Int. J. Human–Computer Studies 203 (2025) 103586

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Between immersion and usability: A comparative study of 2D and mixed reality interfaces for remote music making

Alberto Boem *, Matteo Tomasetti, Luca Turchet

Department of Information Engineering and Computer Science, University of Trento, Italy

ARTICLE INFO

Keywords:
Mixed reality
Music technology
User interfaces
Musical metaverse
Networked music performance
User experience

ABSTRACT

Networked Music Performance (NMP) systems have traditionally focused on optimizing audio transmission for remote collaboration, often relying on basic 2D video feeds that lack the spatial awareness and visual cues essential for coordinated musical performance. While Mixed Reality (MR) technology offers promising enhancements for musical collaboration, its integration into NMP systems remains understudied–particularly regarding the balance between immersion and usability. This study addresses that gap by comparing traditional 2D video conferencing setups with MR environments featuring point cloud representations and spatial audio. Using a simulation-based approach with pre-recorded stimuli, we conducted an experimental study with professional electric guitar players ($N = 16$) to examine how both setups affect musical performance, movement, and sense of presence, combining quantitative motion analysis with qualitative feedback. Head movement analysis revealed significantly reduced motion in MR compared to 2D conditions ($p < 0.001$), while detrended fluctuation analysis indicated more structured movement patterns in MR ($p < 0.01$). Post-task evaluations showed that although MR enhanced immersion and presence ($p < 0.001$), participants reported greater performance coherence and fewer technical interferences in the 2D condition ($p < 0.001$). Spatial audio had minimal impact on user experience, with visual elements drawing most of the participants' attention. These findings underscore a fundamental tension between immersion and physical expressiveness in MR-based music performance, suggesting that future systems should prioritize ergonomics alongside technological innovation. This research contributes to a deeper understanding of the potential and limitations of MR in musical collaboration within the emerging Musical Metaverse.

1. Introduction

The concept of the Musical Metaverse (MM) represents a novel frontier for music collaboration, merging digital and physical worlds to create shared immersive environments where co-located and remote musicians can perform, create, and interact (Turchet, 2023). Within this landscape, Networked Music Performance (NMP) systems enable real-time musical interactions between distant locations (Renaud et al., 2007; Gabrielli et al., 2016; Rottondi et al., 2016). While NMP systems have focused primarily on achieving seamless audio transmission, the MM requires a more comprehensive approach that integrates visual, kinesthetic, and spatial perception, which are essential features for supporting musical collaboration.

Making music together necessitates a communal space grounded in face-to-face interactions (Schütz, 1951). Research has shown that such musical interaction relies heavily on visual cues, including body movements, gestures, and facial expressions, for communication, synchronization, and emotional expression (e.g., Dahl and Friberg, 2007; Bishop et al., 2021; Bishop, 2024). NMP enables two main forms of collaborative music-making: realistic jam sessions and remote recording (Renaud et al., 2007). Realistic jam sessions have emerged as the focal point of research interest (Boem et al., 2025), aiming to transcend geographical barriers by creating an experience where musicians feel as if they are performing together in the same physical space. This requires sophisticated technical solutions to handle rapid audio transmission, sound spatialization, and visual communication between performers (Schiavio et al., 2022). Until now NMP systems have relied on simple webcam feeds to provide visual feedback between musicians. While these 2D video streams allow performers to see each other, they significantly limit spatial awareness, making it difficult for musicians to understand the relationships between performers in space (Renaud et al., 2007; Delle Monache et al., 2019).

Mixed Reality (MR) (Speicher et al., 2019; Billinghurst and Kato, 1999) represents a promising solution by blending physical and virtual elements. Recent advances in XR technologies have established MR

* Corresponding author.
E-mail addresses: alberto.boem@unitn.it (A. Boem), matteo.tomasetti@unitn.it (M. Tomasetti), luca.turchet@unitn.it (L. Turchet).

https://doi.org/10.1016/j.ijhcs.2025.103586
Received 21 February 2025; Received in revised form 16 July 2025; Accepted 22 July 2025
Available online 31 July 2025

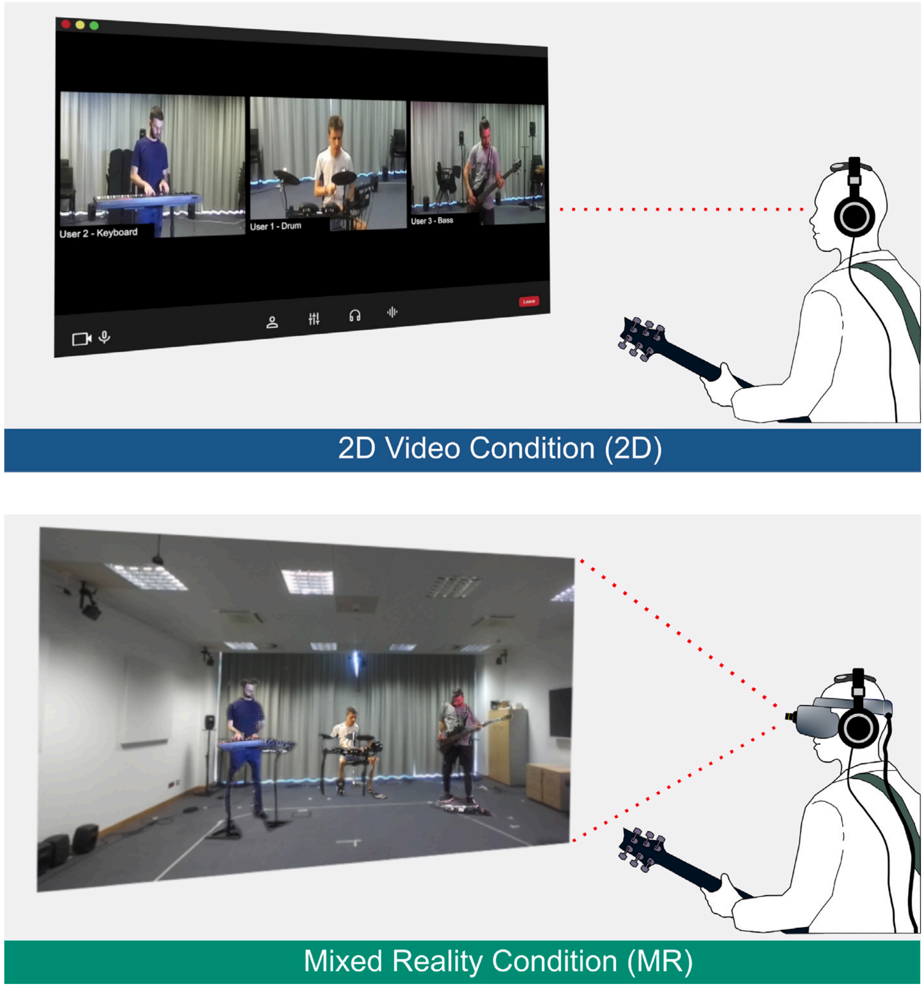

Fig. 1. Comparison of the two systems used in our study: on the left, a telemeeting interface displaying remote musicians as 2D video streams with stereo audio; on the right, a MR system rendering remote musicians as point clouds with binaural audio.

as a powerful tool for remote collaboration, with volumetric video technology enabling photorealistic capture of participants and their surroundings in real-time (Orts-Escolano et al., 2016; Abdallah et al., 2023; Viola et al., 2023; Cortés et al., 2024).

