Touch, Sound, and Space: Exploring Immersive Music Interaction through AI-Generated Environments

Wanfang Xu
Xi'an Jiaotong-Liverpool University
Suzhou, China

Jifan Yang
Xi'an Jiaotong-Liverpool University
Suzhou, China

Fengwen Zhang
Xi'an Jiaotong-Liverpool University
Suzhou, China

Yu Lu
Xi'an Jiaotong-Liverpool University
Suzhou, China

Lijie Yao
Xi'an Jiaotong-Liverpool University
Suzhou, China

Le Liu
Northwestern Polytechnical University
Xi'an, China

Lingyun Yu*
Xi'an Jiaotong-Liverpool University
Suzhou, China

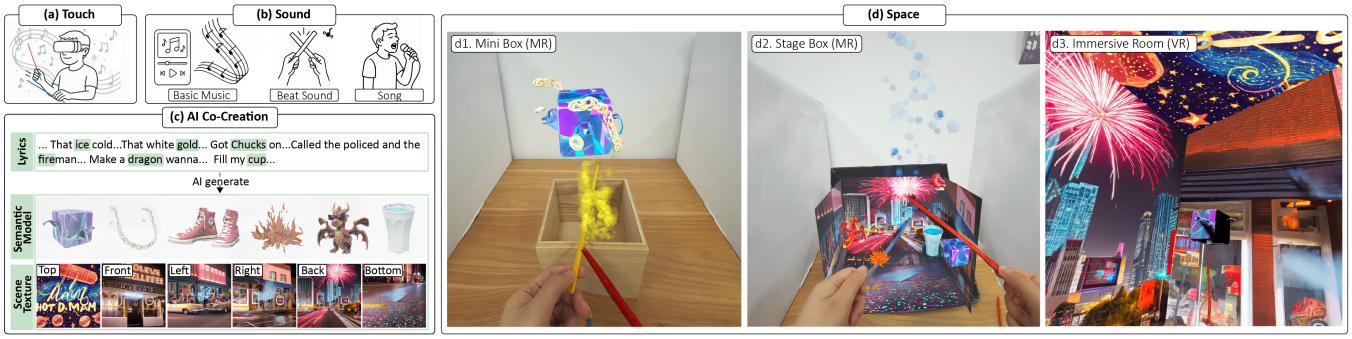

Fig. 1: Our prototype integrates (a) touch interaction, (b) sound, and (c) AI co-creation with music to generate dynamic visuals across (d) three spatial configurations: (d1) a handheld AR music box, (d2) a table-sized AR stage, and (d3) an immersive VR environment.

ABSTRACT

We introduce an interactive music system powered by AI-generated content (AIGC) that enables users to engage with music through multimodal interactions involving touch, sound, and spatial immersion. Motivated by the desire to enhance engagement and emotional connection with music, our system enables users to co-create and interact with musical content. Users upload a song and a descriptive text prompt, from which the system generates 3D visuals. During playback, users can embed their own audio inputs and trigger responsive visual effects such as color-driven point clouds using tangible controls. To explore how spatial scale and embodiment shape user experience, we implement the system across three increasing spatial scale and embodiment: (1) a handheld AR music box, (2) a table-sized stage box, and (3) a fully immersive VR environment. Through a user study, we investigate how different levels of immersion and interaction influence user engagement, emotional response, and sense of presence. Our findings demonstrate the potential of combining AIGC with embodied interaction to enrich creative expression and enhance immersive musical experiences.

*Corresponding author

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 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 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 servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

VINCI '25, December 01–03, 2025, Linz, Austria

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

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Human-centered computing → Systems and tools for interaction design**.

KEYWORDS

Immersive Creation, Music Generation, Visual Design

ACM Reference Format:
Wanfang Xu, Jifan Yang, Fengwen Zhang, Yu Lu, Lijie Yao, Le Liu, and Lingyun Yu. 2025. Touch, Sound, and Space: Exploring Immersive Music Interaction through AI-Generated Environments. In *Proceedings of The 18th International Symposium on Visual Information Communication and Interaction (VINCI '25).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Music is a form of expression shaped by musicians, while its meaning and impact unfold through the audience's perceptions, emotions, concepts, and desires. Interpreting music is inherently subjective, shaped by individual interpretation and personal preference. To deepen individual engagement and enrich the listening experience, we explore how music content and spatial environments can be

co-designed by the audience, and whether their experience is enhanced when the entire musical space responds to their interactions. Specifically, we examine which design elements, beyond the music itself, contribute to a coherent listening environment and elevate the listener's experience. In live performances, dynamic lighting [18], thoughtful stage design [11], and the collective atmosphere of the crowd [29] work together to intensify emotional impact. Likewise, high-fidelity spatial audio (e.g., surround or binaural sound) can deepen immersion and make listeners feel enveloped by the music. Our observations lead us to consider three key design factors: **environment**, **music creation**, and **user interaction**.

