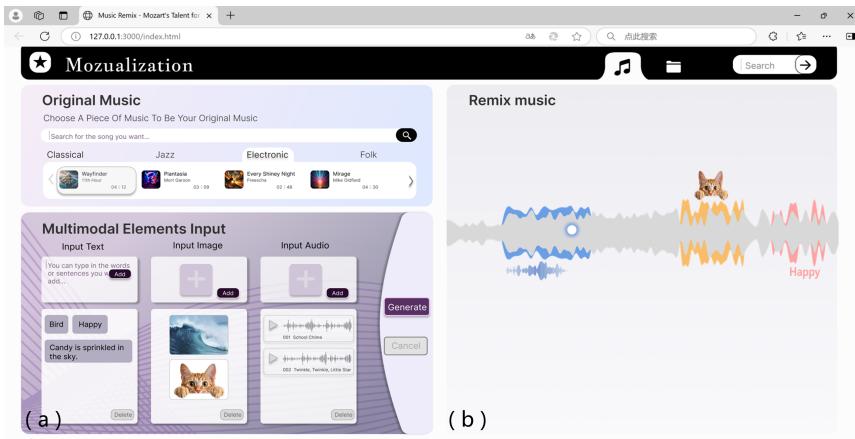


1 Mozualization: Crafting Music and Visual Representation with Multimodal AI

2
3 ANONYMOUS AUTHOR(S)



21 Fig. 1. The interface of Mozualization includes a material upload area for text, images, and audio clips (a), and a visualization of the
22 generated music (b).

23
24 In this work, we introduce Mozualization, a music generation and editing tool that creates multi-style embedded music by integrating
25 diverse inputs, such as keywords, images, and sound clips (e.g., segments from various pieces of music or even a playful cat's meow).
26 Our work is inspired by the ways people express their emotions—writing mood-descriptive poems or articles, creating drawings
27 with warm or cool tones, or listening to sad or uplifting music. Building on this concept, we developed a tool that transforms these
28 emotional expressions into a cohesive and expressive song, allowing users to seamlessly incorporate their unique preferences and
29 inspirations. To evaluate the tool and, more importantly, gather insights for its improvement, we conducted a user study involving
30 nine music enthusiasts. The study assessed user experience, engagement, and the impact of interacting with and listening to the
31 generated music.
32

33 CCS Concepts: • Human-centered computing → Interaction design; Visualization.

35 ACM Reference Format:

36 Anonymous Author(s). 2025. Mozualization: Crafting Music and Visual Representation with Multimodal AI. In *Proceedings of Make sure
37 to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages.
38
39 <https://doi.org/XXXXXXX.XXXXXXX>

41 1 INTRODUCTION

42 The use of Artificial Intelligence Generated Content has significantly advanced art and music creation, catering to the
43 growing demand for personalized and customized designs, including music and visual content. By leveraging deep

44
45 Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not
46 made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components
47 of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on
48 servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

49 © 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

50 Manuscript submitted to ACM

52 Manuscript submitted to ACM

learning and AI algorithms, individuals without professional expertise can easily create unique musical compositions. The generation framework can be broadly divided into two main stages: **Music Creation** and **Music Editing**. In the creation stage, the main task for AI techniques is to generate foundational musical elements, such as melody, harmony, accompaniment, and rhythm, based on user inputs or predefined parameters. In the editing stage, AI technologies are used for refining fine details in existing compositions and augmenting musical elements. Tasks for composers and musicians in this stage include adjusting tempo, modifying instrumentation, adding expressive elements, and integrating additional layers or effects. With AI tools, this process becomes more efficient and interactive, enabling a richer and more personalized editing experience while maintaining the artistic integrity of the piece.

Advanced AI technologies enable music creation and editing not only through traditional tools but also using multimodal inputs such as semantic text and keywords [1, 7, 13, 15]. While these methods leverage simple descriptive words to specify the type of music to be generated, they often rely on overly automatic processes, lacking fine-grained control over musical details and melody selection. Consequently, the generated music tends to be standardized, lacking personalization and emotional depth. Recent research has explored generating music based on images [3] and other visual inputs [6, 12]. These approaches extract themes, colors, and other elements from visual inputs and translate them into musical components. While effective at infusing emotions from diverse forms of input into music, they often fail to clarify how specific visual content influences music composition or how multiple text and visual representations are integrated or merged in the generated results.