However, integrating MR and NMP systems presents significant technical challenges. Even if preliminary, existing studies in this space have consistently encountered fundamental obstacles, particularly regarding the latency and jitter introduced by the network and by the concurrent streaming of audio and video, as well as their synchronization. These limitations have typically restricted experiments to two-person setups and resulted in latency issues that can compromise the musical experience, even in co-located environments (Turchet et al., 2022; Schlagowski et al., 2023). The current technological landscape makes it particularly challenging to achieve the latency required for synchronous musical performance (which needs to be kept below 30 ms (Rottondi et al., 2016)) while simultaneously managing the bandwidth and processing demands of volumetric video streams (Liu et al., 2021).

While real-time interaction remains the ultimate goal for networked musical collaboration, current technological limitations make it challenging to fully assess user experience. Previous research has examined the Quality of Service and the mental and physiological effects of MR NMP systems. However, significant gaps remain in our understanding of how these systems restrict musicians' movements, affect cross-modal perception, and influence spatial awareness in multi-performer scenarios. Furthermore, existing studies have been limited by inconsistent methodologies, involving participants with varying musical expertise and different instruments, along with non-standardized tasks that are difficult to replicate.

To overcome these limitations and enable a systematic study of MR and volumetric video potential in the context of realistic jamming, we developed a simulation-based approach rather than attempting a real-time implementation involving the network. We present a system and an experimental protocol for a controlled environment that allows us to:

- Remove synchronization issues between audio and visual streams, avoiding the technical limitations that have hampered previous real-time studies;
- Support larger musical ensembles (beyond two participants) to better reflect real-world music collaboration scenarios;
- Maintain consistent technical performance across all experimental sessions, ensuring reproducibility and reliable comparisons;
- Create identical conditions for all participants, enabling systematic evaluation of user experience.

Recent research has suggested that using point-clouds of distant musicians in MR can enhance the sense of presence (Schlagowski et al., 2023). However, since the most common way to deliver visual feedback in current NMP systems is through 2D video feedback similar to the ones used in video conferencing systems, it is not clear which are the advantages and limitations of using more immersive setups. Beyond the visual aspect, since the majority of existing NMP systems use stereo audio, it is not clear if spatial audio could add a different dimension to the final user experience. To investigate this, we provide a comparison between traditional 2D interfaces with stereo audio and MR systems with spatial audio for remote jamming. Understanding how MR interfaces affect musical interaction patterns is essential for developing systems that enhance rather than hinder collaborative music-making in multiuser immersive environments.

Our work compares a traditional 2D video setup with stereo audio for NMP telemeetings with an MR environment utilizing point cloud videos and binaural audio (see Fig. 1). Through this controlled comparison, made with pre-recorded materials, we examine how these different approaches impact musicians' performance, engagement, and overall user experience. We combine quantitative metrics — such as Quantity of Motion and head orientation — with qualitative feedback from 16 professional electric guitar players to evaluate the challenges and opportunities presented by MR-enhanced NMP systems.

Specifically, this study addresses four key research questions:

RQ1: How does a MR setup impact the performance and engagement of musicians compared to a 2D video setup?

RQ2: What challenges and opportunities arise when integrating MR in a realistic jamming scenario, particularly concerning the physical comfort of musicians using head-mounted displays (HMDs)?

RQ3: To what extent do musicians' physical movements differ between MR and traditional 2D systems?

RQ4: What are the effects of point cloud videos and binaural audio on the sense of presence and immersion in remote musical collaborations?

Our findings contribute to understanding the potential and limitations of MR technologies in musical collaboration, particularly focusing on how musicians adapt to and interact within immersive environments. This work provides insights for the future development of a more effective and user-oriented combination of Mixed Reality and NMP technologies within the MM.

2. Related work

Most research on NMP systems has concentrated on developing audio-based systems that minimize latency, reduce jitter, and maintain high audio quality (Gabrielli et al., 2016; Rottondi et al., 2016). Latency must be kept below 30 ms. The main outcomes are systems that support the transmission of audio data through a network, as shown by some of the most popular tools such as JackTrip (Cáceres and Chafe, 2010), Jamulus (Jamulus, 2025), or Sonobus (Sonobus, 2025). However, musical collaboration extends beyond sound transmission. Visual and kinesthetic cues, such as body movements and gestures, play a key role in performance, synchronization, and emotional expression (D'Amario et al., 2018; Dahl and Friberg, 2007; Vines et al., 2011; Delle Monache et al., 2019).

2.1. Visual and body cues in music performance

Previous research has found that visual cues alone can effectively communicate expressive information of a performance (Davidson, 1993; Behne and Wöllner, 2011; Tsay, 2014). It was also observed that visuals often dominate the understanding of aspects of music performance such as emotion (Dahl and Friberg, 2007), and expressiveness (Broughton and Stevens, 2009). However, the visual aspect of music performance is dominated by the movements of musicians' bodies. As shown by research on audience perception, independently from their musical role, musicians that exhibit a high amount of movement are better perceived and valued (Davidson, 1993; Nusseck and Wanderley, 2009; Burger and Wöllner, 2023). Recognition of emotions in music performance is also linked to kinesthetic properties of the body, such as speed, regularity, and the quantity of motion (Dahl and Friberg, 2007). Compared to other body parts, the head has proven to be especially effective in conveying expressiveness (Davidson, 1994; Massie-Laberge et al., 2016; Dahl and Friberg, 2007). In group performances, research showed that musicians achieve synchronization more effectively when they can observe each other's body movements compared to when they lack visual contact (D'Amario et al., 2018; Palmer et al., 2019). Therefore, visual and kinesthetic cues play an important role especially when musicians need to synchronize tempo changes or begin sections at the same time (Bishop and Goebl, 2015; Timmers et al., 2014; Williamon and Davidson, 2002). These results point towards the idea that musical coordination has a close relationship with social coordination and may influence musical coordination (Davidson and Good, 2002; Bishop et al., 2021). Furthermore, gesture and posture expressions are considered very important for interpersonal communication towards reaching a common performance goal (Argyle, 2013). The perception of musical performance emerges from the interplay between visual and auditory elements — specifically, how we see musicians' body movements and how these movements align with the sounds we hear. This cross-modal relationship between sight and audition is crucial for creating an authentic performance environment. These findings emphasize why visual communication is essential in technology-mediated music performances — it is not enough to just hear the other musicians, we need to see them too.