First, allowing audiences to contribute to, modify, or reinterpret musical content fosters a sense of **ownership**, **emotional engagement**, and **personal expression**. This engagement shifts users from passive listeners to active co-creators, enabling creativity and deepening their connection to the music. With the advancement of AI technologies, co-creation has become increasingly accessible, particularly in domains such as digital painting and music generation. However, the challenge lies in designing systems that support meaningful **co-creation**, where the process itself is as engaging as the outcome. This concern motivates us to investigate the research question **RQ1: How to facilitate engaging AI-user co-creation?**

Second, the music environment acts as the stage where the musical experience unfolds. To strengthen listeners' experience, we need to consider: beyond the auditory, how should the **surrounding space** be designed? Visual elements, such as symbolic glyphs, dynamic lighting, and expressive color changes, can work in tandem with sound to create a rich, multisensory experience. Spatial factors, including scale and layout, also play a critical role in shaping user perception and emotional response [17]. For example, a small, handheld space like a music box may evoke intimacy and a sense of personal connection, while a large-scale immersive setting, such as a virtual concert hall, may intensify embodiment and emotional impact. These considerations motivate **RQ2: Which elements and spatial scales best enhance user experience in a musical environment?**

Third, within an immersive environment, or through tangible objects like a music box, user interaction serves as the bridge that transforms passive listening into active engagement. Meaningful interaction enables audiences to influence the unfolding musical experience and perceive its evolution in real time [23]. In addition, interaction is deeply intertwined with the other two design factors: it brings music creation into the realm of lived experience and dynamically reshapes the environment through responsive visuals, adaptive spatial audio, and contextual modulation. These connections prompt the question **RQ3: What forms of music interactions and responses do users expect?**

To explore the interplay of design factors and the balance between passive listening and creative participation, our study investigates what makes music interaction meaningful. We conducted a formative study to capture listener preferences on music materials, environments, and interaction modes. Based on these insights, we developed a prototype that combines textual input with music to generate dynamic visuals across different immersive settings, from handheld to full VR. We then conducted user studies to evaluate its impact on interaction, perception, and emotional engagement. In summary, we contribute to the following aspects:

- A formative study identifying design requirements for immersive music environments.
- AI-enhanced environments enabling music interaction across three levels of immersion.
- Analysis of user experience across different immersive settings.

2 RELATED WORK

In this section, we review related work in three key areas: (1) technologies for music creation, (2) environments for experiencing music, and (3) interaction techniques in music systems. These areas inform distinct aspects of music co-creation and provide a foundation for designing systems that support meaningful musical experiences.

2.1 Music Creation Technologies

Music creation is shifting from a creator-dominated model to one emphasizing co-creation with audience participation [23, 28]. This shift redefines the audience as active contributors rather than passive recipients. Previous practices have explored various forms of audience involvement in music creation [14]. For example, Thomas C. Duffy's interactive performance *"The Critic's Choice"* lets audiences influence composition through voting, while Queen's *"We Will Rock You"* turns listeners into performers via stomping and clapping. Such practices foster creativity and innovation, democratize artistic creation, and enhance interactivity in performance.

The advancement of AI technology has created new opportunities in music creation. AI tools can generate complete compositions from simple text prompts [1, 2, 8] or create music inspired by emotionally rich multimodal inputs [12, 13, 33], enabling personalized experiences. While these technologies make music creation more accessible, their limited controllability and customizability [6], along with users' desire to preserve agency over reliance on automated outputs [19], underscore the need for tools that actively support and amplify human creativity. Motivated by this, our research explores novel approaches to integrating AI into music co-creation, rendering the process more engaging, interactive, and participatory.

2.2 Musical Environment Design

Musical experience extends far beyond the auditory domain. For example, visual [26] and tactile [10] elements can be integrated into a musical environment to enrich the listener's sense of immersion. Enhancing live music atmospheres through environmental design has a rich tradition. Typical examples include the dazzling light shows at rock concerts [3] and the intricate stage designs found in operas [22]. In addition, the size and layout of the venue can be adapted to different musical genres and audiences to shape immersive experiences. For instance, Dobson [7] observed that intimate settings in smaller venues strengthen perceived connections between performers and audiences, whereas larger venues facilitate performances emphasizing collective participation and interaction. Furthermore, design of acoustic properties—including the selection of stage materials [16], strategic device placement [15], and thoughtful architectural layout [17]—can further influence how music is perceived and appreciated by audiences.

In recent years, virtual reality (VR) and augmented reality (AR) technologies have been increasingly applied to musical contexts [4]. VR enables the creation of fully virtual environments, allowing users

to engage with music in deeply immersive ways [21]. In contrast, AR overlays visual elements onto the physical environment, producing context-aware and interactive experiences [34]. These technologies raise two key questions: How does the visibility of the physical environment affect musical interaction? Which levels of immersion and spatial scale do users prefer? To address these, we created a set of musical environments varying in immersion and spatial configuration, and investigated their effects on user experience.

2.3 Interaction Techniques in Music Creation

Traditional forms of interaction in music creation have primarily relied on playing instruments. While enjoyable, it generally requires musical theory knowledge and performance skills [32], which may discourage individuals with limited musical backgrounds.

The development of interactive technologies has largely reshaped how music is created, presented, and experienced. Smart music controllers, for example, allow users to manipulate tempo, chords, and sound effects through touch [31] and swipe gestures [9]. Motion-sensing devices further expand interaction by enabling users to control music generation and transformation through body movements [30]. These technologies make music interaction more engaging and accessible, especially for individuals without formal musical training. Building on this trend, we integrated interactions based on touch and sound within musical environments to support creative expression and enrich the overall experience.

3 FORMATIVE STUDY

To understand how users engage with music and envision ideal co-creation, we conducted a formative study to explore user preferences regarding music creation, environment, and interactions.

3.1 Study Settings

Participants. We recruited six participants with varying backgrounds: two (P1, P2) with no formal music training, three (P3, P4, P5) with brief experience learning an instrument, and one expert (P6) who had passed the highest level of piano examinations. All participants identified as frequent consumers of music, with five reporting daily engagement, one reporting weekly engagement.

