

CrowdForge: Effects of Audio and Visual Audience Participation Modalities on Agency, Collaboration, and Engagement in AI-Mediated MR Performances

Torin Hopkins
SolJAMM Research and ATLAS Institute
University of Colorado, Boulder
Boulder, USA
TorinHopkins@gmail.com

Shih-Yu Ma
University of Colorado, Boulder
Boulder, USA
Shih-Yu.Ma@colorado.edu
0000-0001-6686-0079

Zander Trafton-Hancock
ATLAS Institute
University of Colorado, Boulder
Boulder, USA
zatr2033@colorado.edu

Ellen Yi-Luen Do
ATLAS Institute
University of Colorado, Boulder
Boulder, USA
ellen.do@colorado.edu

Luca Turchet
Dep. of Information Engineering
and Computer Science
University of Trento
Trento, Italy
luca.turchet@unitn.it

Abstract—Technology-mediated audience participation (TMAP) enables audiences to influence live performances, yet balancing interaction with meaningful agency remains challenging. While AI-generated music and MR enable new forms of audio and visual audience participation, little is known about how these participation modalities influence audience experiences during live musical performance. In particular, it remains unclear how they shape perceived agency, collaboration, and engagement, and how they complement one another within TMAP systems. We present CrowdForge, an AI- and MR-mediated paradigm where audiences collectively control generative music and virtual environments during a live performance featuring a guitarist and DJ. Through a mixed-methods concert study evaluating three interaction conditions (audio, visual, and combined audio-visual), we demonstrate that observability is the primary driver of audience agency. While visual interaction provided easily observable feedback, audio parameters offered macro-structural control over the music. Audiences tended to use the control enabled by AI audio generation to test musical boundaries, introducing a productive adversarial tension between the crowd and the musicians. This dynamic pushed creative limits, deepened engagement, and generated moments of musical coherence when performers successfully resolved audience-induced challenges. We argue future TMAP systems should prioritize human-centered, highly observable interactions to build agency, while combining the expressiveness of music with the clarity of visual feedback.

Index Terms—Technology-mediated audience participation, mixed reality, Musical Metaverse, Musical XR

I. INTRODUCTION

Traditional live music performances are predominantly unidirectional experiences, with minimal interaction between the audience and performers [1]. Participatory music experiences

shift this focus and prioritize aspects such as collaborative creativity and community building [2], [3]. Though participatory music experiences are ubiquitous across cultures, and come in many different forms [1], [4], [5], in recent years, the role of technology in facilitating audience participation has become an area of academic and artistic interest [2], [6], [7]. These interdisciplinary artistic works ask us to reconceptualize the traditional paradigm, shifting the experience of music performance towards an exploratory and communal effort. The field of technology-mediated audience participation (TMAP) in live music performance [8] simultaneously celebrates the artistry of trained live musicians and the use of technology as a means of communicating the intentions and sentiments of an audience, thus contributing to the creative process.

Mozart famously utilized musical dice to promote audience interaction, having spectators roll dice to rearrange parts of the composition. As technologies became more sophisticated and interconnected [9], [10], engineers, designers, and artists alike have contributed to this growing field by utilizing a myriad of technological tools [6], [11]–[14]. With the growing prevalence of networking, reliance on visual media (e.g., mixed reality, interactive visualizations), and complex engineered systems, TMAPs have begun to touch on many questions central to the musical metaverse [15], [16].

Recently, in [8] Hodl and Turchet outline several dimensions of technologies that circumscribe TMAPs, including light [17], [18], physical devices [19], sensor-based devices [5], [17], [20], and smartphone/tablets [3], [14], [21]. This effort illustrated a comprehensive design space that connects the choice for various technological implementations, as well as the motivations for design decisions given the various technologies and aims of performance.

From a complementary perspective, Wu et al. [2] identi-

This work has been supported by the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379).



Fig. 1. Overview of the audience participation workflow in CrowdForge. (A) Audience selections are aggregated into a collective configuration spanning musical character, tempo, instrumentation, dancer movement, virtual performers, and stage environment. (B) The selected parameters are incorporated into a live musical performance involving human musicians and projected audiovisual content. (C) The visual selections simultaneously configure a mixed-reality performance environment populated by virtual musicians and dancers. Together, these audio and visual modalities allow audience input to become a perceivable component of the resulting performance.

fied four key dimensions of TMAPs: agency (the perceived ability to influence the performance), participation modality (how audiences interact), degree of networking (the role of networked technologies in supporting participation), and scalability (the capacity to accommodate larger audiences). Based on feedback from both audience members and performers, agency, participation modality, and interface design emerged as the most salient themes. Conversely, despite their presence in the TMAP framework proposed in [8], little is known about how participation modality (e.g., interactive audio and visual elements) impacts the sense of agency, collaboration, and engagement.

Additionally, many TMAP systems utilize visualization, sampling, and data points to impart change on musical and visual parameters, either providing information to performers (e.g., Open Symphony [2]), or to inform musical content through sampling (e.g., Pharosphones [22]). Some have even leveraged the variability in output potential of AI-generated music [19], [23] and Mixed Reality (MR) [24]. However, to the best of our knowledge, there has yet to be a systematic presentation of how various creative and interactive modalities impact both audio/visual and perceptual outcomes.

Furthermore, there is an opportunity to explore how different participation modalities enabled by AI-generated music [23], [25]–[30] and MR [31], [32] influence audience experiences during live performance. In particular, we seek to understand how audio and visual participation modalities shape perceived agency, engagement, and social collaboration, and how they complement one another in supporting meaningful audience participation. This addresses a well-known challenge in designing TMAP systems: balancing accessible interaction (i.e., interfaces that are easy to understand and use) with opportunities for audiences to make meaningful creative contributions [2], [11], [33]. AI-generated music and MR are particularly well suited to enable such complementary participation modalities, allowing audiences to influence both musical and visual aspects of a live performance through intuitive, high-level interactions.

In this paper we present *CrowdForge*, an AI and MR-

mediated audience participation paradigm for live musical performance, as well as an accompanying experiment. We specifically aim to answer two research questions:

- 1) How do different audio and visual participation modalities impact an audience’s sense of agency, engagement, and collaboration in a live music performance?
- 2) How can the design of TMAP systems optimize the trade-off between making collaborative elements accessible to a diverse audience while ensuring they can make meaningful creative contributions?