2.2. Visual and bodily cues in NMP

One approach to transmit visual and body cues between peer musicians and enhance the sense of togetherness in NMP was done by integrating audio streams with video feeds of musicians (Carôt and Schuller, 2011; Olmos et al., 2012; Redman, 2020). Examples of NMP applications that include real-time video feeds are LoLa (Drioli et al., 2013), JackTrip Virtual Studio (JackTrip, 2025), and Elk Studio Live (Elk, 2025). In terms of design, these systems often resemble videoconferencing tools, such as Zoom or Microsoft Teams, in terms of user interfaces, functionalities, and layout. However, they lack the spatial depth and resolution needed for richer interactions. While effective in providing immediate feedback, 2D videos present several drawbacks. First of all, flat videos may constrain some important non-verbal cues available in face-to-face situations, including depth, resolution, and field of view (Smith and Neff, 2018). Peers are visible to each other, but not co-present, thus lacking a common spatial reference frame (Fussell et al., 2000). This situation highlights the importance of visual channels in NMP systems, but further research is needed to compare the impact of 2D video with more immersive solutions, such as those offered by MR.

2.3. Mixed reality in NMP

XR technologies have emerged as powerful tools for remote collaboration, offering new ways for people to interact through telepresence and conferencing systems (Kantonen et al., 2010; Smith and Neff, 2018; Qiu et al., 2023). Research has shown that using three-dimensional avatars in these teleimmersive environments can create a sense of social presence that closely matches the experience of face-to-face interactions (Smith and Neff, 2018). Moreover, it was highlighted that when these systems create photorealistic representations of participants, people can communicate naturally and express emotions just as they would in person (Zioulis et al., 2016; Son et al., 2020).

2.3.1. Embodied avatars

While traditional video conferencing has many readily available solutions, XR-based networked music performance systems — particularly those using MR — are still in their early stages. These systems currently require custom-built setups and remain at the prototype level (Boem et al., 2024). Recently, the use of embodied 3D avatars has started to be explored in immersive NMP systems (Hupke et al., 2022; Hopkins et al., 2022; Cairns et al., 2023; Hunt et al., 2023; Van Kerrebroeck et al., 2024). For example, Van Kerrebroeck et al. (2024) found that visual coupling through avatars produced prosocial effects and flow similar to the ones of face-to-face interactions, and better than in an audio-only condition. Hupke et al. (2024) found improvements in social presence and immersion when using avatars in networked rhythmic tasks. However, these systems faced important limitations. First, they were evaluated only with dyads and with very simple rhythmic tasks, which limits their broader applicability. Second, these settings relied on specialized tracking equipment with reflective or active markers, that need to be worn by musicians. Finally, physical musical instruments are difficult to track and render in real-time.

2.3.2. Volumetric videos

Volumetric videos present a promising alternative for achieving photorealistic representation of humans and musical instruments (Young et al., 2022). Point-cloud video technologies have been successfully explored for remote collaboration since they can offer a 6DoF viewing experience (Orts-Escolano et al., 2016; Abdallah et al., 2023; Cortés et al., 2024). Such kind of technologies are considered the next step for video meetings (Jansen et al., 2020; Regenbrecht et al., 2021; Gunkel et al., 2023; Viola et al., 2023; Liu et al., 2021), which can be compressed and transmitted over networks, particularly with the emergence of 5G and 6G connectivity (Liu et al., 2021).

Point-cloud videos have recently gained attention in the NMP community for realistic jamming. However, to date, the examples are limited to just a few instances. Turchet et al. (2022) explored the volumetric representation of remote musicians in MR through a setup that allowed the concurrent streaming of both point-cloud video and audio data. While the general idea was well received by the two involved participants, the experience was negatively impacted by synchronization issues due to video transmission latency. Schlagowski et al. (2023) showed with a similar setup that point-cloud videos in MR can improve the sense of co-presence among musicians. Nevertheless, they found no significant difference in flow and affect compared to a condition when only audio was transmitted between musicians. Differently from the previous one, to avoid synchronization issues, audio latency, and signal degradation, only video data were streamed, while audio was exchanged between peers through audio cables.

While insightful, these studies revealed several technical challenges. Volumetric video streaming demands specialized internet connections and efficient protocols (Orts-Escolano et al., 2016; Liu et al., 2021), as realistic representations require significantly larger data volumes than 2D video streams (Viola et al., 2023). Cortés et al. (2024) found that delays of up to 900 ms are acceptable for collaborative tasks in volumetric MR. However, synchronizing such video feeds with low-latency audio, as required by NMP systems is not feasible.

Moreover, these studies focused on two-person setups, whereas NMP systems must typically support larger musical ensembles, adding further complexity.

2.4. Audio in immersive NMP

Most NMP systems use mono or stereo audio that should be experienced through headphones. Nevertheless, spatial audio techniques, especially binaural and ambisonics (Zotter and Frank, 2019), have been explored to provide more expressive and congruent NMP environments (Gurevich et al., 2011; Turchet et al., 2024). Indeed, it has been recently shown that ambisonics and binaural audio improve the sense of social connectedness and realism compared to stereo setups (Tomasetti and Turchet, 2023). Spatial audio encompasses technologies developed to simulate the experience of detecting single or multiple sound sources in three-dimensional space, reflecting the natural human ability to locate sounds in the real world.

However, the majority of previous work on spatial audio for NMPs has been done in audio-only conditions, and spatial audio is not included in current commercial or open-source NMP platforms. Moreover, the use of spatial audio as a means to increase the sense of presence has not been investigated yet with point cloud videos in MR.

3. Material and methods

To systematically compare 2D and MR interfaces in an ecologically valid settings, we conducted a user study involving 16 professional electric guitar players. For this purpose, we developed a system that allows us to playback pre-recorded audio and video sequences, which serve as stimuli. These stimuli include four songs performed by a trio of musicians, with which a subject can interact. Unlike previous studies focusing on two-person interactions in musical MR, our research explores the full complexity of band interaction by simulating a four-person remote ensemble. This enabled us to avoid technical issues such as network jitter and latency and, therefore, to focus on the user experience of musicians.

Our goal was to simulate the experience of an electric guitar player performing in a band composed of drums, electric bass, and keyboard, with these instruments remotely participating in the performance. We chose the electric guitar because it is a widely used musical instrument that can be easily interfaced with digital equipment. Moreover, it serves both rhythmic and melodic roles, making it suitable for improvisation and accompaniment.