To address these challenges, we developed Mozualization, an AI-based music generation system designed to create personalized music using multimodal inputs. Our approach supports diverse ways to **blend and craft music rich in color, style, and personalization**. **Color** can be exacted from text that conveys a person's emotions and intentions, and/or from an image filled with warm and cool tones, evoking specific emotional atmospheres. **Style** can be highly diverse, derived from patterns in images, the structure of melodies, or the genre and type of music. **Personalization** adds a unique touch by incorporating elements such as a piano melody, a baby's voice, or even the sound of a cat's meow. Our vision is to take these personal and emotional elements, and seamlessly merge them to create a cohesive and expressive piece of music. However, the key challenge lies in how the system can interpret these multimodal inputs and effectively translate them into musical components—such as rhythm, timbre, melody, or harmony—that harmoniously integrate into the composition. Additionally, it is crucial to determine where and how these changes should be applied in the original composition to preserve its coherence while enhancing its richness and individuality.

The main contributions of our work are:

- a user-centric design involving professional musicians and novice enthusiasts.
- an innovative tool for creating multi-style embedded music by integrating diverse inputs.
- a visual representation form to show where and how inputs are applied to enrich the composition.

2 RELATED WORK

In this section, we review relevant works on AI-based music creation and editing.

2.1 AI-based Music Creation

Music creation focuses on generating music from scratch, including tasks such as notation [11], recombination sequencing [10], and symbolic coding for sound synthesis [16]. Integrated music generation tools have been developed to assist composers in structuring their work and creating melodies, harmonies, and rhythms based on predefined musical theories and patterns. Recently, deep learning techniques have been extensively applied to generating music

from diverse music corpora [5]. As early as 1989, Lewis and Todd [17] proposed the idea of using neural networks for automatic music creation, employing RNNs to sequentially generate music. Since then, various deep learning models have been rapidly developed for music generation. MidiNet [19] was designed to learn stylized musical elements from existing melodies for approximate creation, while MuseGAN [9] generated polyphonic accompaniments by creating multi-track instrumental sounds from melodies. These approaches have laid a solid technical foundation for music creation. To enhance the creativity and interoperability of music creation, more advanced AI models have been developed to generate music through textual prompts and images. Google’s MusicLM [1] and Meta’s MusicGen [7] generate music in various styles based on detailed text prompts, while ByteDance’s Seed-Music [2] integrates style descriptions, audio references, and music score as multimodal inputs. Additionally, MeLFusion [6] synthesizes music by using text descriptions and corresponding images. Although these models provide diverse approaches to music creation, they rely on an automated generation process that lacks transparency. Musicians and composers may find it difficult to grasp how music is generated or how they can participate in and influence the creative process.

2.2 AI-based Music Editing

Music editing focuses on modifying and refining existing compositions, including tasks such as audio splicing, pitch and tone adjustment, and applying mixing effects. These tasks typically require musicians to analyze the original material, make creative adjustments, and blend multiple independently recorded tracks, such as instruments and vocals. AI techniques have been introduced to assist musicians in these tasks. For example, Suno [18] employed AI to enable users to splice songs using their own audio input, such as hummed melodies or played chords. Udio¹ features a remix function that adjusts tones, allowing users to transform a song into a different musical style while preserving the original lyrics and main melody. However, these transformations are applied to the entire song, changing its overall style rather than making detailed adjustments to specific attributes or sections of the music. To address this limitation, MusicMagus [20] introduced a zero-sample text-to-music editing approach, enabling users to refine specific musical attributes using text-based prompts. However, their approach requires a certain level of musical expertise to achieve optimal results. Users need to know which attributes should be revised and how to frame them in prompts.

In this work, we aim to design a system that supports users in creating music using multimodal inputs. Through visual representations, users can understand how different elements are merged into the song, enabling them to edit music locally while preserving the overall melody and structure.

3 DESIGN CONSIDERATION

To gain a deeper understanding of how professional musicians create and edit music, we conducted interviews with three target users: a remix creator (1 year of experience), a rapper (4 years of experience), and a self-published media blogger (3 years of experience). The goal of these interviews was to gather diverse insights, expectations, and pain points related to AI music tools. We included the media blogger because many people expressed an interest in associating their blog content with background music, ideally aligning the music with the narrative of the story to enhance engagement and emotional impact.