To answer such questions, we conducted a multi-condition, live performance featuring a DJ, a guitarist, an MR demonstrator in VR, and 19 interactive audience members. Following the performance ¹, a mixed-methods evaluation was conducted.

Our study makes the following contributions:

- 1) CrowdForge: A networked, AI-mediated audience participation performance paradigm in MR;
- 2) A mixed-methods exploration of agency, collaboration, and engagement with respect to interactive modality (generated audio and MR visuals);
- 3) Insights on how to promote a sense of agency, collaboration, and engagement in AI and MR-based TMAPs for live music performance.

II. SYSTEM DESIGN

The CrowdForge performance system consisted of four main components: 1) the web interface for aggregating audience preferences, 2) AI model for music generation, 3) Ableton Live 12 for DJ and AI control, and 4) a projected MR environment (see Figures 1 and 2).

In-person audience members connect to the server via their smartphones and vote on parameters of the performance at the beginning of each set (roughly every 7 minutes). Depending on the condition, participants voted on either audio parameters, visual parameters, or both. Audio parameters included aspects

¹A recording of the performance is available: https://www.youtube.com/watch?v=nbGJML_i9CU

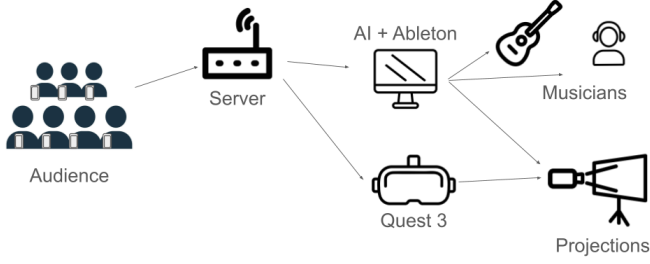


Fig. 2. System diagram and signal flow illustrating the control audience members had over elements of the performance (from left to right). Audience members visit a web page on their mobile devices and vote on parameters they want to see reflected in the performance. The inputs are then aggregated and sent to both a generative transformer model (played through Ableton Live 12 software) and a mixed reality engine (Unity 3D) for rendering both audio and visuals.

that impact the generated AI music (such as genre and tempo) and other audio parameters (e.g., guitar tone of a live human guitar player). Visual parameters impacted the projected MR visual scene, such as virtual characters and the virtual environment (see Table I). On stage a guitar player, a DJ, and an MR demonstrator wearing a Quest 3 headset were featured. Though future performances can include audience members wearing headsets or other MR devices, the test performance only included an MR demonstration of a single perspective with a projected visual scene (see Figures 1 and 2).

A. Web Interface Architecture and User Interaction

The web interface serves as the primary interactive conduit between the live audience and the generative AI performance system. From the audience perspective, users enter the ecosystem through an onboarding portal where they manipulate sliders and dropdown menus to cast votes on musical, visual, or both musical and visual parameters (depending on condition).

The finalized data structure is then visualized on a public-facing aggregator dashboard. The server then simultaneously broadcasts to the Ableton Live API for set control (via AbletonOSC [34]), the causal model for multitrack MIDI generation, and the Unity Engine for 3D visual rendering (see Figures 3 and 2).

B. AI Model for Music Generation

1) *AI Model and Tokenization Scheme*: The system employs Llama-Z, a domain-adapted iteration of the Llama-3.2-1B causal language model. The REMI-z tokenization scheme [35] was utilized for its ability to generate multiple cohesive instrument parts in a fully orchestrated track while enabling audience members to omit an instrument and retain coherence in the generated output. Data representation relies on a specialized variation of the REMI-z tokenization protocol, namely the inclusion of genre.

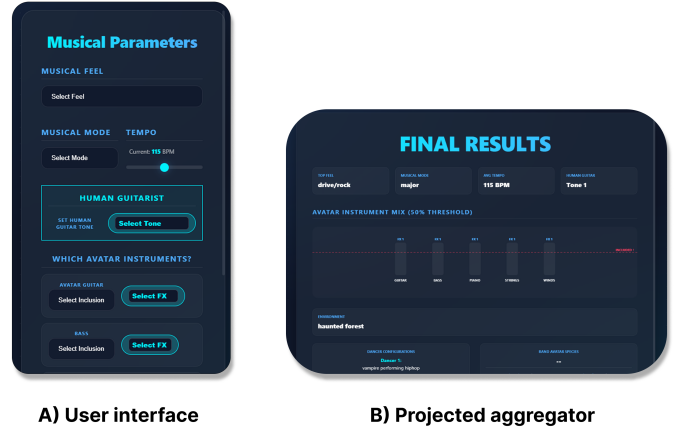


Fig. 3. A) Mobile user interface for audience member during the audio parameter selection condition. B) Sample projected aggregator screen displayed to audience after selections were made.

2) *Pre-processing and Mathematical Normalization*: To reduce the complexity of the data distribution and adhere to the REMI-z tokenization scheme, all incoming multitrack corpora—comprising 20,000 sampled segments from the Los Angeles MIDI Dataset (LAMM) and the Slakh2100 dataset—are normalized by discarding any segments not adhering to a 4/4 time signature.

All harmonic content is mathematically shifted to a C-Major or A-Minor ground truth. Key detection is performed using the Krumhansl-Schmuckler (K-S) correlation algorithm, comparing an extracted 12-tone pitch class distribution histogram.

C. System Integration Pipeline:

Web Interface → *AI* → *Ableton Live*

The inference architecture utilizes a multi-threaded pipeline translating web-based audience requests into arranged multitrack compositions inside Ableton Live.

1) *Autoregressive Generation*: The generation sequence initiates when the web interface transmits an OSC message to the model to initiate generation given the input selected by the audience. The message specifically contains the target genre and the instrumentation (e.g., Drums, Piano, Guitar).

2) *The Ableton Controller Integration*: The controller script acts as a stateful agent directly manipulating Ableton Live’s API, affecting parameters such as tempo and musical mode, as well as copying the content of locally generated MIDI files as launchable tracks in Ableton Live’s session mode. This enables the DJ to manually manipulate the tracks as they are generated, affording control in a manner similar to pre-arranged tracks during live performance.

D. Mixed Reality Viewer

The MR viewer was developed in Unity 3D and deployed on Meta Quest 3 headsets as the primary visual-rendering component of the CrowdForge performance system. The viewer receives near-real-time control messages from the aggregation server and updates the visual scene according to the audience-selected parameters. These parameters include both character

appearances and environmental themes, allowing the projected MR environment to reflect the collective preferences gathered through the web interface.