We designed two playing conditions based on existing technologies as well as research at the intersection of MR and NMP:

1. *(1) 2D NMP:* This condition mimics the experience of a musician using a GUI of an NMP tool such as JackTrip Virtual Studio or Elk Live Studio. Positioned in front of a 2D monitor, the musician plays with three other connected musicians, whose video feeds are displayed in a gallery view. As in conventional NMP systems, audio is presented as a stereo mix of the individual signals from the remote musicians.
2. *(2) MR NMP:* Inspired by the works reported in Schlagowski et al. (2023) and Turchet et al. (2022), this system simulates the experience of playing with an MR-based NMP where remote users are presented as volumetric point-clouds. Here, the musician wears an HMD with a color passthrough, allowing the remote musicians to be visualized as if they were in the same physical space as the participant. We integrated binaural audio based on results reported in Tomasetti and Turchet (2023) because it is considered the next direction for an interface of an immersive NMP (Turchet et al., 2024).

The complete system and related materials are available in a Github repository.[1]

3.1. User study

The experiments were conducted in a laboratory of the University of Trento equipped for studies on musical interaction in XR settings. A within-subjects design was adopted, which involved two independent variables (2D and MR conditions) and five dependent variables related to user performance: mean quantity of motion, mean head quantity of motion, mean yaw, mean pitch, and mean roll. Additionally, variables related to the used questionnaires were considered. Four musical pieces were recorded and utilized as audio and visual stimuli. The order of presentation for the conditions and pieces was randomized.

3.2. Participants

Sixteen participants took part in the evaluation session (2 females, and 14 males), aged between 22 and 51 (mean age = 30.68, SD = 7.19). All participants were highly experienced electric guitar players, with 10 to 40 years of experience (mean years = 17, SD = 8.61). They are active musicians, working as session musicians, songwriters, and teachers. Nine had previous sporadic exposure to XR experiences,

[1] https://github.com/CIMIL/MR_2D_Experiment_System.

Table 1
Summary of the emotional characteristics, musical genre, duration, tonality, and tempo (BPM) of the four pieces performed by the recorded musicians, which were used as stimuli during the experiment.

Song	Emotion	Genre/Style	Duration	Tonality	BPM
1	Angy	Hard rock	03:11	Am	170
2	Relaxed	Pop, Ballad	03:18	Am	73
3	Happy	Rock 'n' Roll	03:28	A7	160
4	Sad	Ballad, traditional	03:14	Am	70

primarily in the form of VR games, and three had tried telemeeting systems such as Zoom, mostly for teaching music during the COVID-19 pandemic. All participants were right-handed. Participants were asked to read and give consent to the study information, in compliance with the provisions of the University of Trento ethical committee.

3.3. Musical content selection

The musical content selection process balanced technical and artistic considerations. We recruited three composers to suggest musical pieces based primarily on their affective characteristics rather than tonality or tempo, while ensuring compatibility with our instrumental ensemble (drums, bass, keyboard, and guitar).

We instructed each composer to select three pieces representing the four quadrants of Russell's valence-arousal circumplex model (anger, happiness, relaxation, and sadness) (Russell, 1980), resulting in 36 pieces in total. To provide clearer guidance, we used Hevner's Adjective Circle (Hevner, 1936), a model used for categorizing emotional responses to music. Therefore, 'anger" can be described with adjectives such as *"agitated, angry, excited"*, "happiness" with *"cheerful, playful, lively"*, "sadness" with *"depressed, sorrowful, serious"*, and "relaxation" with *"dreamy, reflective, calm"*.

Subsequently, the authors and composers collaboratively selected four final pieces using additional criteria such as appropriate complexity and accessibility not only for the participants to our study but also the session musicians who recorded the audiovisual stimuli. We prioritized pieces that offered sufficient harmonic, melodic, and rhythmic interest while remaining approachable for quick learning and performance. Table 1 summarizes the characteristics of the four selected pieces.

3.4. Recordings

To create and deliver a realistic experience, we developed a systematic recording process able to capture the performances of four professional musicians: a drummer, a bassist, and a keyboard player. We recruited three professional musicians who volunteered to participate in these audio and video recordings. Each musician recorded their parts separately, incrementally building the group performance. The process began with the drummer, followed by each subsequent musician performing their part while listening to the previously recorded tracks. This approach ensured tight musical coordination and allowed for multiple takes (typically three to five per song) to select the best performance for each instrument. To enhance the realism of these recordings, we structured each session to mimic authentic jamming scenarios. Each recording began and ended with 20 s of silence, during which musicians were instructed to perform setup actions, such as adjusting their guitar straps, setting drum hardware, or checking keyboard settings. These details contributed to the ecological validity of our simulated performance environment.

3.5. Data acquisition pipeline

For data acquisition, we used a setup based on a Desktop PC with Windows 10 equipped with a GPU NVIDIA GTX 3090 and a RME Fireface UFX soundcard, operated with the TotalMix FX application (v.183). The process and technical pipeline are summarized in Fig. 3.

3.5.1. Video data acquisition

For acquiring the video data we used the Stereolabs ZED 2, a stereoscopic depth-sensing camera. For real-time data acquisition, we developed a custom application with Python (v3.8) using Stereolabs' ZED Python API (v4.0). The recordings were made with a resolution of $2 \times (1280 \times 720$ pixels) and 60 fps. The depth range was set from a minimum of 1 to a maximum of 3 m.[2] The video stream was directly recorded on the hard disk of the host PC, and encoded in a .svo file. This proprietary file format by Stereolabs stores the synchronized left and right video streams captured by the ZED camera. This data is used for further processing through the ZED SDK. The result of the data acquisition for video data is presented in Fig. 2.

3.5.2. Audio data acquisition

In parallel with the video data, we captured the audio signal from each musician's instrument. We used the Digital Audio Workstation (DAW) Cockos Reaper (v6.78) for recording the audio streams. Each track was recorded at a sample rate of 48,000 Hz and saved as a .wav file.

3.5.3. Synchronization of video and audio data

We synchronized the recordings using Open Sound Control (OSC) (Wright, 2005) messages sent from a Python GUI application developed with Tkinter. The data acquisition process was controlled by an operator who could manually start and stop the recordings using buttons in the GUI. When the recording started, the ZED camera was activated, and after a brief delay (corresponding to the time required for the ZED camera to begin acquisition), the video data was captured. An OSC message was then sent to Reaper to initiate the audio data acquisition. Both recordings could be manually stopped and paused as needed.

3.6. Apparatus

For the user study, we developed two distinct presentation systems using Unity (2020.3.48f1 LTS) with the ZED Plugin for Unity (v4.0.6): one for the 2D Condition and another for the MR Condition. Both systems operated on a Desktop PC running Windows 10, equipped with an NVIDIA GTX 3090 GPU and an RME Fireface UFX soundcard, managed via the TotalMix FX application (v.183). This setup allowed us to playback the recorded tracks and manipulate them for presentation in the 2D and MR conditions. The apparatus and layout of the experiment are detailed in Fig. 4.

