generating personalized audio-visual experience and enhancing user engagement by integrating hand gestures. The system supports dynamic visual rendering through TouchDesigner and real-time music generation with MusicGen [1], tuned via a feedback loop interface. Early evaluation via user surveys suggests increased immersion and perceived emotional resonance. While still in a prototyping stage, this work paves the way for deeper exploration of affective computing in interactive media and potential use in therapeutic and entertainment contexts. **Link to Github Repository**

KEYWORDS

Affective Computing, Multimedia, Emotion Detection, Human-Computer Interaction, Generative AI

ACM Reference Format:
Aditya Vikram Singh, Ankit Gundewar, Shishir Kallapur, and Yifan Tang. 2025. Emotion-Driven Audio-Visual Experience: Enhancing Human-Computer Interaction through Real-Time Multimodal Feedback. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 8 pages.

1 INTRODUCTION

Current multimedia experiences are often static and unable to respond to users' real-time emotional states. Our project addresses this limitation by proposing an AI system that dynamically adjusts audio and visual outputs based on the user's mood and interactive cues, aiming to increase engagement and personalization.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1.1 Problem Statement

Traditional multimedia systems lack adaptiveness to real-time emotional states. In domains such as therapy, meditation, and digital entertainment, emotional awareness and personalization are increasingly critical. Despite advancements in generative AI, most systems still lack seamless real-time responsiveness to users' emotional states and intent. This gap hinders meaningful human-computer interaction and user satisfaction. Our goal is to address this through multimodal integration and affective AI.

Challenges include reliably interpreting multimodal emotional cues and integrating generative tools into a real-time system. The technical complexity of connecting diverse systems—emotion detection, gesture recognition, and generative models—creates significant integration challenges, particularly regarding latency management and coherent experience design.

1.2 Summary of Approach

The system is built to analyze real-time input from users via a chat UI and engage with the system for an enhanced experience via hand gestures. By taking the input from the user, our system generates personalized music using MusicGen [1] and visuals using StableDiffusion [14] in TouchDesigner. The UI built in Streamlit allows a feedback loop for tuning the emotional impact.

The novelty of our system lies in combining generative art and audio with real-time emotion and intent recognition, allowing users to feel in control of and connected to the experience. Unlike passive consumption, our system enables users to actively influence outputs based on their gestures and expressions, supporting a collaborative human-AI dynamic.

Key Components.

- Webcam-based video input with hand gestures.
- Real-time visual generation in TouchDesigner.
- Music generation using Meta's MusicGen
- Interactive control using Streamlit interface.

1.3 Hypotheses

Our project mainly focused on the below hypotheses:

- The personalized audiovisual experience generated by AI can enhance user engagement and emotional impact compared to static media.

We base it on the following sub-hypotheses:

- Emotion-driven generative content can lead to deeper user immersion.
- Interactive elements can increase perceived control and satisfaction.

1.4 Summary of Results

Initial user testing revealed a positive correlation between real-time adaptive content and perceived emotional engagement. Users reported higher immersion levels when the system responded to their emotional state, supporting our hypothesis that personalized AI-generated content increases emotional impact. The prototype demonstrates the potential of emotion-aware systems in HCI despite technical implementation challenges. Qualitative feedback indicates particularly strong resonance with the system's ability to match visual aesthetics to emotional states, though audio generation latency remains a limiting factor.