The MR scene contains a virtual ensemble consisting of two dancers, two drummers, one keyboardist, one violinist, one flutist, one guitarist, and one bassist (see Figure 1C). Each instrumental character is configured to respond to OSC messages transmitted from the Ableton Live controller. These messages trigger synchronized performance animations, such as striking the appropriate drum pad, pressing keys within the correct keyboard region, or performing the corresponding instrumental gesture. The dancers are controlled through a combination of style commands from the aggregation server and timing messages derived from Ableton Live playback, enabling them to perform choreographic motions that remain synchronized with the music.

The environmental layer of the MR viewer places the full ensemble on a virtual stage. Similar to the character appearances, the active environment is updated through OSC messages from the aggregation server, allowing the visual scene to transition in response to audience voting results.

During the live performance, one Meta Quest 3 headset was used to project an example view point in VR. The demonstrators view point, gaze direction, and movements were cast to the projection display—enabling the audience to view the virtual space and the performance space simultaneously. Though viewing the VR environment could have been accomplished without a human demonstrator, the demonstrator provided the audience with a dynamic and human perspective within an immersive environment. A full view of the MR performance environment is shown in Figure 1.

III. STUDY DESIGN AND PERFORMANCE

The study consisted of three music sets (roughly 7 minutes each) per performance differentiated by the selections the audience could make: 1) **audio only**, 2) **visual parameters only**, and 3) **both audio and visual parameters**. Audience members voted on performance parameters via web form before the beginning of each set. Two performances consisting of three conditions/sets occurred over the course of one day (N=19 audience members in total, N=5 for performance 1 and N=14 for performance 2). The first performance began with audio and the second with the visual condition.

The study was approved by the University of Colorado Institutional Review Board and took place in a media studio, akin to a black box theater. The session began with a brief explanation of how to participate and all participants logged into the local area network to maintain connection throughout the performance. It was ensured that all participants were able to connect to the network before beginning the performance. Participants faced the performers and were instructed to make selections at the beginning of each condition/set.

There were two live performers: 1) a guitar player, 2) a DJ, and an MR demonstrator wearing a Quest 3 headset and projecting their live point of view of the virtual stage to the audience. Behind the performers was the projection of the

virtual environment and a projection of the aggregator screen (see Figure 1B). While the audience made selections (see Table I) the aggregator screen showed user-chosen names for the audience contributors, enabling a dimension of self expression and a visual check to ensure all members of the audience remained connected to the network.

After the audience members finished making their selections individually on their mobile devices the music generation began. This initiated the aggregation screen, which was immediately made visible to the audience, enabling them to see the results of the collective voting (see Figure 3). Within approximately 10 seconds the DJ began observing and playing clips which were generated (based on audience selections) and programatically populated in the Ableton Live set. The guitar player improvised with the DJ-selected generated tracks and each set lasted approximately 7 minutes.

After the three sets concluded, the audience members were instructed to complete an online questionnaire. Following the questionnaire, six audience members total from the two performances volunteered to stay for a brief semi-structured audio-recorded interview (10-20 minutes).

IV. AUDIENCE MEMBER EVALUATION

Our evaluation is based on feedback from 19 total audience members ($N = 19$; 17 men, 2 women; $M_{age} = 28.89$, $SD = 9.89$) and the two musicians taking part in the performance. Audience member feedback was assessed through questionnaire (N=14, 5 opted out), behavior as recorded in the application (N=17, two experienced connection issues), and semi-structured interviews (N=6, of the 19 volunteered to stay for interviews). The performers provided self-reflective accounts of the performance experience.

A. Behavioral Analysis

Skip rates provided behavioral validation of participants' engagement, aligning closely with the subjective ratings. When isolated in single-modality conditions, participants heavily favored the active domain. In the Audio condition, participants actively engaged with audio elements—showing a 0% skip rate for tempo and a 6.7% skip rate for genre and human guitar tone—while almost entirely ignoring the inactive visual configurations (93.3% skip rate for characters, moves, and environments). In the Visual condition, participants engaged with visual settings (only an 11.1% skip rate for characters and moves). In the “Both” condition where all options were available, engagement balanced out but prioritized perceived high-impact features. Participants interacted heavily with core structural elements across both domains: tempo (0% skip rate), visual characters and moves (7.7% to 15.4%), and genre/environment (15.4%). More granular audio configurations—such as specific effects choices for AI instruments—saw much higher skip rates ranging from 30.8% to 38.5%. This behavioral data mirrors the subjective Influence and Agency findings, which suggest elements that participants felt gave them high influence and control (like tempo, genre, and human musician instrument tone) were

Condition (No. of Parameters)	Available Options
Audio (14) - Genre / Feel - Tempo - Mode - Human Guitar Tone - Instruments - Piano FX - Guitar FX - Bass FX - Strings FX - Winds FX	Genre / Feel: drive/rock, pulse/hiphop, groove/funky, elegant/atmospheric Tempo: 90–140 BPM Mode: major, minor Human Guitar Tone: clean, distorted, atmospheric Instruments: guitar, bass, piano, strings, winds Piano FX: default, dist, echo Guitar FX: default, dist, airy Bass FX: default, dist, reverb Strings FX: default, dist, chorus Winds FX: default, dist, phaser
Visual (14) - Environment - Character Skins for 9 characters - Dance Moves for 2 dancers	Environment: haunted forest, mars scape, laser grid Character Skins: vampire, werewolf, moose orc, goblin elf, flesh eater, dark noire, lizard species, webbed and armored, x-bot, y-bot Dance Moves: hiphop, breakdance, latin, party
Both (28) - All Audio Parameters - All Visual Parameters	Description: The “both” mode combines all available parameters from the audio and visual conditions.

TABLE I
PARAMETERS AND AVAILABLE OPTIONS PER APP CONDITION

rarely skipped, whereas the AI details that scored poorly on influence metrics were frequently bypassed even in a fully multimodal environment (see Figure 6).

B. Exploratory Questionnaire Results

To understand how modality impacted the interactive audience experience, we analyzed participant responses ($N = 14$) across four primary constructs: Sense of Agency, Social Presence, Emotional Response via the Self-Assessment Manikin (SAM), and the perceived influence of audio parameters.