3.6.1. 2D condition

In the 2D condition, participants were presented with a familiar video conferencing interface, displaying three video streams in a gallery layout, complete with non-functional UI elements to enhance authenticity. The design was inspired by applications such as JackTrip Virtual Studio, and Elk Studio Live, as well as videoconferencing tools such as Zoom. In this application, the .svo files were displayed as monoscopic videos. To ensure synchronization, audio and video files for each musician were loaded simultaneously.

3.6.2. MR condition

The MR condition employs a Stereolabs ZED Mini camera mounted on an HTC Vive headset, enabling high-quality color passthrough video and 6-DoF positional tracking. The passthrough camera operates at 60 fps, with a resolution of 1280×720 pixels per lens. The three remote musicians are displayed as point clouds using the shader provided by the ZED SDK (see Figure). The MR implementation requires careful attention to the spatial configuration of the pre-recorded musicians. We use the left and right HTC handheld controllers as positional

[2] https://www.stereolabs.com/docs/depth-sensing/depth-settings#depth-modes, (Last accessed 05/01/2025).

Fig. 2. The result of the Video Data acquisition: (1) the video stream of the bass player recorded on the ZED 2 camera (.svo) is rendered in Unity as a video frame and cropped (for 2D condition); (2) is rendered as a point-cloud through a shader.

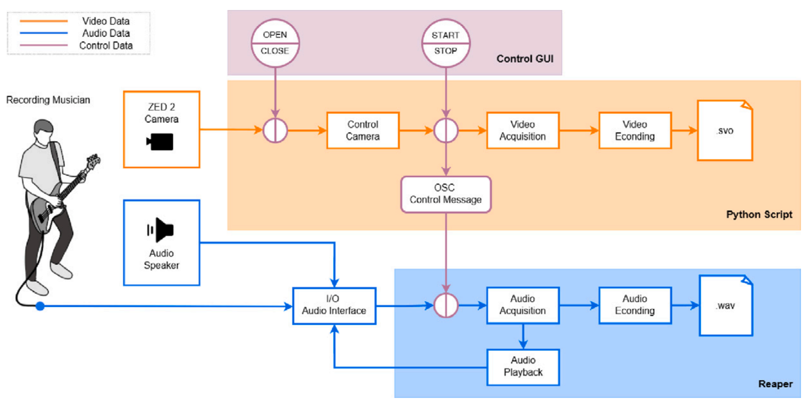

Fig. 3. Diagram of the system used for capturing video and audio data from musicians during recording sessions. The setup includes the integration of various hardware and software components to ensure synchronized data acquisition.

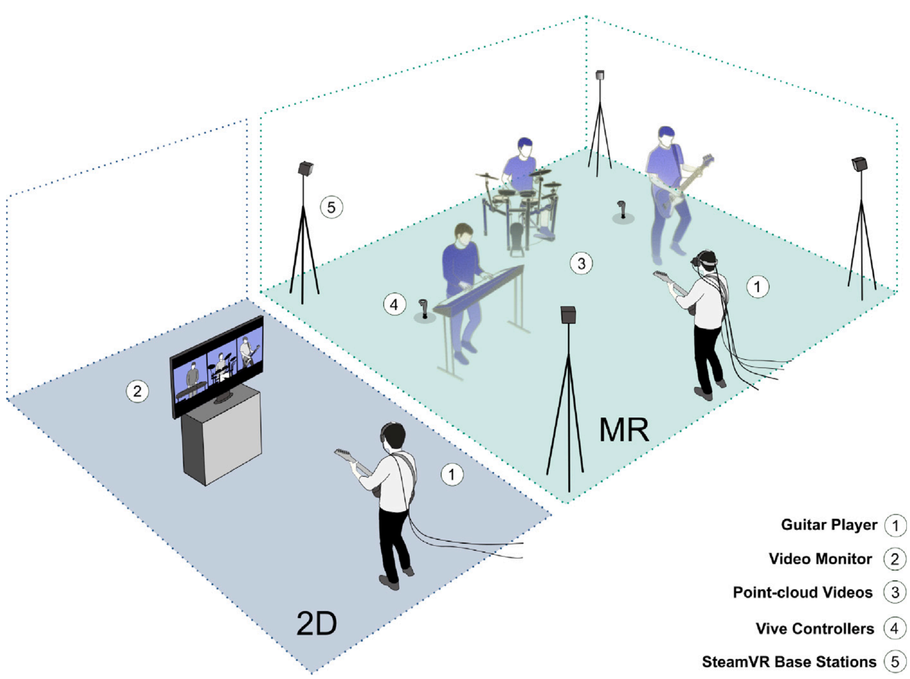

Fig. 4. Experimental setup illustrating the configuration for both 2D and MR conditions. The layout includes the positioning of the display devices and the spatial arrangement of the virtual musicians in the MR environment.

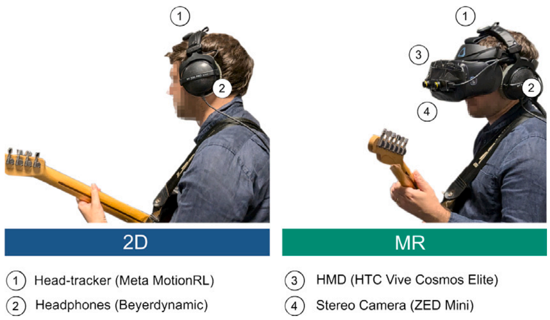

Fig. 5. Equipment setup for participants during the experiment: The 2D condition (left) involves a standard monitor and headphones, while the MR condition (right) uses an HMD with a ZED Mini camera for passthrough, providing an immersive experience.

anchors. We placed them on the floor with custom 3D-printed stands. Their position defines the placement of the bass player (right) and keyboardist (left), with the drummer's position (center) calculated as the midpoint between them. This setup, tracked by four SteamVR Base Stations 2.0, ensures consistent performer placement across sessions while maintaining flexibility for different room configurations. With this setup, we aim to display the musicians in the space visible to a user wearing an HMD. To achieve this, we use a combination of tracking provided by the ZED Mini and OpenXR through SteamVR. During playback, we achieved a stable frame rate of approximately 90 fps.