Apparatus. We provided an illustrative VR environment featuring a skybox. The environment was created using the Unity 3D engine and deployed on a desktop (Intel Core™ i9, 3.5GHz, 64GB RAM, GeForce RTX3090, 24 GB video memory). Participants experienced the environment through a Meta Quest 3 headset. Upon entering the environment, a predefined music track was automatically played, offering an integrated audiovisual experience.

Procedure. Participants first provided informed consent and described their regular music listening habits, including the typical contexts in which they engage with music. Next, they were introduced to the VR environment. Finally, participants reflected on this experience, comparing it with listening through traditional methods and discussing their willingness to use immersive environments for music listening. Participants who expressed interest were encouraged to further elaborate on the environment and interaction designs they envisioned or desired in immersive settings.

3.2 Feedback

We audio recorded participant feedback and transcribed verbatim, followed by a thematic analysis. Two investigators independently coded the collected data, and then reviewed the initial codes and grouped into potential themes. After iterative refinement and discussion among the investigators, we identified three main themes: multisensory elements, immersive environment, and interaction.

Multisensory elements enhance music creation. Participants reported engaging with music not only via traditional streaming platforms but also from background audio within multimedia contexts such as videos and games. Three participants (P1, P2, and P4) emphasized that accompanying visual or narrative content enhanced their attentional and emotional connection to the music. P2 further remarked, *"Audio-centric platforms are not my first choice for experiencing music. I only use them when I am multitasking and cannot look at a screen."* Reflecting this preference, participants expressed strong interest in incorporating additional media—such as **visuals**, **text**, and **narrative**—into music creation to spark imagination. Four participants (P2, P4, P5, and P6) emphasized that these elements must be **thoughtfully integrated and emotionally aligned with the music**. As P6 explained, *"For example, combining tense, dramatic music with footage of a storm can create a cohesive atmosphere that amplifies the underlying emotional message of the music."*

Overall, participants were not simply drawn to multimedia for novelty. Rather, they sought cohesive, multisensory experiences in which each element meaningfully contributes to the emotional and narrative arc of the music.

Immersive environments strengthen music engagement. Participants compared everyday music listening and live performances, noting that the rich, multi-sensory nature of live shows adds an emotional depth often missing from solitary listening. As P5 noted, *"At a concert, the waving glow sticks and the energy of the crowd make the music experience more alive and immersive. I wish I could have that kind of feeling in everyday context."* After experiencing the VR environment, participants described it as a powerful way to recapture some of that lost intensity. The idea of creating immersive spaces that mirror the emotional or narrative qualities of the music was well received. However, participants expressed a desire for **varying levels of immersion**. For example, P1 commented, *"I enjoy the immersion that VR offers, but sometimes I want to see my real surroundings so I do not feel completely disconnected."*

Therefore, immersive environments can strengthen music engagement, but flexibility in immersion levels is required to accommodate different user preferences.

Real-time interaction fosters active involvement. Rather than remaining recipients, participants envisioned dynamic and responsive environments that would allow for **active involvement**. P5 proposed, *"I want to interact with the environment, like I can move or touch some artifacts within the scene while listening."* P6 further noted that the environment should adaptively respond to both music properties and user behavior: *"Visual effects that shift with the rhythm of the music and respond to my gestures in real time would make the experience much more engaging."*

Consequently, compelling music experiences go beyond passive reception; participants sought environments that respond to and evolve with their interactions and affective states in real time.

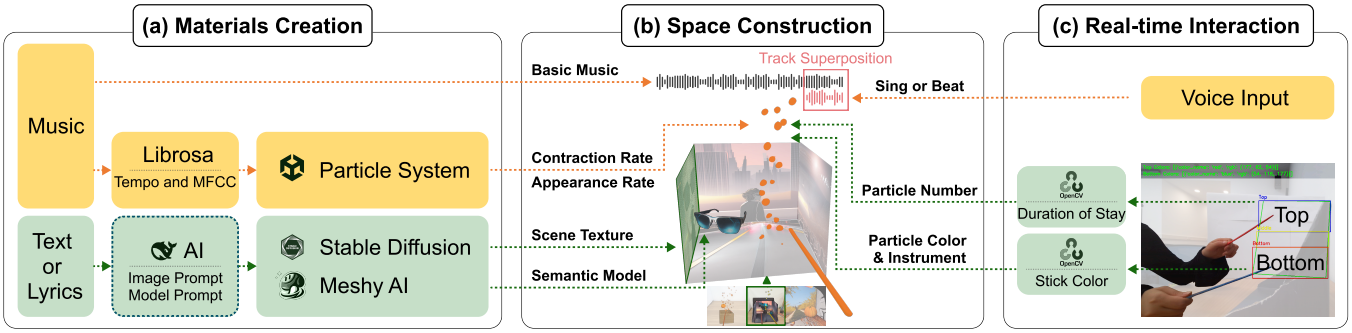

Fig. 2: Framework: (a) Materials Creation—extract music features, drive particles, and generate AI prompts; (b) Space Construction—generate textures/models, overlay sounds, and adjust particles; (c) Real-time Interaction—control color sticks with sound input.

3.3 Design Considerations

Drawing from the formative study, we identified three key considerations for designing an immersive music interaction system:

DC1: Leverage multisensory elements to support creation. Participants emphasized the importance of multimedia elements in music creation—not as decorative features, but as integral tools for deepening emotional and imaginative engagement. To support this, the system should analyze musical features (e.g., tempo, tonality, dynamics) and generate synchronized visual and narrative components that reflect and enhance the emotional tone of the music.