To evaluate differences across the experimental conditions (Audio, Visual, Both) and elements, non-parametric statistical tests were employed due to the ordinal nature of the data and distribution assumptions. A Friedman Omnibus Test was first conducted to identify if there were overall significant differences among the related groups. When the Friedman test indicated significance (or near significance), Wilcoxon signed-rank tests were used for pairwise post-hoc comparisons. To account for the family-wise error rate, standard Bonferroni corrections were applied when evaluating the three main conditions (3 comparisons), while a step-down Holm-Bonferroni correction was applied when evaluating the audio sub-elements to preserve statistical power across the larger set of tests (15 pairwise comparisons among 6 elements). Additionally, one-sample Wilcoxon signed-rank tests were used to evaluate whether absolute scores in each condition significantly deviated from a neutral baseline (e.g., 4 on a 1–7 Likert scale; 5 on a 1–9 SAM scale). Effect sizes for all Wilcoxon tests are reported as r , where values of 0.1, 0.3, and 0.5 indicate small, medium, and large effects, respectively.

1) *Sense of Agency*: To evaluate sense of agency, we analyzed responses from the Agency questionnaire alongside Dominance (i.e., perceived control) scores from the SAM. For the Agency measure, no significant overall differences were observed across the modalities (Friedman $p = 0.721$).

However, the SAM Dominance scores—which measure the emotional feeling of being in control of the situation—demonstrated significant impact. While the omnibus test for Dominance did not show strict overall significance across conditions (Friedman $p = 0.368$), and pairwise differences were marginal (Audio vs. Visual $p = 0.075, r = 0.48$), comparisons against the neutral baseline revealed the Audio condition ($p = 0.006, r = 0.73$) and the Both condition ($p = 0.025, r = 0.60$) significantly elevated feelings of Dominance above the neutral baseline with large effect sizes (see Figure 4). In contrast, the Visual-only condition did not elevate Dominance beyond neutral ($p = 0.612$). This indicates that while explicit, cognitive ratings of agency remained stable across modalities, the underlying affective sense of dominance and control was significantly driven by the presence of audio.

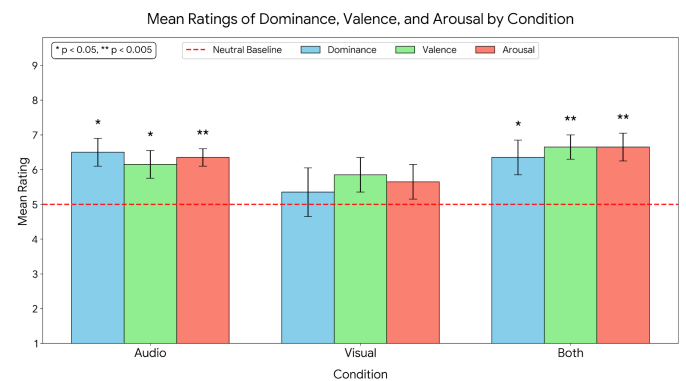


Fig. 4. Indicates dominance, valence, and arousal scores on the self-assessment manikin, comparing Audio, Visual, and Both conditions. Each measure is rated 1-9. Significance figures are demonstrated with respect to the neutral baseline of 5 (shown in red).

2) *Sense of Social Presence Between Audience Members*: The experience of Social Presence—the feeling of being connected or co-present with the other audience mem-

bers—showed a marginal overall effect across the three conditions (Friedman $p = 0.092$). Post-hoc pairwise testing uncovered a significant difference specifically between the Visual and Both conditions ($p = 0.015$), driven by a large effect size ($r = 0.65$). This suggests that adding audio elements to a visual experience significantly enhances the feeling of social presence compared to experiencing visual elements alone (see Figure 5).

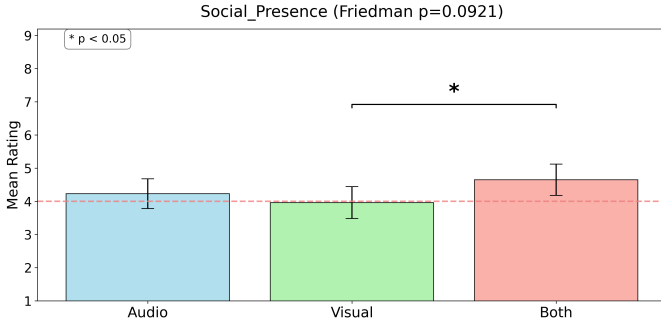


Fig. 5. Pairwise testing indicates a significant difference between the Visual and Both conditions ($p = 0.015$). Red line indicates neutral score.

3) Measures of Valence and Arousal Across Conditions:

Conditions containing audio successfully elevated emotional engagement above the neutral baseline. Both the Audio and “Both” conditions significantly increased Valence (Audio: $p = 0.021, r = 0.62$; Both: $p = 0.003, r = 0.80$) and Arousal (Audio: $p = 0.002, r = 0.84$; Both: $p = 0.004, r = 0.76$) with large effect sizes (Figure 4). The Visual-only condition did not significantly exceed the baseline for either metric ($p > 0.10$). However, direct comparisons between the Audio and Visual-only conditions were not statistically significant ($p > 0.32$). Thus, while audio elevates affective states above a neutral experience, it is not strictly superior to visual elements alone.

4) *Perceived Influence and Agency of Specific Audio Components:* To further investigate the specific impact of sound, various audio elements were evaluated for their individual contributions to feelings of Agency and overall Influence (see Figure 6). When assessing which specific audio features gave participants a sense of agency, “Guitar tone for human player” stood out as the only element rated significantly above the neutral baseline ($p = 0.007$, large effect size $r = 0.72$). Other elements, such as musical mode, tempo, and choosing AI player effects, did not significantly exceed the baseline for agency (see Figure 6). However, when evaluating broader influence on the experience, participants rated several core musical elements as significantly impactful compared to baseline. These included Guitar tone ($p = 0.002, r = 0.85$), Musical feel/genrre ($p = 0.002, r = 0.82$), Musical mode ($p = 0.020, r = 0.62$), and Choosing to include AI players ($p = 0.032, r = 0.57$). An omnibus test confirmed an overall significant difference among the influence ratings of these specific elements (Friedman $p = 0.016$). Uncorrected pairwise p-values highlighted strong trends for core musical features over granular settings; for instance, customizing the Guitar tone

was rated much more influential than choosing specific sound effects for generated instrument parts ($p = 0.009, r = 0.70$).