3.6.3. Audio rendering

Audio implementation differs significantly between the two conditions. In the 2D setup, conventional stereo playback is used, with the three audio tracks mixed together using the Unity-integrated Audio Mixer. In contrast, the MR condition employs spatial audio, where each audio source corresponds to the position of each musician's point cloud. We utilized the VIVE3DSP audio SDK (v1.2.8) for 3rd High Order Ambisonics, ensuring better integration with our software and hardware workflow. The MR audio system takes into account room acoustics based on measured dimensions and materials, creating a cohesive spatial audio experience that aligns with the visual representation of the performance space. The room size is $7 \times 10 \times 3$ m, with materials such as "*plaster absorber concrete*" for three walls, floor, and ceiling and "*curtains*" for the wall behind the musicians. We applied a reverb level of 3 dB.[3]

3.7. Setup

The setup used for the experiment is detailed in Fig. 5. In both conditions 2D and MR, audio was delivered using a pair of Beyerdynamic DT-770 Pro 80 Ohm headphones (weight: 346 g), connected to the sound card. For the 2D Condition, the video stream was presented through a video monitor (LG 42lm3450, resolution: 1920×1080). For the MR Condition, the visual stimuli were delivered through an HTC Vive Cosmos Elite HMD (resolution per-eye: 1440×1700, refresh rate: 90 Hz, horizontal field of view (FOV): 97 deg, vertical FOV: 94 deg, weight: 600 g), connected to the host PC, with a ZED Mini camera attached on its front to provide color passthrough (weight: 62 g). The total weight was 662 g. Both the monitor and positional trackers (representing the position of the point clouds) were placed at 200 cm in front of the participants and at a distance of 300 cm between each other. The monitor was placed at a height of 150 cm to cover the average eyesight. These measurements were obtained after a pilot test with three proficient guitarists, who were not included in the user study.

[3] See the details https://hub.vive.com/storage/3dsp/vive_3dsp_audio_sdk_unity_plugin.html (Last accessed: 01/10/2025).

Fig. 6. Participants improvising with their guitars and effects in both 2D and MR conditions.

3.8. Task

The experimental task was designed to simulate a realistic remote band performance scenario. Each participant was asked to improvise with a pre-recorded three-piece band across four distinct musical pieces in two conditions: 2D and MR (see Fig. 6).

In the 2D Condition, participants faced a monitor displaying the three video streams of the band members in a standard video conferencing layout, with stereo audio delivered through headphones.

In the MR Condition, participants wore an HMD through which they could see point-cloud representations of the band members positioned in their physical space, with binaural audio providing spatial cues corresponding to each musician's position.

In both conditions, participants used the same guitar setup and audio monitoring system to ensure consistency in their playing experience. The differences in visual presentation and audio spatialization between conditions allowed us to examine how these interface elements affected their performance, engagement, and sense of presence during the musical interaction. Each participant performed all four pieces in both conditions, with the order of conditions (2D and MR) and pieces (Angry, Sad, Relaxed, Happy) counterbalanced to control for learning effects. Between conditions, participants were given a 10 min break to prevent fatigue and maintain consistent performance quality throughout the experiment.

3.9. Procedure

Participants were welcomed into the room equipped with the system and briefed about the experiment's purpose and goals. After signing the consent form, they set up their equipment, including their guitar, pedal effects, and amp simulators. Participants then connected their instruments to our audio interface using mono TRS-TRS cables and donned headphones, through which they received a mix of the system's sound and their guitar. A sound-check session with an example track (not used in the experimental sessions) followed, allowing participants to adjust the volume and tone of their instruments.

The experiment involved two conditions, 2D and MR, presented in a randomized order among participants. Each condition included four pieces, also presented in a randomized order, resulting in a total of eight trials (four per condition). While four pieces were used as stimuli, they were treated as variations to allow participants to engage with full musical pieces while avoiding mental and physical fatigue from repetition.

An experimenter assisted participants throughout the experiment, acting as both director and audio engineer, and helped with wearing and removing headphones and the HMD when needed. Participants took a 10 min break between each trial. After completing each trial,

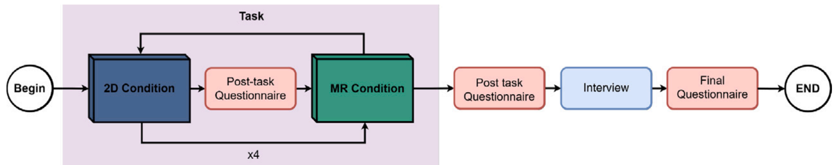

Fig. 7. The procedure used in the experiment, illustrating the sequence of tasks participants engaged in during the study.

they filled out an online questionnaire on a laptop. Upon completing all trials, participants participated in a semi-structured interview conducted by the experimenter and filled out a demographic questionnaire, including age and experience with guitar playing, networked music systems, and immersive systems. The experiment lasted for approximately 2 h. The procedure is summarized in Fig. 7.

3.10. Collected data

During the experiment, we collected two types of data: quantitative and qualitative, which helped us to uncover different aspects of the user experience.

3.10.1. Quantitative data

We collected two types of quantitative data: the Quantity of Motion (QoM) of participants' bodies and heads, and the rotation of each guitar player's head. The QoM serves as an overall measure of the amount of motion detected in the musicians. The body QoM was obtained by analyzing video recordings of each participant, captured with a digital video camera (Canon Legria HF R46) mounted on a tripod, which was manually started and stopped at the beginning and end of each trial. The rotation and QoM of participants' heads were derived from recordings obtained with a MetaMotionRL (MMRL) head tracker by Mbient Lab,[4] placed on the top-center of the headphones. This setup allowed us to compare head motion and rotation between the two systems. To log the data in real-time, we used the MMRL2OSC application (Tomasetti and Turchet, 2023). The head tracker transmitted the rotations as quaternions to a custom C# application every 20 ms, due to the Bluetooth connection's latency and the refresh rate of 200 Hz introduced by the MMR2OSC software used for OSC streaming.

3.10.2. Qualitative data

While objective measurements can effectively evaluate performance outcomes, subjective feedback can capture participants' perspectives on their musical experience. We adapted a questionnaire from Delle Monache et al. (2019), specifically developed for evaluating presence in NMP systems. Presence is also a crucial element in MR. Although the original questionnaire was used in 2D video-mediated NMP, it is based on established presence questionnaires, namely the Witmer–Singer Presence Questionnaire (Witmer and Singer, 1998) and the IPQ Igroup Presence Questionnaire (Schubert et al., 2001), which are extensively used in MR scenarios. Items were selected, rephrased, and adapted to suit the specific context of performing music in a networked environment.

The original questionnaire comprised 27 items. However, since it was designed for real-time scenarios, we removed questions related to network latency and jitter on audio and video streaming. This resulted in a questionnaire composed of 14 items. It covers three main areas: the feeling of co-presence (Q1–Q5), the coherence of the experience (Q6–Q9), and the sense of immersion produced by the systems (Q10–Q14). Participants were asked to respond to each question using an 11-point Likert Scale. The fourteen items of the questionnaire are presented in Table 2.