2 BACKGROUND / RELATED WORKS / JUSTIFICATION

- **Multimodal Emotion Processing**: Gerdes et al. demonstrated how emotional cues across different modalities enhance perception and user engagement [4]. Our work extends this by applying these principles to real-time interactive experiences. Their research established that combining visual and auditory emotional cues creates stronger affective responses, which our system leverages through synchronized multimedia generation.
- **Art and Music Emotion Linkage**: Hisariya et al. explored music generation from visual/emotional cues [5], which we build upon by adding user feedback loops and real-time adjustment. Their approach relied on static art while ours incorporates dynamic facial expressions. Their work established important mappings between visual emotional cues and musical parameters that we adapt for our real-time system.
- **Generative Multimedia Tools**: Liu et al. showed potential for aligning generative visuals to audio [12]. Our approach integrates similar techniques but adds emotion detection as an input vector, creating a more personalized experience. Their work demonstrates technical feasibility but lacks the interactive component our system introduces.
- **Affective Computing in HCI**: Building on Picard's foundational work in affective computing, recent research has demonstrated benefits of emotional awareness in interface design. Systems that adapt to user emotional states have shown higher engagement levels and user satisfaction in contexts ranging from education to entertainment. Our work contributes to this growing field by exploring real-time adaptation in creative multimedia experiences.
- **Therapeutic Applications**: Research in arts therapy has established connections between creative expression and emotional regulation. Digital tools offer new possibilities for therapeutic interventions, with early studies showing promise for guided multimedia experiences in managing anxiety and mood disorders. Our system could potentially serve as a technological platform for such applications.

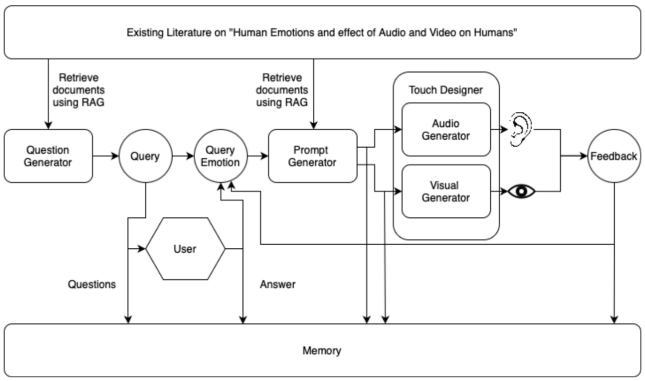

Figure 1: System Overview

Our system builds on these foundations by combining them in real-time user-guided experiences, addressing the gap in interactive emotion-driven multimedia systems.

3 MODEL, ARCHITECTURE AND IMPEMENTATION

3.1 System Overview

Fig 1 illustrates the high-level architecture of our system pipeline, which is composed of four core components:

3.1.1 ***Question Generator.*** The Question Generator module is implemented using a Retrieval-Augmented Generation (RAG) framework coupled with a Large Language Model (LLM). This module is responsible for two key functions:

- Formulating and presenting a set of three personalized questions to the user, strategically designed to elicit information about the user's current emotional state and underlying intent.
- Parsing the user's responses to these questions in order to infer a structured representation of their emotional state and intent using affective reasoning.

The LLM is augmented with access to a curated corpus of psychological and affective computing literature 8, enabling it to retrieve relevant theoretical context during both question formulation and emotional inference. By leveraging domain-specific knowledge through RAG, the module generates semantically rich and psychologically grounded queries and interpretations that reflect real-world affective models.

3.1.2 ***Prompt Generator.*** The Prompt Generator is another RAG-enhanced LLM module that consumes the structured outputs of the Question Generator—namely the inferred emotional state and user intent—to synthesize task-specific prompts for downstream generative models:

- An audio prompt for the audio generation model.

- A visual prompt for the visual (image/video) generation model.

This module accesses domain literature on the psychophysiological effects of media stimuli, including music and visuals, to guide the prompt synthesis process 8. The generated prompts contain detailed multimodal descriptors that are affectively aligned with the user's emotional context—either to reflect, modulate, or amplify their current state in accordance with the user's expressed intent. This grounding ensures that the generative outputs are both semantically coherent and emotionally resonant.

3.1.3 ***Audio Generator.*** The Audio Generator module is responsible for synthesizing audio outputs that are semantically and affectively aligned with the inferred emotional state and intent of the user. To achieve this, we utilize the MusicGen model introduced by Meta [1]. MusicGen is a single-stage autoregressive transformer architecture that directly maps textual descriptions to audio waveforms by generating melodic features conditioned on natural language prompts.

In our implementation, we deploy the lightweight variant of MusicGen due to computational constraints on the local inference environment. This model variant provides a favorable trade-off between generation quality and resource efficiency. The input to MusicGen consists of prompt text generated by the Prompt Generator module, which encodes affective cues and semantic descriptors aligned with the user's current emotional state and desired modulation (e.g., calming, energizing, reflective).