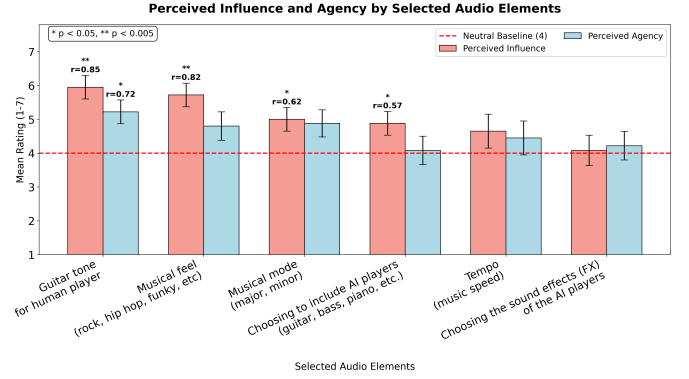


Fig. 6. Indicates perceived influence and agency on the performance when selecting specific audio elements. Both measures use a rating scale of 1-7. Significance figures and effect sizes (r) are demonstrated with respect to the neutral baseline of 4 (shown in red).

C. Thematic Analysis

Thematic analysis is based on free response questions and interview data ascertained post-performance. Participants discussed agency, sense of social presence with other audience members, and commented on how to improve the performance paradigm. Both free responses and interview data were used for thematic analysis.

We recorded, transcribed the interviews, and isolated responses from the 14 participants. Transcriptions and responses were then coded using inductive thematic analysis, rooted in grounded theory [25], [36], [37]. The first and second authors independently coded and conducted the analysis of all 14 participants. We then discussed the topics considerably and converged on the most present topics in the dataset [26], [38]. MAXQDA [39] software was used to keep track of independent coding databases which facilitated discussion. The coding databases were merged after discussion to produce the following final themes.

1) Sense of Agency:

a) *Subtheme: Audio Parameters Require Macro-Adjustments:* The audience’s sense of agency was heavily hindered by the difficulty of tracking individual changes within the aggregated soundscape. Participants observed that while the “big, general elements felt responsive,” more granular audio settings “weren’t present enough in the music for me to tell” (P13), making it “hard to track every parameter” (P4). This lack of clarity was compounded by the fact that aggregating disparate audio choices often led to conflicting results; for instance, combining an “ethereal or atmospheric” vibe with incompatible instrument tones resulted in what P11 described as a “jumbled musical idea”. To reclaim a sense of control, users expressed a desire for deeper musical intervention, wishing for the ability to “control chord progressions, drum parts and other arrangement pieces” (P3) or to dynamically “adjust all the parameters live” (P6).

b) Subtheme: Visual Prominence and Engagement:

When analyzing the sensory feedback of the performance, the visual components often had a more immediate and noticeable impact on the audience than the audio manipulations. Participants frequently commented on the striking nature of the visual changes and how the crowd reacted to them. One participant observed, “I feel audience is more affected by the visual change, (P1)” concluding that “the visual is more controllable and noticable”. The MR components and character animations provided clarity and evidence that the audience’s choices were reflected.

c) Subtheme: Boundary Testing and Extreme Audio Selections: A prominent sub-theme regarding agency was the audience’s tendency to test the limits of the AI-mediated system rather than striving for traditional musicality. This accomplished two main aims: 1) testing the limits of the system, and 2) choosing extreme selections to maximize individual influence over the performance. One participant indicated in their survey response, “I ended up picking extreme options to show much influence I had. (P5)” This sentiment was echoed strongly during the interviews, with one creator noting that instead of making choices that “would make the most musical sense,” the audience’s mentality was often to “take this to the max...with the scariest sound possible. (P2)” This behavior highlights a playful, almost adversarial relationship between the user and the system, where the goal was not necessarily to create a cohesive song, but to understand the boundaries of the technology and attempt to exert personal agency. The sheer novelty of the combinatorial space encouraged this experimental approach, allowing users to stretch the system until they wondered, “when’s it gonna fall apart? (P1)”

2) Social Dynamics and Connection:

a) Subtheme: Emergent Collective Collaboration and Engagement: Voting on preferences before the beginning of a set was noted as an unoptimized method of garnering social collaboration. As one participant stated, “it was cool that we were able to vote. But it didn’t particularly feel like it like bonded me with other people in the audience. (P6)” Furthermore, participants suggested improvements to make this mechanic feel more dynamic and responsive, such as implementing a live voting window where users “can see how many people are voting each option in real time (P3)” to increase active engagement. While voting successfully facilitated interaction, the resulting sense of agency was highly dependent on whether a participant’s individual choice aligned with the majority and affected human performers. One survey respondent captured this sentiment, stating, “I felt we could all create a unique performance, made me more attentive. (P14)” The realization that the performer “received input from the audience and the opinion of the audience actually had an effect on the performance (P3)” fostered a sense of shared responsibility for the outcome.

V. THEMES IN MUSICIAN EVALUATION

Musicians provided a self-reflective written statement pertaining to their experience throughout the two performances.

The musicians included an improvising guitar player who played alongside the AI generated music, while the DJ was arranging and playing AI-generated clips as that were programmatically populated in Ableton. Both were practicing musicians with over 10 years of experience. The guitar player also has experience in improvised performances including technology-mediated audience participation.

Throughout the performance there was non-verbal communication between the two musicians, particularly with regard to ending each song (roughly every 7 minutes). However, each musician made improvised musical decisions based on the generated music, set parameters chosen by the audience (e.g., tempo), and each other’s respective musical choices.

A. Self-reflective Analysis of Performer Dynamics

By examining the firsthand experiences of both the live improvising musician and the DJ, several themes emerge regarding agency, systemic constraints, and the necessity of highly adaptive collaboration.

1) Relinquishing Agency and Adversarial Audience Dynamics: A primary theme characterizing the performance was the live musician’s loss of creative agency to an unpredictable, audience-driven generative system. The Guitarist described the experience as stripping them of traditional control, noting, “All the agency is taken away from me... I have almost none.” Rather than building upon predictable elements in other improvised performances the live musician was forced to react to what they perceived as an adversarial audience actively seeking to test the performer’s limits. The Guitarist observed an explicit “push towards chaos” from the crowd, where participants frequently selected “intentionally wonky” combinations simply to “push my capabilities as far as they could possibly go.”