DC2: Create personalized immersive environment. While immersive environments were generally appreciated, participants expressed varying preferences regarding the intensity of immersion. To accommodate these differences, the system should support adjustable levels of immersion, from subtle ambient overlays to fully immersive VR settings, allowing users to adjust the experience according to their context, comfort level, and creative goals.

DC3: Facilitate real-time, interactive engagement. Participants envisioned music experiences that are not only immersive but also interactive and responsive. To foster this sense of agency, the system should enable users to actively engage with the environment——such as moving within the space and adjusting visual effects in real time. Features such as rhythm-responsive animations, gesture-based controls, and behavior-adaptive visuals can be utilized to transform passive listening into a dynamic, embodied experience.

4 PROTOTYPE DESIGN

Guided by the aforementioned design considerations, we developed a prototype system that integrates music, text, and real-time visuals within immersive environments, aiming to create a cohesive and engaging audiovisual experience. As illustrated in Figure 2, the system's workflow comprises three core phases:

(a) Materials Creation. Based on **DC1**, our prototype leverages AI to analyze musical features, with textual input as a complement, generates three types of visual elements—particle effects, scene textures, and 3D models—to reflect the atmosphere of the music.

The workflow begins with the user providing a music track along with textual information, such as lyrics or a brief thematic description. The system extracts the tempo and Mel-Frequency Cepstral Coefficients (MFCCs) from the provided music using the Librosa library [5]. These audio features are subsequently mapped to parameters of particle effects—specifically, contraction rate and appearance frequency—which are passed into Unity's particle system for later rendering. For textual inputs, the system employs a two-stage process facilitated by DeepSeek AI [1]: selective extraction and description generation. In the **selective extraction** stage, the AI extracts keywords from the textual input that best capture the music's theme and conveyed sentiment. If a keyword appears in the lyrics, its corresponding timestamp is also recorded. In the **description generation** stage, the AI uses the previously extracted keywords to generate a set of descriptions. These descriptions are then processed by Stable Diffusion [24] and Meshy AI [2] to produce scene textures that reflect the musical atmosphere and 3D models corresponding to the lyrics.

(b) Space Construction. To implement the personalized immersive environments highlighted in **DC2**, the previously introduced materials are strategically integrated, constructing box-shaped interactive spaces with three designs of increasing scale—**Mini Box**, **Stage Box**, and **Immersive Room**—each offering a distinct level of immersion. In the Mini Box setup, particle effects are superimposed on a hand-sized physical box situated in the real world, as shown in Figure 1 (d1). The Stage Box features four AI-generated scene textures mapped onto the sides of a table-sized stage model (Figure 1 (d2)), enhanced by dynamic particle effects at its center. The Immersive Room utilizes six AI-generated scene textures to construct a fully immersive VR environment (Figure 1 (d3)). Across all modes, if lyrics are provided, AI-generated models derived from their semantics appear synchronously as the corresponding words are played. In the Mini Box, models emerge from within the physical box and float outward. In the Stage Box, models are generated from a portal at the bottom-right corner of the scene and revolve around the central particle effects. In the Immersive Room, models appear directly at the user's gaze position.

(c) Real-time Interaction. Building on **DC3**, our prototype supports two forms of real-time interaction: sound-based interaction and touch-based interaction. For sound-based interaction, users can sing along or beating while listening to the music, with the system capturing these inputs and layering them onto the original track to achieve personalized music enhancement. For touch-based interaction, inspired by the use of batons in orchestral conducting and glow

[1] https://www.deepseek.com/
[2] https://www.meshy.ai/

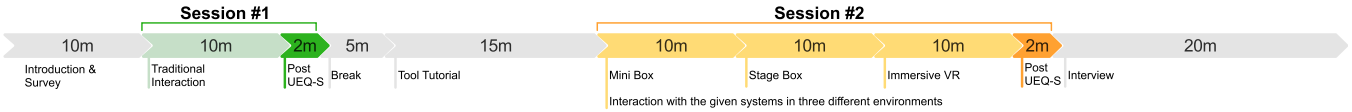

Fig. 3: Overview of the user study procedure: two sessions and an semi-structured interview.

sticks at concerts, the system leverages sticks for interactions. In the Mini Box and Stage Box—where the real-world environment remains visible—users can manipulate physical sticks of different colors to personalize the particle effects. In contrast, the Immersive Room features virtual sticks for interaction. Waving a stick in the upper half of the interactive space gradually increases the number of particles, whereas waving it in the lower half alters particle color to match the stick. Moreover, each stick, distinguished by its color, is assigned a specific instrument. When a MIDI file is provided, waving a stick additionally modifies the instrumentation of the music—for example, the blue stick triggers the sound of an accordion.

5 A USER STUDY

We conducted a user study to gather feedback and insights on music interaction within the prototype system.

Participants. We recruited 12 participants for the study. Six reported daily use of music interaction tools, four used them three to five times per week, one used them one to two times per week, and one rarely used such tools. Prior to the study, participants were asked to provide a piece of music they preferred. The collected pieces reflected diverse personal tastes, spanning genres such as pop, rock, and electronic music. Additionally, participants had the option to provide supplementary textual materials—such as lyrics, thematic descriptions, or original narratives and poems—as creative input for generating scene textures and models.