After three hours of in-depth interviews, we gathered diverse insights into their opinions and needs (as summarized Fig. 2). Our target users shared their approaches to creating and editing music, highlighted challenges in sourcing suitable sound materials, and emphasized the need for more diverse and efficient methods to assist in music generation.

¹<https://www.udio.com/>

157		Lin	"Finding the right position is important"	The key to mixing different elements is to find the right place for the instruments, the drums, the rhythmic variations and so on. This requires creators to have a certain knowledge of music theory in order to get a harmonious sounding piece.
158			"Randomness is beneficial in music creation"	Unknown elements are added that may have unexpected effects in the music. I expect random fusion effects to arise, and sometimes I use some AI tools for inspiration.
159			"Challenges in getting sampling elements"	Rap and electronic music rely heavily on sampling, but without the accumulated elements it can be a pain to find the right material on multiple platforms!
160		Zhen	"Expect quick sample demo mixing tools"	In collaboration with beat makers, it's difficult to get others to understand your intentions correctly and quickly, when you can only communicate and listen to feedback repeatedly to make sure that the mood and feeling you want is being expressed correctly.
161			"Library determines sound effects and BGM"	Without a professional editor, the choices I can have for soundtracks are very limited. It's not uncommon to not be able to find a soundtrack that stylistically matches the content of my videos!
162			"Looking forward to more input methods"	I've tried adding some film dialogue to the BGM, and audience really resonates with what I'm creating. It would be great if I could input the environment I was shooting in and the story of the script to co-influence BGM!
163				
164				
165				
166				
167				
168				
169				
170				
171				
172				
173				
174				
175				
176				
177				
178				
179				
180				
181				
182				
183				
184				
185				
186				
187				
188				
189				
190				
191				
192				
193				
194				
195				
196				
197				
198				
199				
200				
201				
202				
203				
204				
205				
206				
207				
208				

Fig. 2. Suggestions and feedback from the target users.

Based on the interview results, we identified the following design requirements:

- (1) **Multimodal Input.** The system should support diverse inputs such as text, images, and music clips, providing personalized pathways to convey users' emotions and intentions. While each input type has been used individually in music generation, combining these inputs, as well as translating and integrating their information into the music generation process, remains a challenge to be addressed.
- (2) **Visual Representations and Guidance.** The tool should offer visual representations showing how these inputs influence the generated music, along with guidance on where users can make changes or add elements. This feature is essential for supporting users with varying levels of expertise and accommodating different musical styles, helping them navigate and progress in their creative process.
- (3) **Real-time Feedback.** The tool should enable real-time adjustments and previews, allowing users to modify parameters such as tempo, volume, and sound during music generation. Instant feedback on adjustments ensures that the music aligns with the user's creative vision, making the process more intuitive and efficient.

4 MOZUALIZATION

We created Mozualization, a music creation and editing tool, as illustrated in Fig. 1. The interface is structured into two main sections. On the left side, users can search through a music library to select a foundational track and provide custom inputs, such as text, images, or audio files, to convey their creative intentions. The system then processes these elements to synthesize a personalized piece of music. The generated music is visualized and presented in the playback area on the right, where users can see how information extracted from diverse inputs is integrated into the composition. A moving dot progresses along with the music playback, creating a bridge between what users perceive visually and what they hear aurally. This dual-section design allows users with little to no professional music knowledge to engage in the creative process, offering an accessible platform for musical expression.

As depicted in Fig. 3, the workflow of Mozualization consists of three main components: multimodal data sonification, mixing processing, and music visualization.

Multimodal data sonification. This component focuses on how information is extracted from user-provided text and images and translated into audio clips. The system processes user inputs by converting text and image data into audio representations for subsequent mixing. For text data, color-related words are extracted and stored for mapping to