2) Structural Mediation and Systemic Constraints: While the Guitarist contended with the audience’s extreme selections, the DJ stated that much of their attention focused on actively managing the system’s output to maintain musical viability. The DJ reported feeling “a bit limited in the ways in which I could build up the song... especially in certain genres...” The Guitarist echoed this sentiment regarding the AI’s structural homogeneity, observing that despite the parameters changing, the underlying tracks could “feel [like they’re] getting stale at times.” These issues could be mitigated in future systems by increasing the number of musical options generated, or even enabling the audience to weigh in on “real-time decisions to change up the music” as was mentioned by the guitar player.

3) Creative Expansion and the Reward of Coherence: Despite the distinct frustrations of navigating restricted agency, adversarial audience inputs, and system constraints, both performers reported highly rewarding creative outcomes. The sheer unpredictability of the generative outputs forced the Guitarist to combine highly disparate stylistic elements—such as blending delicate acoustic atmospheres with heavy metal distortion—which pushed them to “creative levels... that are beyond the scope of what I would even consider doing.” The

tension between the audience’s push for chaos and the performers’ aim for structure resulted in many opportunities for a satisfying sense of musical cohesion. As the Guitarist reflected, “Anytime there’s a moment of coherence or a moment of, like, oh, wow, that sounds really good, that’s a magical moment in an improvised jam session.” These accounts show that while AI-mediated audience participation introduces operational friction, the necessity for adaptive collaboration yields a unique and creatively generative performance environment.

VI. DISCUSSION

Our findings highlight the dynamics of multimodal audience participation in MR musical performances. A tension emerged between the affordances of generative AI, the behavioral limitations of form-based input, and the audience’s desire for continuous involvement and social connection through collective decision making. By analyzing behavioral data, psychometric questionnaires, and qualitative thematic analysis, several key implications for the design of interactive live music systems became apparent.

A. The Illusion of Choice: Parameter Hierarchy and Behavioral Limits

The behavioral analysis suggests a clear cognitive **threshold regarding the time and effort participants are willing to dedicate to asynchronous, form-based input**. Despite being presented with a large combinatorial space of generative AI parameters, engagement exhibited diminishing returns. When all options were available in the “Both” condition, participants consistently bypassed granular AI configurations (e.g., AI instrument effects, showing skip rates up to 38.5%) in favor of macro-structural choices like tempo (0% skip rate) and visual characters (7.7% skip rate) indicating a decision fatigue for less impactful parameters.

This behavioral data aligns with our subjective Influence metrics, where foundational musical elements significantly outweighed nuanced AI settings. Notably, customizing the “Guitar tone for human player” yielded the highest perceived influence ($p = 0.002, r = 0.85$) and was the only parameter to significantly elevate feelings of specific agency above the neutral baseline ($p = 0.007, r = 0.72$). This human-centric focus reveals that **AI in live musical contexts functions most effectively as a supplement for performance manipulation rather than the primary object of attention**. Thematic data corroborates this, indicating that audiences actively sought to challenge the live human improviser—often selecting extreme parameter combinations to test the boundaries of the performer’s adaptability. Consequently, **agency was heavily dependent on observability**; participants needed to see or hear the human musician react to their specific choices to feel a genuine sense of control.

The findings suggest that perceived agency is driven less by the number of parameters offered to participants than by the *observability* of their influence on the performance. Audience members consistently favored parameters whose effects were recognizable, and attributable to their actions, such as tempo,

musical style, human guitar tone, or visible changes in the virtual environment. Conversely, granular AI parameters were often ignored because their individual contribution to the resulting performance was difficult to perceive. This suggests that the design of TMAP systems **should prioritize making the consequences of audience actions explicit rather than simply increasing the number of controllable parameters**.

B. Sensory Asymmetries: Affective Audio vs. Cohesive Visuals

Our results suggest **complementary roles for the audio and visual participation modalities**. Within the evaluated CrowdForge performance, audio participation was associated with affective engagement, whereas visual participation primarily provided clearly observable feedback. These findings suggest that, in this context, the audio components became the main focus of participants’ engagement—especially due to the impact on human musicians—while the visual components complemented the experience by making audience influence more readily observable. Both the Audio and the “Both” conditions significantly elevated Valence ($p < 0.05, r > 0.60$), Arousal ($p < 0.01, r > 0.70$), and SAM Dominance ($p < 0.05, r > 0.60$) above a neutral baseline. Although tangential to the primary focus, and perceptibly “programmable”, visuals have the opportunity to provide a sense of agency at scale as they conflict less generally than audio parameters and may have had more perceived influence if more tightly coupled to a human performers output.

Auditory manipulation, is highly susceptible to structural conflict. As highlighted in the qualitative analysis, disparate musical choices frequently resulted in chaotic, dissonant noise. This is likely due to constraints imposed by the nature of sound waves, which superimpose in unpredictable ways, unlike visuals. For example, applying musical effects to various instruments results in a combinatorial explosion that is intractable to test pre-performance (e.g., 5 instruments with 5 states including ON, OFF, and effects is 5^5 or 3,125 possible combinations). While some participants enjoyed the “sheer weirdness” of this dissonance, it often disrupted musical cohesion if too many auditory options are available. Some auditory parameters can be non-conflicting, such as genre, mode, and tempo, though these macro-structural parameters are limited. Conversely, thematic feedback revealed that visual changes (e.g., avatars, dancers, environments) were perceived as “most apparent” and less confounding. Because the visual domain enables multiple non-conflicting elements to be present simultaneously, it can offset the dilution of agency that participants experienced when some auditory choices were less apparent. Thus, **a multimodal system is optimal**: it utilizes audio to drive affective arousal while leveraging visuals to provide clear, parallel feedback that validates individual choices without compromising the overall aesthetic experience.