Apparatus. The experimental environments were developed in C# using Unity 3D. Participants interacted with the system via a Meta Quest 3 headset, leveraging its passthrough capability and immersive rendering functionality. The headset was connected to a desktop computer (Intel Core™ i9, 3.5GHz, 64GB RAM, GeForce RTX3090, 24 GB video memory). User-generated input—including singing, rhythmic beating, and the color and position of handheld sticks—was captured using a Microsoft Xbox Kinect sensor.

Procedure. Each participant completed a 1.5-hour study following the procedure illustrated in Figure 3. The process began with an introduction to the study and the collection of participants' background information related to music. Participants then interacted with the music piece they had previously submitted using a traditional smartphone-based music player and completed a questionnaire about their experience. After a 5-minute break, they engaged in a tutorial introducing the prototype, and interacted with the same music piece under the three environment designs—**Mini Box**, **Stage Box**, and **Immersive Room**—for 10 minutes each, with freedom to choose when and how to interact. Following these sessions, participants completed the same questionnaire based on their experience with the prototype. The study concluded with a semi-structured interview to collect additional insights.

Measures. We employed the short version of User Experience Questionnaire (UEQ-S) [25], using a seven-point Likert scale ranging from -3 to 3, with higher scores indicating more positive feedback

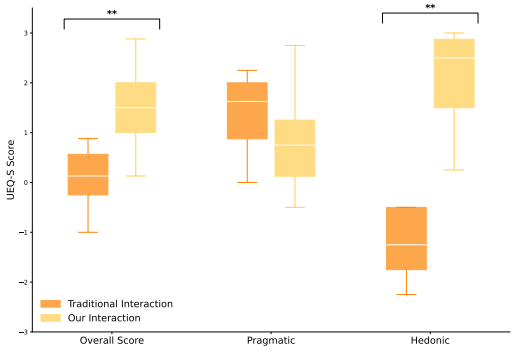

Fig. 4: UEQ-S scores comparing the traditional music player and our prototype.

to evaluate participants' subjective experience across two conditions: the traditional music player and our prototype. The questionnaire consists of two subscales—Pragmatic Quality and Hedonic Quality—each comprising four items. Pragmatic Quality items assess usability aspects including supportiveness, complexity, efficiency, and clarity, while Hedonic Quality items measure excitement, interest, innovativeness and novelty. To supplement the questionnaire, participants were asked to evaluate AI-generated content and their overall experience in a post-study interview. Specifically, for the three immersive environments with distinct spatial scales, participants reported their interaction preferences and how frequently they would engage with each environment.

6 RESULTS

In this section, we present a quantitative analysis of the questionnaire responses and report qualitative insights derived from the interviews.

6.1 Questionnaire Analysis

6.1.1 Analysis Methods. We first conducted a preliminary screening of the collected questionnaire data and identified three participants whose responses exhibited notable discrepancies within a single subscale, with score differences exceeding three points. After cross-referencing these responses with the corresponding interview data, we confirmed that the ratings aligned with their subjective accounts. Therefore, these responses were retained for the final analysis. We then conducted Shapiro-Wilk tests [27] on each subscale across both conditions to assess the normality of data. Except for the Pragmatic subscale under the traditional music player condition ($p = 0.0455$, all data met the assumption of normality ($p > 0.05$). Given this violation and the modest sample size, we used Mann-Whitney U tests [20], a nonparametric statistical test method, for all subsequent analyses to ensure robust group comparisons. Effect sizes were also reported to complement the significance testing results.

6.1.2 Results. Figure 4 illustrates the distribution of questionnaire ratings. In terms of the **overall UEQ-S score**, our prototype

Tab. 1: Mean item scores and differences (Δ) between the traditional music player and our prototype in UEQ-S. Positive Δ values indicate higher ratings for the prototype.

Construct	Item	Traditional (M)	Prototype (M)	Δ
Pragmatic Quality	Supportiveness	0.500	1.250	0.750
	Complexity	2.250	0.583	-1.667
	Efficiency	1.417	1.083	-0.334
	Clarity	1.417	0.083	-1.334
Hedonic Quality	Excitement	-0.167	1.750	1.917
	Interest	-0.167	2.083	2.250
	Innovativeness	-2.167	2.333	4.500
	Novelty	-2.417	2.500	4.917

received higher ratings (M = 1.458, SD = 0.804) compared to traditional interaction with the traditional music player (M = 0.083, SD = 0.594). An independent samples t-test confirmed that this difference was statistically significant (p = 0.001), with a large effect size (Cohen's d = 1.946). The 95% confidence intervals further supported this result, showing no overlap between the two conditions (prototype: [1.004, 1.913]; traditional: [-0.253, 0.420]). Regarding **Pragmatic Quality**, our prototype received lower ratings (M = 0.750, SD = 0.935) than the traditional interaction condition (M = 1.396, SD = 0.808); however, this difference was not statistically significant (p = 0.082), though it reflected a medium effect size (Cohen's $d = -0.739$). In contrast, participants' ratings of the **Hedonic Quality** subscale revealed a significant difference between the two conditions. The prototype scored substantially higher (M = 2.167, SD = 0.862) than the traditional interaction ($M = -1.229$, $SD = 0.661$), with $p < 0.001$ and an extremely large effect size (Cohen's $d = 4.422$). The non-overlapping 95% confidence intervals (prototype: [1.679, 2.654]; traditional: [–1.603, –0.855]) further confirm the significance of the effect.