Due to the transformer's autoregressive nature and local compute limitations, the inference pipeline does not achieve real-time performance. The latency implications and possible optimization strategies are further discussed in Section 6.2.

3.1.4 ***TouchDesigner-Based Visual Generator.*** TouchDesigner [3] is a node-based visual programming platform designed for real-time, interactive multimedia development. It enables rapid prototyping and deployment of audio-visual experiences by chaining modular operators in a dataflow graph. Leveraging its high-performance rendering engine and extensible interface, we built a custom visual generation pipeline that supports multimodal interactivity and real-time feedback.

The custom TouchDesigner network powering our visual generator consists of the following components:

- **Inputs:** We integrated two types of visual inputs into the pipeline:
  - A real-time webcam feed to enable user interaction and emotion-sensitive modifications.
  - A video stream input (e.g., gameplay footage), allowing us to demonstrate emotion-conditioned transformations of static media, as discussed in Section 4.3.
- **Hand Gesture Controller:** We implemented real-time hand tracking to provide intuitive and fine-grained control over both audio and video outputs. Specifically, the distance between the index finger and thumb is continuously tracked:
  - It dynamically modulates the audio output volume.
  - It defines a region of interest for the visual output—highlighting specific areas in the video.

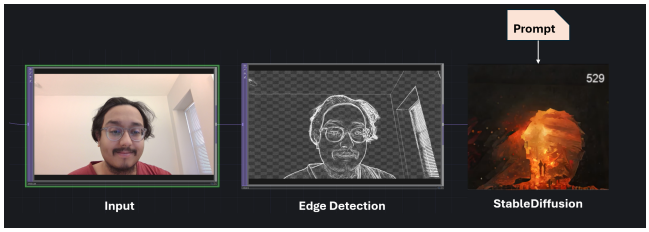

Figure 2: Stable Diffusion with ControlNet used within StreamDiffusion (placeholder image)

- **StreamDiffusion:** The core component of our visual generator is StreamDiffusion [9], a diffusion-based image generation pipeline optimized for real-time applications. Built on top of Stable Diffusion models [15], StreamDiffusion incorporates ControlNet to condition the generative process on structural inputs (e.g., edge maps). It transforms the incoming visual input (from the webcam or gameplay video) by:
  - Extracting edge features from the source image.
  - Feeding them into the diffusion model alongside prompt embeddings.
  - Denoising a latent noise tensor to generate semantically aligned visual outputs.

  This process is illustrated in Figure 2.
- **Audio-Driven Noise Injection:** A dedicated submodule extracts temporal features from the generated audio and injects them as noise signals into StreamDiffusion. This enables temporal synchronization between audio and visual dynamics, ensuring cohesive multimodal outputs.
- **Interaction Metric (Proof-of-Concept):** We implemented a quantitative interaction score based on real-time gesture analysis. The score increases with frequent and dynamic gestures (e.g., rapid finger motion or varied positioning), serving as a proxy for user engagement.
- **Final Integration:** The final system renders the audio-reactive visuals and generated audio in a dedicated output window, creating a unified and immersive user experience. 3 shows the complex network of modules separated by their individual functions

3.1.5 ***Feedback System.*** The complete pipeline depicted in Figure 1 is designed as an iterative, memory-augmented feedback loop. Following each generation cycle, the user is given the opportunity to provide explicit feedback on the generated media output.

To support iterative refinement, both the Question Generator and Prompt Generator modules incorporate a persistent memory mechanism that maintains contextual history across interaction rounds. This memory includes:

- Previous user responses and inferred emotional states.
- Prior generated prompts and corresponding media outputs.
- Explicit user feedback on those outputs.