C. A Conceptual Model of Perceived Agency

Across the quantitative and qualitative findings, perceived agency depends primarily on the observability of the consequences of audience actions. We propose a conceptual model

in which interactive modalities first determine the observability of system responses, which in turn shape perceived agency and ultimately influence engagement and collaboration. Furthermore, parameters whose effects are high-impact and easily attributable—such as genre, guitar tone, or visible changes in the virtual environment—produce a stronger sense of agency than parameters whose influence is subtle or distributed across multiple components. Human performers further reinforce this process because their visible adaptation provides a socially meaningful confirmation that audience input has influenced the performance. This outlook would explain why audience members were keen to test parameter boundaries especially with respect to parameters that directly impact human performers. Thus, audio and visual modalities contribute differently to the audience experience and suggests that **future TMAP systems should consider prioritizing multi-modality and observable, high-impact parameters to enhance agency.**

D. Mediated Participation and Social Isolation

Despite the system’s collaborative premise, the asynchronous voting mechanism proved unsubstantial for fostering true social presence. Both our statistical and qualitative findings highlight that technology-mediated participation can concurrently enable mass interaction while inhibiting interpersonal connection. While the “Both” condition significantly improved Social Presence compared to the Visual condition ($p = 0.015, r = 0.65$), none of the three conditions managed to elevate Social Presence significantly above an absolute neutral baseline (all $p > 0.20$). The aggregation and averaging of votes obscured individual contributions, leading to frustration and a sense of isolation when participants’ choices were not explicitly reflected.

E. Towards Continuous Co-Creation and World Building

These insights advocate for interactive performance design to move away from episodic parameter selection toward continuous, collaborative world-building. Participants explicitly expressed an **intense desire for real-time influence and transparency of live audience voting**, suggesting that form-based voting should be replaced or augmented by dynamic, continuous interaction methods. Thematic feedback strongly supported interfaces where audiences can observe near-real-time voting trends. By integrating live, dynamic, collaborative, and transparent contributions—focusing on high-impact elements—future TMAPs can greatly enhance shared impact.

F. The Role of AI and MR: Enhancing Agency and Creativity

The primary benefit of AI-mediated audience participation is derived from the substantial combinatorial possibility space for which audiences can leverage to participate in novel music creation while challenging the creative boundaries of human musicians. This process was described by the guitarist as seemingly adversarial, while the DJ struggled to manage the system’s output to keep it viable. This tension furthered a sense of agency and arousal among the crowd, while supporting creative expansion beyond the musicians habitual range.

While the auditory domain appears the primary driver of affective engagement, visual elements play an equally important role in validating audience participation. Because visual changes (such as avatars, environments, and dancer animations) can coexist without the same level of perceptual clashing as audio, they provide recognizable feedback. MR greatly enhances this possibility space by introducing spatial, non-conflicting visual layers that can be continuously altered without affecting the structural integrity of the live music.

VII. LIMITATIONS AND FUTURE WORK

While this study provides insights into the dynamics of audience agency and social presence in AI-mediated performances, several limitations present clear avenues for future research. Primarily, the relatively small sample size ($N = 17$ for quantitative behavioral metrics), uneven audience distribution and diversity between sets, and limited number of performances necessitates larger participant pools to enhance statistical power and alleviate ordering effects between audio and visual parameters. This aspect is additionally confounded by a lack of influence of visuals on human performers presenting a further limitation. Given the focus of audio on human performer reactions and influence of human reaction on the perceived affect of the audience, future research could include improvised human dance to make the modalities more comparable. Additionally, observational data suggests a mathematical relationship between the number of configurable parameters and the total number of audience members. Future research should formally model this parameter-to-audience ratio to determine the optimal ratio required to maintain a robust sense of individual agency within crowds of varying sizes. Such studies should parse the impact of observability and agency with respect to a central versus peripheral focus.

The current evaluation was constrained to a specific co-located setup, limiting our understanding of how alternate viewing modalities affect the audience experience. Future work must explore fully immersive virtual reality, individualized MR, and remote participation to test the boundaries of this paradigm. Additionally, extending this system to remote participants would be a logical next step. This would enable us to explore how physically distributed audiences experience social presence and interpersonal connection, informing whether this technology can facilitate communal world-building outside of traditional, co-located concert environments.

VIII. CONCLUSIONS

This study evaluated a novel TMAP paradigm leveraging generative AI and MR to facilitate bidirectional interaction in live music. We assert that in TMAPs, observability is the primary driver of audience agency and engagement. Within our evaluated paradigm, MR visuals fostered agency by making audience contributions visible. Generative AI parameters, however, provided observable, macro-structural control over the musical environment, influencing the live musicians. Because audiences actively seek observable outcomes, they often used these AI parameters to test musical boundaries, introducing

a compelling adversarial tension between the audience and the performers. Rather than disrupting the performance, this dynamic pushes creative limits and drives deeper audience engagement. It also generates moments of musical coherence when the musicians successfully navigate and resolve the challenges presented by the audience. Future TMAP systems should leverage observability, continuous interaction, and human performer-affecting parameters to build agency while embracing the tension-resolution dynamic to elevate collaborative live music experiences.