Manuscript submitted to ACM

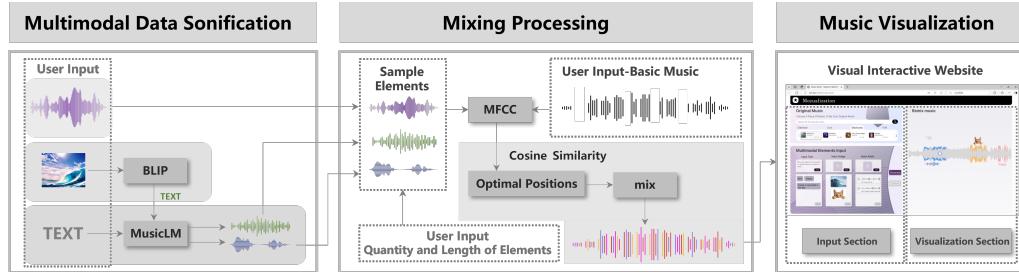


Fig. 3. The workflow consists of three main components: multimodal data sonification, mixing processing, and music visualization.

the music visualization. For image data, the system employs the BLIP model [14], which analyzes the image to extract key features and generates corresponding textual descriptions in natural language. For example, if a user uploads an image of a cat, the BLIP model might generate a description such as “a yellow and white striped cat,” from which the colors “yellow” and “white” would be extracted. Thus, from both images and user-provided text, two key pieces of information are extracted: descriptive text and colors present in the image or mentioned in the text. The descriptive text is then transformed into audio clips using the MusicLM model [1], while the extracted colors are stored and mapped to the music visualization. This integration creates a seamless connection between visual and auditory elements, enhancing the coherence and expressiveness of the generated music.

Mixing processing. This component focuses on determining where the generated audio clips should be integrated into the original music. To achieve this, the system analyzes each audio clip using the audio loading and beat tracking functions from the Librosa library [4], enabling the extraction of fundamental parameters such as audio duration, pitch, and tempo. Next, the system employs the Mel-Frequency Cepstral Coefficients (MFCC) method [8] for feature extraction, capturing essential acoustic characteristics of both the generated audio clips and the original music. To assess their compatibility, the system calculates the cosine similarity between the MFCC feature parameters of the audio segments. Cosine similarity measures the alignment between two feature vectors, with values ranging from -1 to 1 , where values closer to 1 indicate a higher degree of similarity. Using these similarity results, the system identifies optimal positions within the original music for integrating the generated audio clips. It then performs mixing operations at these points, ensuring a seamless and coherent audio composition that maintains the integrity of the original piece.

Music visualization. This component shows where and how the extracted information is embedded into the original music. We developed a stacked streamgraph visualization to represent the generated composition, as shown in Fig. 4. Each input, such as keywords, text, or images, is represented as a layer in the streamgraph. The extracted color information is mapped to the layer’s color, while the layer’s position indicates where the information is merged into the music. The thickness of the layer reflects the intensity of the added audio clip. Additionally, input elements such as text or images are displayed near their corresponding layers for clarity. A moving dot progresses along the streamgraph in sync with the music playback, helping users connect what they see with what they hear. This visualization enables users to understand how diverse inputs influence the original music, empowering them to make further edits. For example, users can upload new text or images and observe how both the music and visualization change.

5 USER EVALUATION AND RESULTS

We conducted a user evaluation to assess the usability of Mozualization and, more importantly, to gather feedback through interviews to identify areas for improvement in the next iteration of the design.

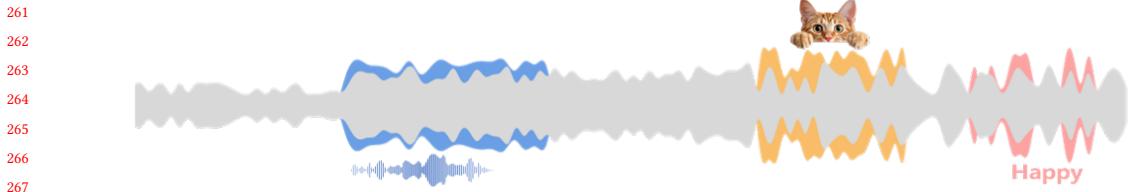


Fig. 4. A stacked streamgraph visualization is used to represent the generated composition, with each layer corresponding to an input element. Preview images and text are embedded within the layers to enhance clarity and provide context.

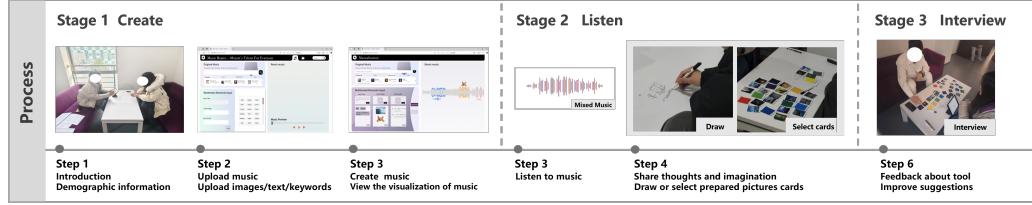


Fig. 5. The evaluation consisted of three stages: create, listen, and interview.