The Question Generator utilizes this memory to update its affective reasoning process. When sufficient prior context is available, it avoids redundant re-questioning by re-analyzing emotional state

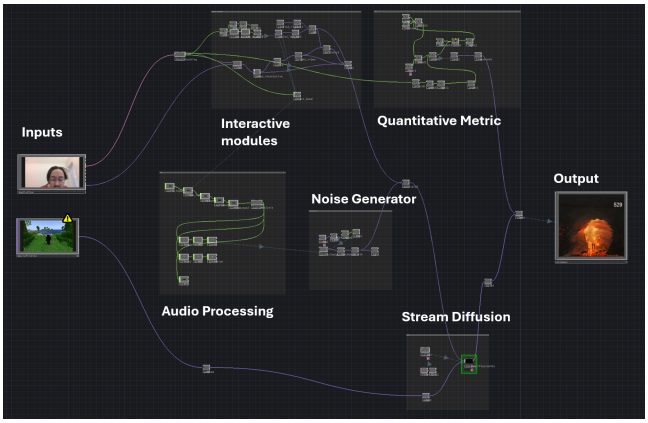

Figure 3: TouchDesigner Setup

and intent from historical interactions, thus streamlining the user experience.

Similarly, the Prompt Generator leverages historical prompt-response pairs and user feedback to adapt its next set of generated prompts. This allows the system to iteratively converge on outputs that are more aligned with the user's preferences and affective goals.

This memory-driven feedback loop enables fine-tuning of the audiovisual outputs without restarting the pipeline, making the system adaptive, personalized, and capable of evolving with the user over time.

3.2 System Integration

We developed a fully integrated end-to-end application by combining custom Python-based modules for emotional state analysis, prompt generation, and audio synthesis with a real-time multimedia pipeline built in TouchDesigner. The entire application is wrapped in an interactive front-end using Streamlit, providing a seamless user interface for multimodal generation.

Upon launch, the system executes the following sequence:

(1) The `QuestionGenerator` module is invoked to generate three affectively oriented questions for the user. These are designed to infer the user's current emotional state and intent.
(2) The user responds via the Streamlit interface. The responses are returned to the `QuestionGenerator`, which processes them and outputs a structured representation of the emotional state and intent.
(3) This affective representation is passed to the `PromptGenerator` which synthesizes:
   - An audio generation prompt.
   - A video generation prompt.

   The LLM is instructed to return its output in a structured JSON format to facilitate downstream parsing.
(4) The audio prompt is fed into a locally hosted instance of the MusicGen model, which generates the corresponding audio waveform. The result is saved as a `.wav` file in the project root directory.
(5) Simultaneously, the video prompt is stored in a plain text file in the same directory.
(6) The TouchDesigner environment is configured to continuously monitor and read these two files:
   - The audio file is used as an input to generate audio-reactive visual effects.
   - The video prompt text is passed to the StreamDiffusion module, which guides the transformation of incoming visual input (from either a webcam or pre-recorded gameplay footage) based on the described emotional context.
(7) The final output is a synchronized, emotion-aligned audiovisual experience, dynamically constructed based on user input and continuously rendered in real time via TouchDesigner.

This tightly coupled system enables real-time interaction and personalization, making it adaptable for various use cases such as affective gaming, immersive storytelling, or emotionally intelligent multimedia applications.

4 EXPERIMENT / EVALUATION APPROACH

4.1 Experiment Design / Plan

Participants experienced the application in a controlled setting with consistent lighting and positioning. Each session lasted approximately 15 minutes, beginning with a brief explanation of the system followed by free interaction. Users were instructed to express different emotions and observe how the system responded. After the interaction period, participants filled out a Likert-scale based survey. We collected subjective responses on emotional alignment, enjoyment, and perceived control.

4.2 Demos

4.2.1 Demo 1. Initially, the system generates and asks three questions to the user:

- *"Could you share how you're feeling right now and what has been influencing those emotions?"*
- *"What's been on your mind lately that might be affecting your emotional state?"*
- *"Are there any recent events that are bringing you hope or is causing concern?"*

The user then responds with: *"I am feeling very anxious and stressed. I have been applying to a lot of jobs but none of them are sending back responses. I am scared about my future. I feel like crying out "*. Figure 4 corresponds to the output for this input. The audio generated encapsulated the anxiousness and fear. The visuals depict lots of swirls and denoting hopelessness.