REFERENCES

- [1] T. Hopkins, "Computer-supported music improvisation," Ph.D. dissertation, University of Colorado at Boulder, 2024.
- [2] Y. Wu, L. Zhang, N. Bryan-Kinns, and M. Barthet, "Open symphony: Creative participation for audiences of live music performances," *IEEE MultiMedia*, vol. 24, no. 1, pp. 48–62, 2017.
- [3] L. Zhang, Y. Wu, and M. Barthet, "A web application for audience participation in live music performance: The open symphony use case," in *Proceedings of the Conference on New Interfaces for Musical Expression*, 2016, pp. 170–175.
- [4] G. Weinberg, "Interconnected musical networks: bringing expression and thoughtfulness to collaborative group playing," Ph.D. dissertation, Massachusetts Institute of Technology, 2003.
- [5] N. Amarie, G. Fadda, A. A. Simiscuka, V. Popescu, and M. Murroni, "Emoloop: A bi-directional system for emotion-driven interaction between remote audiences and live performers," in *2025 IEEE 6th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2025, pp. 1–5.
- [6] O. Hödl, G. Fitzpatrick, and F. Kayali, "Design implications for technology-mediated audience participation in live music," in *Proceedings of the Sound and Music Computing Conference*, 2017, pp. 28–34.
- [7] D. Baldo and R. Masu, "State of the art on audience research within nime: a systematic," in *NIME '26, June 23–26, 2026, London, UK*, 2026.
- [8] O. Hödl and L. Turchet, "The tmap framework: describing the design space of technology-mediated audience participation in live music," *IEEE Access*, 2025.
- [9] T. Hopkins, S. S. Bae, J. Uhr, A. Banić, C. Zheng, and E. Y.-L. Do, "User interfaces in smart cities," in *Handbook of smart cities*. Springer, 2021, pp. 1–33.
- [10] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of Musical Things: Vision and Challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.
- [11] O. Hödl, F. Kayali, and G. Fitzpatrick, "Designing interactive audience participation using smart phones in a musical performance," in *Proceedings of the International Computer Music Conference*, 2012.
- [12] O. Hödl, C. Bartmann, F. Kayali, C. Löw, and P. Purgathofer, "Large-scale audience participation in live music using smartphones," *Journal of New Music Research*, vol. 49, no. 2, pp. 192–207, 2020.
- [13] M. Jung and I. Clester, "Hack the show: Design and analysis of three interaction modes for audience participation," *Journal of the Audio Engineering Society*, vol. 71, no. 11, pp. 769–779, 2023.
- [14] A. Xambó and V. Goudarzi, "The mobile audience as a digital musical persona in telematic performance," in *International Conference on New Interfaces for Musical Expression*. PubPub, 2022.
- [15] L. Turchet, "Musical Metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, pp. 1–17, 2023.
- [16] C. Rinaldi and C. Centofanti, "The musical metaverse: Advancements and applications in networked immersive audio," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–7.
- [17] T. Hopkins, E. Doherty, N. Ofer, S. C. C. Weng, P. Gyory, C. Tobin, L. Hirshfield, and E. Y.-L. Do, "Stringesthesia: Dynamically shifting musical agency between audience and performer based on trust in an interactive and improvised performance," *Audio Mostly 2023, Edinburgh, UK*, 2023.
- [18] J. Freeman, "Large audience participation, technology, and orchestral performance," in *Proceedings of the International Computer Music Conference*, 2005.
- [19] T. Hopkins, C. C. S. Weng, and E. Y.-L. Do, "Rhythmic Conversations: Using AI-Driven Virtual Musicians to Enhance Audience Participation in Musical Performance," *AIMC 2023*, aug 29 2023, <https://aimc2023.pubpub.org/pub/qot7sw4x>.
- [20] M. Hirabayashi and K. Eshima, "Sense of space: the audience participation music performance with high-frequency sound id," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2015, pp. 58–60.
- [21] N. Weitzner, J. Freeman, S. Garrett, and Y. Chen, "massMobile-an Audience Participation Framework," in *Proceedings of the Conference on New Interfaces for Musical Expression*, 2012, pp. 21–23.
- [22] Z. Ma, I. Kuzmin, D. Ruilei, and R. Masu, "Pharosphones: interactive audience participation using light," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2024, pp. 480–485.
- [23] S. Brade, T. Ma, L. Blanchard, K. Lecamwasam, C. M. Salcedo, S. Kim, P. Naseck, A. Li, M. R. Michalek, S. Franjou *et al.*, "Agents in concert: A case-study of bringing ai to the stage in practice," in *Proceedings of the 31st International Conference on Intelligent User Interfaces*, 2026, pp. 1340–1361.
- [24] F. Berthaut, D. M. Plasencia, M. Hachet, and S. Subramanian, "Reflats: Combining and revealing spaces for musical performances," in *New Interfaces for Musical Expression (NIME)*, 2015.
- [25] T. Hopkins, A. Jude, G. Phillips, and E. Y.-L. Do, "Virtual AI Jam: Ai-Driven Virtual Musicians for Human-in-the-Loop Musical Improvisation," *AIMC 2023*, aug 29 2023, <https://aimc2023.pubpub.org/pub/9064kmvw>.
- [26] T. Hopkins, R. Sun, S. C. C. Weng, S.-Y. Ma, J. Crum, L. Hirshfield, and E. Y.-L. Do, "Brain jam: Neural signal-informed adaptive system for drumming collaboration with an ai-driven virtual musician," *Frontiers in Computer Science*, vol. 7, p. 1570249, 2025.
- [27] Y. Kim, S. Brade, A. Wang, D. Zhou, H. Kim, B. Wang, S.-J. Lee, H. F. Flores Garcia, A. Huang, and C. Donahue, "A design space for live music agents," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026, pp. 1–36.
- [28] G. Hoffman and G. Weinberg, "Shimon: an interactive improvisational robotic marimba player," in *CHI'10 extended abstracts on Human Factors in Computing Systems*, 2010, pp. 3097–3102.
- [29] M. Bretan and G. Weinberg, "A survey of robotic musicianship," *Communications of the ACM*, vol. 59, no. 5, pp. 100–109, 2016.
- [30] D. Baldo and R. Masu, "State of the art on audience research within nime: a systematic," *NIME*, 2026.
- [31] T. Hopkins, S. C. C. Weng, R. Vanukuru, S. Novack, C. Tobin, E. Wenzel, A. Banic, M. Gross, and E. Y.-L. Do, "Xr jam: Design considerations for music networking with ai in xr," in *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE, 2024, pp. 97–101.
- [32] T. Hopkins, S. C. C. Weng, R. Vanukuru, E. A. Wenzel, A. Banic, M. D. Gross, and E. Y.-L. Do, "AR Drum Circle: Real-Time Collaborative Drumming in AR," *Frontiers in Virtual Reality*, vol. 3, 2022.
- [33] S. A. Bin, N. Bryan-Kinns, and A. P. McPherson, "Risky business: Disfluency as a design strategy," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2018, pp. 45–50.
- [34] D. Jones, "Abletonosc: A unified control api for ableton live," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Mexico City, Mexico, Article, vol. 60, no. 5, 2023.
- [35] L. Ou, J. Zhao, Z. Wang, G. Xia, Q. Liang, T. Hopkins, and Y. Wang, "Unifying symbolic music arrangement: Track-aware reconstruction and structured tokenization," *Advances in Neural Information Processing Systems*, vol. 38, pp. 14 721–14 750, 2025.
- [36] O. Badreddin, "Thematic review and analysis of grounded theory application in software engineering," *Advances in Software Engineering*, vol. 2013, no. 1, p. 468021, 2013.
- [37] T. Hopkins, S.-Y. Ma, S. C.-C. Weng, M.-Y. Pai, E. Y.-L. Do, and L. Turchet, "Mr-daw: Towards collaborative digital audio workstations in mixed reality," in *2025 IEEE 6th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2025, pp. 1–10.
- [38] G. Terry, N. Hayfield, V. Clarke, V. Braun *et al.*, "Thematic analysis," *The SAGE handbook of qualitative research in psychology*, vol. 2, no. 17-37, p. 25, 2017.
- [39] U. Kuckartz and S. Rädiker, *Analyzing qualitative data with MAXQDA*. Springer, 2019.