5.1 Evaluation

We invited nine music enthusiasts to use our prototype to create their own music and share their insights on the interface design and interaction techniques provided by the tool. These participants came from different majors at the university and all reported experience in searching for and listening to songs on various music applications. Additionally, all participants expressed an interest in posting blogs with pictures as part of their daily routines.

As shown in Fig. 5, the evaluation process consisted of three stages. In the first stage, we introduced the goal of the evaluation and provided an overview of our music generation tool. Participants were asked to share demographic information, such as their interests and experience with music applications. They were then instructed to choose a piece of music and upload it into the system. Additionally, participants were asked to upload images or input text/keywords to describe their mood or the elements they wished to incorporate into the music. After the music was generated, participants were invited to listen to it and view the visual representations, which demonstrated how their inputs were integrated into the result. They were encouraged to use the system to edit the music as needed. In the second stage, we played a pre-generated piece of music. While listening, participants were encouraged to share their thoughts and imagination regarding what they heard, as well as what they liked or disliked about the music. Furthermore, participants were invited to use colorful pencils to draw or select prepared pictures to visually represent how the music made them feel. All discussions, drawings, and selected pictures were recorded for further analysis. In the third stage, participants were invited to use our system to create additional music if they wished. Following this, we conducted interviews to gather feedback about their experiences with the tool and their suggestions for improvement. The entire evaluation process lasted approximately one hour. The study received IRB approval for its protocol (University Research Ethics Review Panel, anonymized for peer review).

5.2 User Experience in Music Creation

Participants actively engaged with Mozualization and expressed a high level of satisfaction with its interface and functionality. The system was described as user-friendly and intuitive. The combination of text, images, and audio input facilitated flexible idea expression, allowing participants to explore and realize their creativity. Among the nine

313 participants, eight had no prior knowledge in music composition and editing, while one had only engaged in basic
314 music editing. However, following their interaction with Mozualization, all participants developed new insights into the
315 process of music creation. They iteratively refined their compositions, experimenting with various input combinations to
316 better align with their personal preferences and creative goals. In addition, all participants indicated a strong willingness
317 to use Mozualization again. They noted that the system's interactive stimulated their curiosity and fostered a sense of
318 engagement, increasing their motivation to use the tool in the future.
319

320 5.3 Perception of Generated Music and Visualization

321 Participants provided positive feedback on the generated music, noting that Mozualization effectively facilitated the
322 fusion of musical components, resulting in a cohesive and engaging auditory experience. All participants found it
323 exciting to identify the elements they had provided and were intrigued by how these inputs were transformed into
324 musical components and seamlessly merged with others. They appreciated the system's ability to integrate diverse
325 elements into harmonious compositions, even when the individual components differed significantly in style. For
326 instance, a peaceful melody and Peking Opera were successfully merged with complementary rhythm and placement.
327 One participant expressed initial skepticism about incorporating bell chimes, fearing they might create a harsh or
328 discordant effect. However, after listening to the generated composition, he noted that the bell chimes enriched the
329 auditory experience: "The appearance of the bell with the music feels like someone is praying," perfectly aligning with
330 the narrative conveyed by the music. Additionally, three participants highlighted that combining real natural sounds
331 with the music elicited stronger emotional responses. They emphasized that this approach enhanced the music's ability
332 to evoke imagery, creating an immersive and emotionally resonant experience for listeners.
333

334 Participants also provided positive feedback on the music visualization, emphasizing its utility in enhancing their
335 understanding of the generated compositions. By observing waveform changes and input elements, they were able to
336 better comprehend how the diverse elements were integrated into the final output. Participants noted that this feature
337 provided an intuitive way to connect auditory and visual elements. The visualization not only enriched the music's
338 sense of imagery but also effectively conveyed the musical qualities of the text and images, further deepening user
339 engagement with the creative process.
340