4.2.2 Demo 2. This demo includes the feedback loop. Initially, the system generates and asks three questions to the user. This time the response was: *"I am so damn fed up right now. I am frustrated and angry and nothing is going my way. Everything seems irritable and I feel like screaming out loud and punch a hole in the wall."* Figure 5 corresponds to the initial output for this input. The audio generated encapsulated the some major tones, depicting a positive aligned music. The visuals had jagged aggressive lines but the colors were a mix of warm and cool colours. The user then enters the feedback *"This is not helping, i want to vent out my anger right now as it is the only thing that I feel."*. The system takes this feedback and tunes

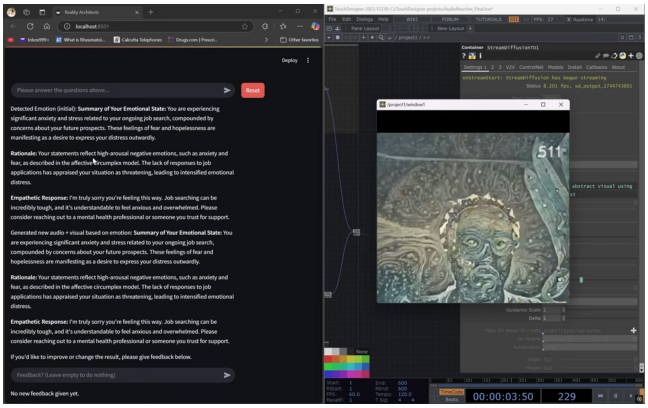

Figure 4: Output for an emotional state of fear and anxiousness

Figure 5: Initial Output for an emotional state of anger and frustration

the audio and visual to give us a better output as seen in Figure 6. The music was aggressive and inspired by heavy metal. The visuals included fire and bright red colors .

4.3 Potential Application: Gaming

One of the biggest potential applications of this system could be in creating personalized game levels based on user emotions. We tested a proof-of-concept with our system by replacing the webcam input by a gameplay video of *Minecraft*, as discussed in 3.1.4.

Figure 7 shows the prrof-of-concept application that we came up with. We can see how based on the emotion and the intent, the output was a tuned version of the base game. With better and powerful models and fine grain control of the game enviornment, we can scale this up to create more immersive experiences.

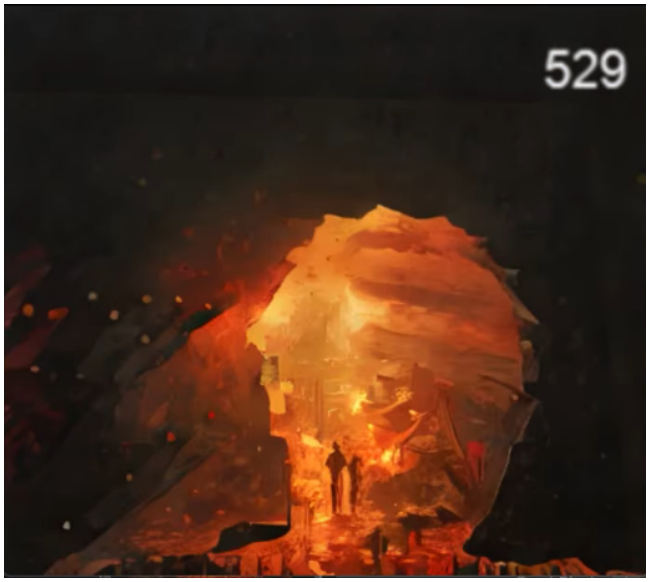

Figure 6: Tuned Output for an emotional state of anger and frustration

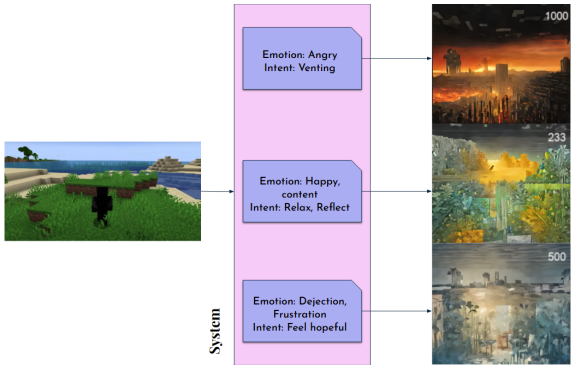

Figure 7: Potential application of our system in gaming

4.4 Participants

Our initial study used an informal sampling of 9 participants, including classmates and close friends. While not statistically representative, these early interactions offered valuable qualitative feedback. Future evaluations will adopt a more rigorous experimental protocol.