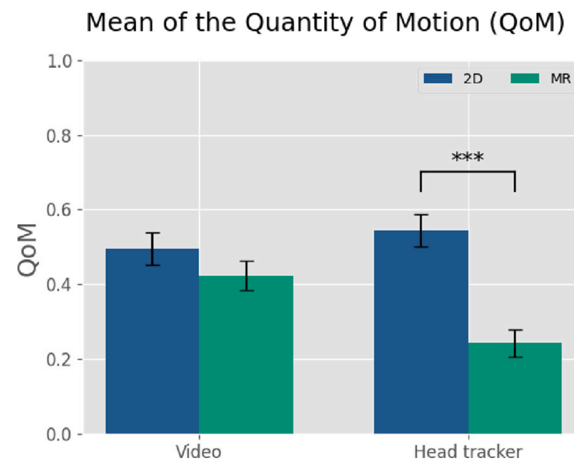

Fig. 8. Mean and standard error of the QoM results from the video recordings and the head tracker. The symbol *** refers to $p < 0.001$.

Additionally, to gain further insights from the participants, we conducted a series of semi-structured interviews at the end of each experiment. These interviews were recorded as audio and later transcribed.

4. Results

4.1. Quantity of motion

The results of the analysis are presented in Table 3 and Fig. 8.

4.1.1. Analysis of video data

To assess the differences in overall body movement between the 2D and MR conditions, we calculated the QoM from the video recordings of participants. We developed a custom Python script (v3.10.4) using the OpenCV library (v4.9.0.80) to batch-process all the video data collected during the experiment.

The QoM was determined by computing the absolute difference in pixels between consecutive frames. A threshold was applied to the difference image to create a binary image, where pixel values above the threshold indicated motion and values below indicated no motion. The number of non-zero pixels in the thresholded image, representing detected motion, was counted. This count was accumulated over a specified number of 10 frames to calculate the total motion in the sequence. The resulting value was then normalized, and we calculated the average of all values to obtain the total QoM. Our approach was based on the work of Camurri et al. (2005) and Jensenius et al. (2005), who proposed similar workflows for computing QoM in the musical context from monocular video recordings of performers.

The resulting data were first normalized across subjects and then analyzed using a linear mixed-effects model, with the QoM variable and conditions (2D, MR) as fixed factors, and the subject as a random factor. An ANOVA was performed on the fitted model. Although the mean QoM for the 2D condition appeared slightly higher compared to the MR condition, no significant effect was found.

[4] https://mbientlab.com/metamotionrl/ (Last accessed: 01/10/2025).

Table 2
The fourteen items composing the post-task questionnaire used during the user study, which was adapted from the one described in Delle Monache et al. (2019).

Post-task questionnaire	
Q1	*While I was playing, I felt like I was in the same room with the other musicians*
Q2	*I felt connected to other musicians*
Q3	*The visual representation made me feel involved in the remote environment*
Q4	*The auditory representation made me feel involved in the remote environment*
Q5	*I felt involved in playing in this condition*
Q6	*I played naturally with this system*
Q7	*How aware were you of the physical world around you during the performance?*
Q8	*How much did your musical experience in the remote environment seem consistent with your real world experiences?*
Q9	*How realistic did the remote environment feel to you?*
Q10	*How quickly did you adapt to the experience in the remote environment?*
Q11	*How aware were you of the devices used in the experience?*
Q12	*How well could you concentrate on the music performance rather than on the mechanisms required to perform?*
Q13	*How much did the quality of the visual representation interfere with or distract you from the performance?*
Q14	*How much did the quality of the auditory representation interfere with or distract you from the performance?*

Table 3
Mean and standard deviation for the QoM of body and head, along with the results of the statistical analysis for the 2D and MR conditions.

QoM	Condition		ANOVA		Effect size
	2D	MR	F(1,11)	p-value	Cohen's d
Video	0.496 ± 0.353	0.423 ± 0.338	1.4071	0.2378	0.21
Head tracker	0.544 ± 0.313	0.244 ± 0.289	32.01	0.001	1.0

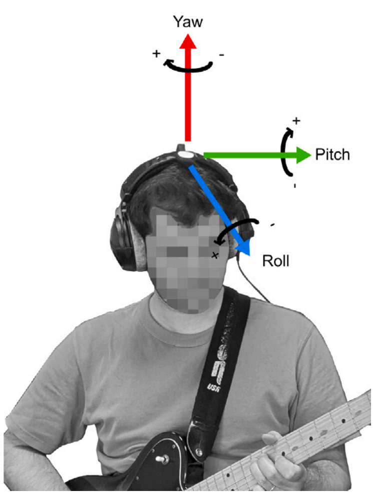

Fig. 9. The Yaw, Pitch, and Roll data derived from the head tracker, illustrating the participants' head orientation and movement patterns during the experiment.

4.1.2. Analysis of head tracker data

We calculated the QoM for each trial to investigate the amount of participants' movement within the 2D and MR conditions. First, we computed the cumulative sum of the absolute values of the derivatives of the norm of the quaternions (saved at intervals of 100 ms). Then, the data were normalized across participants and analyzed using a linear mixed-effects model, with the QoM variable and condition (2D, MR) as fixed factors, and the subject as a random factor. An ANOVA was performed on the fitted model, revealing a significant main effect ($F(1, 11) = 32.01$, $p < 0.001$; Cohen's $d = 1.00$). A post-hoc test adjusted with the Tukey correction showed that the 2D condition had a higher QoM compared to the MR condition ($p < 0.001$).

These results suggest that the type of apparatus used had a significant effect on the head QoM, with the 2D condition outperforming the MR condition. This indicates that users may experience greater movement when using the 2D condition compared to the MR condition.

4.2. Head orientation

Head orientation provides valuable information, such as the direction of participants' gaze, which can be a useful measure of attention. This can be inferred from the head tracker's rotation and direction data. To determine the direction participants' heads were pointing and assess any effects between the 2D and MR conditions, we converted the head tracker data from quaternions to Yaw, Pitch, and Roll, allowing us to handle the data in degrees (see Fig. 9).

The distribution of the mean orientation for Yaw, Pitch, and Roll for the two conditions is presented in Fig. 10. For Yaw, the distribution centers around a positive value, particularly in the MR condition, with the mean Yaw above zero for both 2D and MR. This suggests a slight tendency for head turns to the right in MR. For Pitch, the distribution skews toward negative values, with the mean slightly below zero in both conditions, indicating a tendency for the head to tilt downward. For Roll, the distribution is symmetrical around zero, with the mean close to zero in both 2D and MR, indicating no significant tilt.

We assume that when participants looked straight ahead, the data for Yaw, Pitch, and Roll should show a peak between $-10°$ and $10°$. However, it can be observed that, especially for Yaw in the 2D condition, the distribution of angles is more spread out, in contrast to MR, where they are more concentrated around zero, indicating less movement.