341 5.4 Improvement Suggestions

342 Some participants provided constructive feedback for enhancing Mozualization to better support the generation of
343 high-quality musical compositions. One participant recommended implementing measures to avoid frequency masking
344 during audio mixing, which occurs when two sounds occupy the same frequency range, leading to their mutual
345 concealment and a reduction in the clarity of the final mix. Moreover, two participants noted that when blending a
346 melodic audio clip with another piece of music, a more nuanced mixing processing approach might be required to
347 coordinate the tuning discrepancies between them. Such adjustments would help ensure a smoother integration of the
348 audio components, further improving the overall cohesiveness of the generated music.
349

350 6 DISCUSSION

351 Mozualization represents an initial design for generating customized music using diverse inputs. Our concept is that text,
352 images, and audio clips serve as powerful mediums for expressing emotions, each carrying a unique "color" and "sound"
353 when abstracted from their original forms. By encouraging innovative ways to merge these inputs, Mozualization
354 enables the creation of outputs that effectively convey users' emotions, whether in a musical or visual form. Visualization,
355 in particular, plays a crucial role in showing how and where diverse inputs are integrated into the music. This not
356

365 only helps users understand the generated results but also provides a platform for further refinement. Users can edit
366 a wide range of variables, including the color, shape, position, and texture of the visualization, as well as the tone,
367 rhythm, timbre, and volume of the music. These customized features open up opportunities for future exploration and
368 innovation in creative expression. Such versatility also makes Mozualization suitable for a wide range of applications.
369 For instance, in art galleries, it can generate music to complement exhibit pictures, enhancing the viewer's experience.
370 In opera, it can create background visuals that immerse the audience in the narrative, making performances more
371 engaging and emotionally impactful.
372

373 While Mozualization enhances music creation, it has limitations requiring further exploration. Converting images to
374 textual descriptions may lose details, limiting their expressive potential in music composition. Additionally, the current
375 editing capabilities lack precision, making fine-grained adjustments and complex creative expression challenging. To
376 address these issues and support more intuitive music creation, we propose three future research directions:
377

378 **Enhancing Visualization Editing Capabilities.** It is necessary to enhance the current visualization editing
379 functionality to support fine-grained control over added elements. For example, adjusting the height of elements in
380 the visualization could modify pitch, while altering their horizontal position could fine-tune their timing within the
381 composition. Additionally, users might wish to make localized adjustments, updating the melody of particular sections.
382 These advanced editing capabilities would enable users to explore and refine their creative ideas with greater precision,
383 striking a balance between algorithm-driven generation and user-driven customization. Such functionality would be
384 particularly valuable for users with prior musical knowledge, allowing them to leverage their expertise to achieve more
385 sophisticated and nuanced compositions.
386

387 **Extending to Immersive Environments.** Expanding the system to immersive environments, such as VR or AR,
388 could create a more engaging and sensory-rich platform for music creation. In these environments, music elements could
389 be designed as virtual objects that users can place, move, and manipulate within 3D space, allowing users to compose
390 music in a dynamic way. Additionally, immersive environments could enable users to experience their compositions
391 from a variety of perspectives, moving through the virtual space to hear how different arrangements and spatial
392 placements of musical elements affect the overall sound, fostering exploration and discovery in ways that traditional
393 2D interfaces cannot replicate.
394

395 **Expanding Input Modalities.** Incorporating additional input modalities could greatly enhance user interaction
396 within the music generation process, particularly in immersive environments. Unlike in 2D interfaces, where editing is
397 limited, users themselves could become active control points in the generated music. For example, users' positions,
398 movements, and gestures could dynamically influence the playback, such as adjusting volume or tempo as they move
399 through the space or interact with virtual objects. This would enable a more embodied and intuitive way of editing and
400 personalizing the music, deepening the connection between the user and their creation.
401

402 7 CONCLUSION 403

404 In this work, we developed Mozualization, a multimodal music generation tool designed to inspire music and art
405 enthusiasts to explore innovative approaches to their creations. We believe this tool has significant potential for
406 refinement, with its resulting innovations applicable to a variety of scenarios. By delving into future improvements
407 and research directions, we can further advance multimodal generation technology, unlocking new opportunities and
408 fostering greater innovation in the fields of music and art creation.
409