Ideally, we would expand to a larger, diverse population of 25-30 participants, primarily from Northeastern University, ensuring a balanced gender distribution and varied technical backgrounds. A formal study would include participants from varied age groups (18-55) and technical backgrounds to assess the system's accessibility as well.

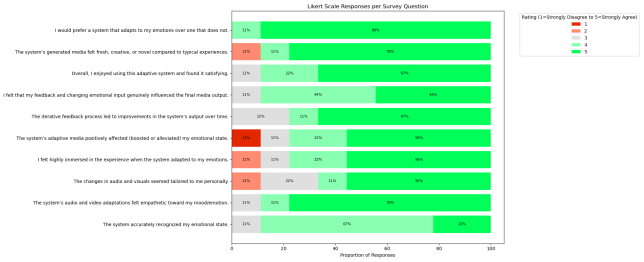

Figure 8: Survey results

4.5 Instruments Used

- 5-point Likert scale survey assessing emotional resonance, immersion level, system responsiveness, and overall satisfaction
- Interaction logging as engagement metric, tracking frequency of gestures and emotion changes

The questions on the survey were as follows:

- *The system accurately recognized my emotional state.*
- *The system's audio and video adaptations felt empathetic toward my mood/emotion.*
- *The changes in audio and visuals seemed tailored to me personally.*
- *I felt highly immersed in the experience when the system adapted to my emotions.*
- *The system's adaptive media positively affected (boosted or alleviated) my emotional state.*
- *The iterative feedback process led to improvements in the system's output over time.*
- *I felt that my feedback and changing emotional input genuinely influenced the final media output.*
- *Overall, I enjoyed using this adaptive system and found it satisfying.*
- *The system's generated media felt fresh, creative, or novel compared to typical experiences.*
- *I would prefer a system that adapts to my emotions over one that does not.*

4.6 Survey

The evaluation consisted of three phases:

- **Exploration Phase**: Participants freely interacted with the system without specific directions to establish baseline engagement patterns.
- **Directed Task Phase**: Users were instructed to express specific emotions and gestures to test system recognition accuracy.
- **Answering questionnaire**: After experienceing the enitre system, participants were asked to complete the likert scale based survey

5 RESULTS

We conducted a Likert-scale survey to evaluate the perceived effectiveness, adaptiveness, and emotional alignment of our affective multimedia system. 9 Participants rated their agreement (1 = Strongly Disagree, 5 = Strongly Agree) with a set of statements spanning usability, personalization, emotional resonance, and system responsiveness.

Key Findings:

- **User Preference and Satisfaction:** A majority (89%) of participants expressed a strong preference for systems that adapt to emotions over static systems. Furthermore, 67% of users reported overall satisfaction with the adaptive system.
- **Creativity and Novelty:** 78% of participants found the system-generated media to be fresh, creative, or novel compared to traditional experiences.
- **Emotional Adaptation and Immersion:** 56% of users agreed that the adaptive media positively affected their emotional state, and an equal proportion reported high immersion when the system responded to their emotions.
- **Feedback Loop Effectiveness:** 67% of users indicated that the iterative feedback process improved the output over time. 88% felt their feedback and emotional changes influenced the final media, validating the memory-augmented feedback loop.
- **Empathy and Personalization:** 78% agreed that the system's adaptations felt empathetic to their mood, and 67% perceived the changes in audio and visuals as personally tailored.
- **Emotion Recognition:** 89% of participants believed the system accurately recognized their emotional state.

Overall, the results suggest strong user endorsement of emotionally adaptive systems, supporting the effectiveness of our multimodal, feedback-driven pipeline.