We further analyzed participants' ratings on each item of the UEQ-S questionnaire. The detailed result is presented in Table 1. The prototype scored significantly higher than the traditional interaction across all four **Hedonic Quality** items, with mean differences ranging from 1.917 to 4.917. The largest difference was observed in the item "Novelty" where the prototype received a mean score of 2.5 compared to -2.417 for the traditional interaction ($\Delta = 4.917$). This was followed by "Innovatieness" ($\Delta = 4.5$). The remaining items "Interest" ($\Delta = 2.25$) and "Excitement" ($\Delta = 1.917$) also demonstrated clear distinctions. However, the traditional interaction outperformed the prototype on the "Complexity" item, with a mean score of 2.25 compared to 0.583 for the prototype ($\Delta = -1.667$), and a similar trend was observed in "Clarity". In summary, while the prototype may be perceived as slightly less intuitive in certain usability aspects, its pronounced strengths in emotional engagement, perceived innovativeness, and overall appeal led to enhanced user experience.

6.2 Interview Results

For the interview data, we followed the same thematic analysis procedure as in the formative study and identified three themes regarding participants' experiences.

Experience of co-creation with AI. All participants submitted music with lyrics for materials creation. The majority of participants (11 out of 12) showed a high level of acceptance toward AI-generated materials from this content. As P1 stated, *"The AI-generated content matched the music I provided—I did not expect the effects to be this good."* However, three participants pointed out that the AI-generated elements did not always meet their expectations, highlighting the need to support further customization of the created content. For example, P3 commented, *"At times, the generated textures and models did not fully reflect the meaning of the lyrics or showed a somewhat inconsistent visual style, particularly when the lyrical content was more abstract."* Despite this, participants generally felt that AI brought novelty and a sense of participation to their co-creation experience, sparking their inspiration. Three participants expressed interest in exploring more imaginative settings in future use, such as experimenting with different musical styles or incorporating more imaginative text materials, rather than relying solely on existing lyrics. One participant further suggested saving the generated scenes or recording the interaction process as a form of creative output.

Perception of Environments. Participants provided positive feedback on the three immersive environments, noting that they effectively enhanced the musical atmosphere through the integration of music-related textures and models. Ten participants specifically highlighted their appreciation for the particle effects, stating that these dynamic elements helped them better perceive the rhythm of the music. The ability to control the particles' color also enabled them to adjust the ambiance of the environment according to their own interpretation. Participants also praised the audio features provided by the environment. All of them experimented with beating to enhance the original music track. Two participants further hummed along with the music during the experience. P11 commented, *"The immersive visuals combined with audio feedback gave me a KTV-like experience, which made me want to sing along to the rhythm."*

Participants expressed diverse preferences regarding spatial scale and embodiment. Six participants favored the fully immersive VR environment, describing it as *"offering the highest level of immersion and the most powerful atmosphere"* (P1). Five participants found the table-sized stage to be the most engaging. P2 remarked, *"I found the blend of virtual and real particularly fascinating, because the virtual elements emerged into the physical world in such unexpected ways! The stage-like design also made it feel like a live performance unfolding right in front of me."* One participant preferred the handheld physical box, holding and rotating it during the experience. This participant saw potential for broader applications, commenting, *"The box could also be replaced with other physical objects, like a drum, which could provide more tangible, musical interactions."* These responses suggest that spatial scale and the embodiment of environment influence user perception and engagement, with different setups offering distinct experiential qualities.

Engagement of interactions. Participants demonstrated varying preferences for interaction styles. Seven participants interacted more actively, frequently utilizing voice input and manipulating sticks to engage in the music. P2 pointed out that our prototype lowered the barrier for music interaction: *"Traditionally, interacting with music often required playing an instrument; but now, I can engage directly with the music with just a simple gesture."* In contrast, the remaining five participants spent more time listening to the music and observing the environment. P7 explained, *"I find adjusting the audio-visual effects through interaction to be a very novel experience; however,*

when I want to focus on simply enjoying the music, it may be a bit distracting." These individual differences highlight the need for the system to offer customizable interaction levels to better accommodate varying user needs and preferences.

7 DISCUSSION

Drawing on the findings, we revisit the research questions and highlight the challenges and opportunities for designing adaptive, user-centered musical environments.

7.1 Design Implications

In the context of **co-creation**, users exhibited interests in collaborate with AI to create visual effects aligned with the music context, showed general acceptance with the AI-generated contents, and valued the novelty and creativity inspired by AI, reflecting an emerging mindset in which users increasingly embrace AI as a partner in creative expression. Notably, when it comes to subtle emotional expression, AI-generated content does not always align with users' expectations. AI is thus most effective when used as supportive tool, preserving user's creative intent. This addresses **RQ1: How to facilitate engaging AI-user co-creation?** In the current version, the co-creation of scene textures and 3D models is limited—users have minimal interaction with these visual elements during music playback. Supporting **real-time modification** would enhance user agency and engagement, allowing them to continuously shape the visual atmosphere in sync with the music.

Takeaway: AI is most effective in co-creation when it supports and amplifies user intent. Future systems should prioritize real-time, fine-grained control so that AI functions not as an autonomous creator, but as a responsive partner in shaping immersive, music-driven experiences.

For the second question—**RQ2: Which elements and spatial scales best enhance user experience in a musical environment?**—we found that users require a combination of sensory elements to shape their experience, including sound effects, dynamic visual effects, and immersive environmental design, presented in a cohesive and structured manner that seamlessly aligns with the music's progression and emotional trajectory. Our prototype presents textures and models that align with the musical content, complemented by rhythm-synchronized and customizable particle effects. However, the presentation still requires refinement, as models currently appear abruptly in the Immersive Room. While the Stage Box introduces them more deliberately using portals, the aesthetic may not be suitable for all musical genres. Exploring more subtle **transition techniques** could enable a more context-sensitive and harmonious integration of visual elements.