417 REFERENCES

- 418 [1] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts,
 419 Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. MusicLM: Generating Music From Text. arXiv:2301.11325 [cs.SD]
 420 <https://arxiv.org/abs/2301.11325>
- 421 [2] Ye Bai, Haonan Chen, Jitong Chen, et al. 2024. Seed-music: A unified framework for high quality and controlled music generation.
 422 arXiv:2409.09214 [cs.SD] <https://arxiv.org/abs/2409.09214>
- 423 [3] Mattia Bellini. 2022. Interactive digital narratives as complex expressive means. *Frontiers in Virtual Reality* 3 (2022), 854960. <https://doi.org/10.3389/frvir.2022.854960>
- 424 [4] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music
 425 Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, Kathryn Huff and James Bergstra (Eds.). 18 – 24. <https://doi.org/10.25080/Majora-7b98e3ed-003>
- 426 [5] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. 2019. Deep Learning Techniques for Music Generation – A Survey.
 427 arXiv:1709.01620 [cs.SD] <https://arxiv.org/abs/1709.01620>
- 428 [6] Sanjoy Chowdhury, Sayan Nag, K J Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. 2024. MELFuSION: SynthesizinFg Music from Image
 429 and Language Cues Using Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26816–26825.
 430 <https://doi.org/10.1109/CVPR52733.2024.02533>
- 431 [7] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music
 432 generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) (NIPS '23). Curran
 433 Associates Inc., Red Hook, NY, USA, Article 2066.
- 434 [8] S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.
 435 *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 4 (1980), 357–366. <https://doi.org/10.1109/TASSP.1980.1163420>
- 436 [9] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. MuseGAN: multi-track sequential generative adversarial networks for
 437 symbolic music generation and accompaniment (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 5, 8 pages.
- 438 [10] Matthew Duignan. 2008. *Computer Mediated Music Production: a Study of Abstraction and Activity*. Ph.D. Dissertation. Open Access Te Herenga
 439 Waka-Victoria University of Wellington. <https://doi.org/10.26686/wgtn.16945879.v1>
- 440 [11] J Julia, P D Iswara, and T Supriyadi. 2018. Song creation by using computer music notation. *IOP Conference Series: Materials Science and Engineering*
 441 434, 1 (nov 2018), 012055. <https://doi.org/10.1088/1757-899X/434/1/012055>
- 442 [12] Jaeyong Kang, Soujanya Poria, and Dorien Herremans. 2024. Video2Music: Suitable music generation from videos using an Affective Multimodal
 443 Transformer model. *Expert Syst. Appl.* 249, PC (Sept. 2024), 17 pages. <https://doi.org/10.1016/j.eswa.2024.123640>
- 444 [13] Max W. Y. Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, Jitong Chen, Yuping Wang,
 445 and Yuxuan Wang. 2023. Efficient neural music generation. In *Proceedings of the 37th International Conference on Neural Information Processing
 446 Systems* (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 766, 14 pages.
- 447 [14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language
 448 Understanding and Generation. arXiv:2201.12086 [cs.CV] <https://arxiv.org/abs/2201.12086>
- 449 [15] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2023. Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion.
 450 (2023). arXiv:2301.11757 [cs.CL] <https://arxiv.org/abs/2301.11757>
- 451 [16] Julius O. Smith. 2004. Virtual Acoustic Musical Instruments: Review and Update. *Journal of New Music Research* 33, 3 (2004), 283–304. <https://doi.org/10.1080/0929821042000317859>
- 452 [17] Peter M Todd. 1989. *A connectionist approach to algorithmic composition*. Vol. 13. JSTOR, 27–43.
- 453 [18] Víctor Villalpando. 2024. Inteligencia artificial al alcance de todos: plataformas y proyectos destacados. *CLIP de SEDIC: Revista de la Sociedad Española
 454 de Documentación e Información Científica* (06 2024), 14–26. <https://doi.org/10.47251/clip.n89.143>
- 455 [19] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. 2017. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music
 456 Generation. *ArXiv* abs/1703.10847 (2017). <https://api.semanticscholar.org/CorpusID:2002865>
- 457 [20] Yixiao Zhang, Yukara Ikemiy, Gus Xia, Naoki Murata, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024.
 458 MusicMagus: Zero-Shot Text-to-Music Editing via Diffusion Models. arXiv:2402.06178 <https://arxiv.org/abs/2402.06178>
- 459
- 460
- 461
- 462
- 463
- 464
- 465
- 466
- 467
- 468