6 ANALYSIS / DISCUSSION

6.1 Successes

Real-time interaction and feedback loop proved engaging, with participants reporting a sense of agency and connection with the generated content. The multimodal approach (combining visual and audio) created a more holistic emotional experience than either modality alone would provide. The system's ability to recognize and respond to emotional shifts created what one participant described as "a conversation with the technology," suggesting the emergence of a rudimentary emotional intelligence.

6.2 Challenges and limitations

Difficulty with APIs for music generation created latency issues that occasionally disrupted the immersive experience. Setting up TouchDesigner pipelines required significant technical expertise, limiting potential for widespread adoption without further simplification. We also observed notable individual differences in how emotions were expressed and interpreted, highlighting the need for calibration periods or more sophisticated personalization algorithms.

The current implementation struggles with rapidly changing emotional states, as the music generation component requires approximately 1-2 minutes to create new content constraint to hardware limits. This creates a noticeable lag between emotional expression and audio response. The visual components, while more

responsive, occasionally exhibited rendering artifacts during complex scenes.

6.3 Areas of Improvement

Integrating Emotion detection using webcam input under varied lighting conditions would be a vast improvement since the current system only considers the user's text input to gauge the emotional state and intent. Smoother integration of music generation with reduced latency would enhance the real-time experience. Future iterations should explore a wider emotional palette beyond basic emotions. Implementation of server-side rendering and optimization to reduce network-related latency. We also identified the need for a more comprehensive design language mapping emotions to audiovisual parameters, potentially drawing from established research in color psychology and music theory.

7 CONCLUSION AND FUTURE DIRECTIONS

We developed a prototype emotion-aware multimedia experience system that integrates real-time input, generative AI, and user feedback. Our preliminary results indicate that personalized, responsive multimedia can create more engaging and emotionally resonant experiences. The addition of gesture control provides intuitive interaction mechanisms that participants found natural and expressive.

Beyond creative expression, this technology may be used in therapeutic settings to help users recognize and regulate emotions, especially in youth or clinical populations. Our long-term vision includes adaptive systems for mindfulness, emotion training, and even game industry.

Future plans include testing with a larger, more diverse participant pool to understand cultural variations in emotional expression and perception. We aim to refine the generative components to reduce latency and improve responsiveness. Additionally, integrating more robust emotional measurement tools such as GSR (Galvanic Skin Response) sensors could provide objective physiological data to complement subjective reports.

Technical improvements will focus on three primary areas:

- Optimizing the music generation pipeline for lower latency and more seamless transitions
- Developing more sophisticated emotion-to-media mapping algorithms, potentially employing reinforcement learning to discover optimal parameters based on user feedback
- Enhancing the gesture recognition system to support more complex interactions and continuous control paradigms

We also plan to explore domain-specific applications in areas such as:

- Mental health: Developing targeted experiences for anxiety reduction and emotional regulation
- Education: Creating emotionally responsive learning environments that adapt to student engagement
- Gaming and entertainment: Integrating our approach into interactive narrative experiences

This prototype demonstrates the feasibility and potential of emotion-driven multimedia experiences while highlighting the challenges of creating truly responsive systems. As generative AI continues to advance, we anticipate that the gap between emotional expression and media generation will continue to narrow, enabling increasingly seamless and meaningful human-computer interactions.

8 ACKNOWLEDGEMENTS

8.1 Software Used

TouchDesigner for visual generation, MusicGen for audio generation, Streamlit for UI development, MediaPipe for gesture recognition.

8.2 Help and Support

Feedback from classmates and professor during design reviews significantly shaped our approach. We also acknowledge the online TouchDesigner community for technical assistance with complex rendering pipelines and performance optimization.