In terms of **spatial scale**, each configuration we explored offers distinct experiences. The handheld **Mini Box** supports continuous awareness of the physical world and potentially extending to physical instruments for tactile engagement. The **Stage Box** delivers a semi-immersive experience, blending physical and virtual elements. By contrast, the VR **Immersive Room** isolates users, providing a heightened sense of immersion. User feedback highlights diverse preferences: some prioritize tactile realism and physical grounding, while others seek full sensory immersion in VR,indicating that a single approach cannot meet all needs. An adaptive environmental design capable of **seamlessly switching between AR and VR** could offer a more inclusive experience.

Takeaway: User experience in musical environments depends on a seamless combination of sensory elements and flexible spatial scales. Since preferences range from tactile realism to full sensory immersion, adaptive designs that seamlessly transition between AR and VR can better accommodate diverse needs.

For the third question—**What forms of music interactions and responses do users expect?**—our study showed that users expect the musical environment to respond to their actions in real time, for example by allowing visual elements to move or change style in response to gestures. They also benefited from integrating personalized sounds, such as humming or beating, into the creative process, fostering a live performance-like atmosphere. Our prototype allows users to layer auditory elements onto the musical track and utilizes colorful sticks as physical controllers to manipulate visual effects. By offering a playful experience, it encourages user interaction and supports the creation of rich, dynamic outputs that enhance the overall audiovisual experience. These interaction paradigms could be further extended—for instance, by incorporating full-body kinesthetic interactions—to provide additional opportunities for creative exploration in music-driven environments.

Notably, users revealed a spectrum of engagement levels, suggesting that interaction designs of musical environments must carefully balance **stimulation** with **attentional demand**. Excessive interaction complexity or sensory input risks overwhelming or distracting users who seek a contemplative experience, while limited interaction options may frustrate those who desire frequent engagement and great agency. **Flexible interaction modes** should be offered, allowing users to fluidly shift between passive listening and active participation according to moment-to-moment preferences.

Takeaway: Users expect musical environments to respond to their actions in real time, blending gesture-driven visual changes with the integration of personalized sounds to create a live performance-like atmosphere. Playful, embodied interactions, such as physical controllers or future full-body engagement, enhance creativity and foster richer, more dynamic audiovisual experiences

7.2 Limitations and Future Directions

We acknowledge certain limitations in the current experimental design. Our evaluation of user experience was confined to questionnaires and interviews. Future work could incorporate quantitative measures, such as interaction accuracy or task completion time, to gain additional insights into system usability. Moreover, we did not investigate potential variations across different demographics, such as musical background, gender, or age. Exploring these dimensions could provide more nuanced insights into user engagement and system effectiveness, ultimately informing design improvements and enhancing inclusiveness across diverse user groups.

Looking ahead, several promising directions warrant further exploration. One avenue is extending the designs to support multi-user environments, enabling collaborative music creation and real-time interaction among multiple participants, which could provide insights into social dynamics, engagement, and collective creativity. Another important area is the study of AI-generated content (AIGC) and its effects on user engagement. Systematic comparisons between AI-generated and manually designed musical environments could help

identify the strengths and address the limitations of AI-supported co-creation. Future research could also explore adaptive designs that tailor the complexity, feedback, and control options to match users' skill levels and learning trajectories. Together, these directions can contribute to a deeper understanding of interactive music systems and the potential for AI to augment creativity and user experience.

8 CONCLUSION

We investigate how AI-driven co-creation, immersive environments, and multimodal interactions shape user experiences with music, revealing diverse preferences in interaction styles, spatial embodiment, and levels of immersion. While AI offers valuable support in facilitating the expression of thoughts and emotions, it should not replace the process of transforming imagination into artistic reality. Instead of a provider of finished works, AI is better positioned as a tool that enriches and personalizes the creative journey. The same principle applies to the design of visual and environmental elements, as well as interactions: systems should empower individuals to define their own level of engagement within the co-creation experience. As the saying goes, "There are a thousand Hamlets in a thousand people's eyes." Such customization acknowledges the multiplicity of user interpretations, allowing diverse forms of creativity and engagement to coexist within a musical environment.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 62272396).

REFERENCES

[1] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. MusicLM: Generating Music From Text. arXiv:2301.11325 doi:10/grp69x

[2] Ye Bai, Haonan Chen, Jitong Chen, et al. 2024. Seed-music: A unified framework for high quality and controlled music generation. arXiv:2409.09214 doi:10/p243

[3] Esben Oxholm Bonde, Ellen Kathrine Hansen, and Georgios Triantafyllidis. 2018. Auditory and Visual based Intelligent Lighting Design for Music Concerts. *Eai Endrosed Trasactions on Creative Technologies* 5, 15 (4 2018). doi:10/p3xw

[4] Jennifer L. Breese, Mark A. Fox, and Ganesh Vaidyanathan. 2020. Live Music Performances and the Internet of Things. *Issues in Information Systems* 23, 3 (2020), 179–188.

[5] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, Kathryn Huff and James Bergstra (Eds.). 18 – 24. doi:10/gf4wxc

[6] Jean-Pierre Briot and François Pachet. 2020. Deep learning for music generation: challenges and directions. *Neural Computing and Applications* 32, 4 (Feb. 2020), 981–993. doi:10/ggtmj6

[7] Melissa Dobson. 2010. *Between stalls, stage and score: an investigation of audience experience and enjoyment in classical music performance.* Ph. D. Dissertation. University of Sheffield.