8.3 Knowlededge Base of Literature on Human emotions and effects of media on emotions

Z. Xiong, et.al [17], A. Ho, et.al [6], Stamkou, E. et.al [16], Lerner, Jennifer S. et.al [11], Y. Huang et.al [7], Raglio, A. et.al [13], Kokinous, J. et.al [10], Cuadrado, F. et.al [2], S. P. Hutchinson et.al [8]

WORK DIVISION

- Aditya : Worked with Shishir in setting up the visual pipelines of StreamDiffusion and Noise Generation. Also worked on setting up the Audio pipeline in TouchDesigner along with Yifan. Worked on creating the QuestionGenerator and PromptGenerator RAG based LLM models. Setup App in Streamlit with Shishir.
- Ankit : Worked on initial testing of various LLMs during project ideation. Tested audio-generative models suitable for project use-case that would be cost effective. Helped set up MusicGen by Meta locally. Fine-tuned prompts to get efficient audio output from the music generator.
- Shishir : Worked with Aditya in setting up the visual pipelines of StreamDiffusion, noise generation in TouchDesigner. Worked on Integrating the music generation model by Meta into the system locally.Helped develop the UI on Streamlit with Aditya.
- Yifan : Gesture recognition algorithms for audio-visual content manipulation, exploration of server-based visual generation approaches instead of local pretrained models, and experimental evaluation.

REFERENCES

[1] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and Controllable Music Generation. arXiv:2306.05284 [cs.SD] https://arxiv.org/abs/2306.05284

[2] Francisco Cuadrado, Isabel Lopez-Cobo, Tania Mateos-Blanco, and Ana Tajadura-Jiménez. 2020. Arousing the sound: A field study on the emotional impact on children of arousing sound design and 3D audio spatialization in an audio story. *Front. Psychol.* 11 (May 2020), 737.

[3] Derivative.ca. [n.d.]. Derivative — derivative.ca. https://derivative.ca/. [Accessed 21-04-2025].

[4] Antje B. M. Gerdes, Matthias J. Wieser, and Georg W. Alpers. 2014. Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains. *Frontiers in Psychology* Volume 5 - 2014 (2014). https://doi.org/10.3389/fpsyg.2014.01351

[5] Tanisha Hisariya, Huan Zhang, and Jinhua Liang. 2024. Bridging Paintings and Music – Exploring Emotion based Music Generation through Paintings. arXiv:2409.07827 [cs.SD] https://arxiv.org/abs/2409.07827

[6] A. Ho. 2023. The Role of Visual Aesthetics in Emotional Response. https://doi.org/10.54941/ahfe1004170

[7] Y. Huang. 2024. A theory of emotion based on a universal model. https://www.nature.com/articles/s41599-024-02869-x#citeas

[8] Samantha P. Hutchinson. 2019. The Effects of Visual Stimuli on Induced Emotional Responses in Popular Music. https://api.semanticscholar.org/CorpusID:231853744

[9] Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, and Kurt Keutzer. 2023. StreamDiffusion: A Pipeline-level Solution for Real-time Interactive Generation. (2023). arXiv:2312.12491 [cs.CV]

[10] Jenny Kokinous, Sonja A Kotz, Alessandro Tavano, and Erich Schröger. 2015. The role of emotion in dynamic audiovisual integration of faces and voices. *Soc. Cogn. Affect. Neurosci.* 10, 5 (May 2015), 713–720.

[11] Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and Decision Making. *Annual Review of Psychology* 66, Volume 66, 2015 (2015), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

[12] Vivian Liu, Tao Long, Nathan Raw, and Lydia Chilton. 2023. Generative Disco: Text-to-Video Generation for Music Visualization. arXiv:2304.08551 [cs.HC] https://arxiv.org/abs/2304.08551

[13] Alfredo Raglio, Lapo Attardo, Giulia Gontero, Silvia Rollino, Elisabetta Groppo, and Enrico Granieri. 2015. Effects of music and music therapy on mood in neurological patients. *World J. Psychiatry* 5, 1 (March 2015), 68–78.

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752

[16] Corona R. Stamkou E., Keltner D. 2024. Emotional palette: a computational mapping of aesthetic experiences evoked by visual art. *Nature* (2024). https://doi.org/10.1038/s41598-024-69686-9

[17] Zhiyong Xiong, Xinyu Weng, and Yu Wei. 2022. Research on the Influence of Visual Factors on Emotion Regulation Interaction. https://doi.org/10.3389/fpsyg.2021.772642