[8] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. MuseGAN: multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *AAAI Conference on Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, New Orleans, Louisiana, USA, Article 5, 8 pages. doi:10/p242

[9] Jules Françoise. 2013. Gesture–sound mapping by demonstration in interactive music systems. In *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona, Spain) *(MM '13)*. Association for Computing Machinery, New York, NY, USA, 1051–1054. doi:10/gmvcv5

[10] George Charles Higgs. 2018. *The Sense Ensemble: An Approach to Music Composition for Deaf and Hearing Alike*. Ph. D. Dissertation. Trinity College Dublin.

[11] Oliver Hödl, Geraldine Fitzpatrick, Fares Kayali, and Simon Holland. 2017. Design Implications for Technology-Mediated Audience Participation in Live Music. In *14th Sound and Music Computing Conference SMC 2017 (Proceedings of the 14th Sound and Music Computing Conference 2017)*. 28–34. http://eprints.cs.univie.ac.at/5451/

[12] Jaeyong Kang, Soujanya Poria, and Dorien Herremans. 2024. Video2Music: Suitable music generation from videos using an Affective Multimodal Transformer model. *Expert Systems with Applications* 249 (2024), 123640. doi:10/p244

[13] Yewon Kim, Sung-Ju Lee, and Chris Donahue. 2025. Amuse: Human-AI Collaborative Songwriting with Multimodal Inspirations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 187, 28 pages. doi:10/g9hpt2

[14] Sang Won Lee. 2012. *Audience participation using mobile phones as musical instruments*. Ph. D. Dissertation. Georgia Institute of Technology.

[15] Jorge Lensing. 2023. The future of audio-visual designers with a focus on sound. *International Journal of Film and Media Arts* 8, 3 (2023), 38–58. doi:10/p3xx

[16] Yangjingyi Li and Fenghsu Lee. 2023. Sound effect of transmitting specific emotions in music performance based on the acoustic properties of conjugated materials. *Frontiers in Chemistry* Volume 11 - 2023 (2023). doi:10/p248

[17] Linsong, Li. 2023. The influence of concert hall architecture on acoustics (the effect of musical performance). *E3S Web of Conf.* 389 (2023), 06016. doi:10/p249

[18] Hao Liu. 2024. Contemporary Stage Lighting Art: The Multi-Element Composition of Post-Drama Stage Space. *American Journal of Arts and Human Science* 3, 3 (Aug. 2024), 211–217. doi:10/p246

[19] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10/gjzbww

[20] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50 – 60. doi:10/fgp7wb

[21] Hafizhah Insani Midyanti and Yudi Sukmayadi. 2021. The intersection of local wisdom and digital innovation in performing arts. *Interlude: Indonesian Journal of Music Research, Development, and Technology* 1, 1 (2021), 43–60. doi:10/p3xz

[22] Aya Mohamed Kamal ElDean Amean. 2019. The interaction between materials technology and theatrical scenography. *International Journal of Multidisciplinary Studies in Art and Technology* 2, 2 (2019), 29–47.

[23] Kouyou Otsu, Jinglong Yuan, Hisato Fukuda, Yoshinori Kobayashi, Yoshinori Kuno, and Keiichi Yamazaki. 2021. Enhancing Multimodal Interaction Between Performers and Audience Members During Live Music Performances. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 410, 6 pages. doi:10/p247

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. 10684–10695 pages.

[25] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (01 2017), 103. doi:10/gf4zsf

[26] Michael Schutz. 2008. Seeing music? What musicians need to know about vision. *Empirical Musicology Review* 3, 3 (2008), 83–108. doi:10/gddxt7

[27] S. S. Shaprio and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3-4 (12 1965), 591–611. doi:10/dkpp9q

[28] Alina Striner and Brenna McNally. 2017. Transitioning Between Audience and Performer: Co-Designing Interactive Music Performances with Children. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2115–2122. doi:10/gkh2s2

[29] Dana Swarbrick and Jonna Katariina Vuoskoski. 2023. Collectively Classical: Connectedness, Awe, Feeling Moved, and Motion at a Live and Livestreamed Concert. *Music & Science* 6 (2023), 20592043231207595. doi:10/p245

[30] Richard Vogl, Hamid Eghbal-Zadeh, Gerhard Widmer, and Peter Knees. 2018. GANs and Poses: An Interactive Generative Music Installation Controlled by Dance Moves. *Interactive Machine-Learning for Music@ Exhibition at ISMIR. Paris, France* (2018), 1–5.

[31] KatieAnna E. Wolf and Rebecca Fiebrink. 2019. Personalised interactive sonification of musical performance data. *Journal on Multimodal User Interfaces* 13, 3 (01 Sep 2019), 245–265. doi:10/gmvcwg

[32] Yongmeng Wu. 2018. *Design and Evaluate Support for Non-musicians' Creative Engagement with Musical Interfaces*. Ph. D. Dissertation. Queen Mary University of London.

[33] Wanfang Xu, Lixiang Zhao, Haiwen Song, Xinheng Song, Zhaolin Lu, Yu Liu, Min Chen, Eng Gee Lim, and Lingyun Yu. 2025. Mozualization: Crafting Music and Visual Representation with Multimodal AI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 407, 7 pages. doi:10/p24z

[34] Akmaljon Yuldashev. 2024. Music in Virtual Reality: New Opportunities for Composers and Performers. *Jurnal Pendidikan Non formal* 1, 3 (2024), 8–8. doi:10/p3x2