To investigate the effect of the 2D and MR conditions on the mean head orientation as measured by the head tracker (for Yaw, Pitch, and Roll), a linear mixed-effects model was fitted with the two conditions as a fixed effect and subjects as a random effect. An ANOVA was performed on the fitted model, which revealed no significant effect.

4.3. Head movement dynamics

To investigate the temporal dynamics of head movements across the two conditions we employed Detrended Fluctuation Analysis (DFA). DFA is particularly suitable for quantifying long-range temporal correlations in physiological signals and has been used to study motor control behaviors in music performance (González Sánchez et al., 2020; Moumdjian et al., 2020). This method provides a scaling exponent (α) that characterizes the temporal correlation structure of the movement patterns. The results of the analysis are summarized in Table 4.

For this analysis, we used the head movement data recorded in three rotational dimensions (Yaw, Pitch, and Roll). First, the angular measurements were unwrapped and normalized (z-score), then we computed a global movement metric as the Euclidean norm of the three dimensions. Temporal correlations in head movements were analyzed using the DFA. It computed the cumulative sum of the mean-centered signal, divided it into non-overlapping segments, and calculated fluctuations around linear trends at different time scales. The scaling exponent (α) was determined through linear regression in log–log space.

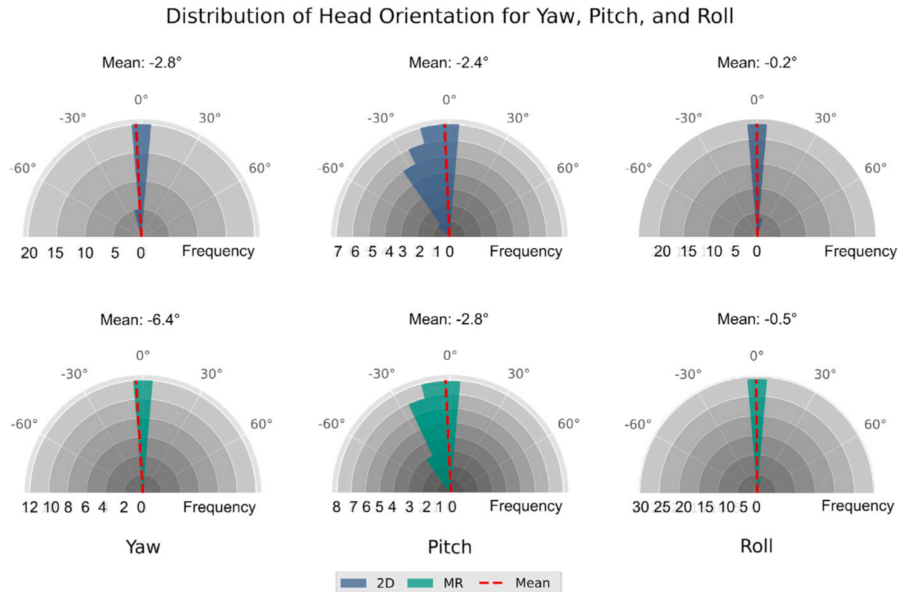

Fig. 10. A visualization of the head tracker's orientation and rotation (Yaw, Pitch, Roll) for both the 2D and MR conditions, highlighting the differences in head movement and orientation between the two experimental setups.

Table 4
Detrended Fluctuation Analysis (DFA) results by conditions.

DFA	Condition		Mann–Whitney U test	
	2D	MR	u-statistic	p-value
DFA scaling exponent (α)	1.404 ± 0.01	1.445 ±0.008	1413.0	0.0028
R-squared	0.984 ± 0.001	0.987 ± 0.001	1413.0	0.003

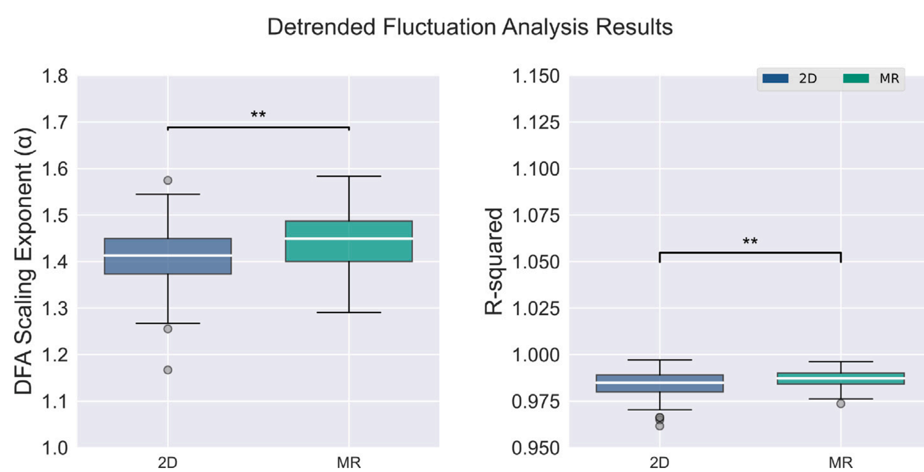

Fig. 11. The results of the Detrended Fluctuation Analysis for the 2D and MR conditions. On the left are the results of the scaling exponent (α), on the right are the results of the R squared. The symbol ** refers to p < 0.01.

Statistical comparisons between conditions were performed using a non-parametric Mann–Whitney U test.

The DFA analysis revealed small but statistically significant differences in head movement patterns between 2D and MR conditions (see Fig. 11). In the MR condition, participants exhibited a marginally higher scaling exponent (α = 1.445 ± 0.008) compared to 2D (α = 1.404 ± 0.01; p = 0.0028). Both conditions showed persistent head movement patterns, with very high R-squared values (2D: 0.984 ± 0.001; MR: 0.987 ± 0.001), suggesting slightly more robust scaling behavior in the MR condition.

4.4. Post-task questionnaire

We analyzed the results of each questionnaire item using generalized linear mixed-effects models, with the condition as a fixed factor and subjects as a random factor. An ANOVA was performed on the

Table 5
The table includes the mean and standard deviation of participants' answers for both 2D and MR conditions. The table also includes the results of the statistical analysis, highlighting significant differences between the conditions.

Questionnaire item	Condition		Type II Wald Chi-square test		Effect size
	2D	MR	$\chi^2(2)$	p-value	Cohen's d
Q1	4.48 ± 2.66	6.58 ± 2.16	32.612	0.0001	−1.01
Q2	4.67 ± 2.67	5.92 ± 2.3	14.341	0.0002	−0.669
Q3	4.92 ± 2.33	6.55 ± 2.56	24.065	0.0001	−0.867
Q4	7.22 ± 2.21	7.76 ± 2.13	3.7083	0.0568	−0.343
Q5	6.2 ± 2.31	6.55 ± 2.32	1.2817	0.26	−0.2
Q6	7.7 ± 1.98	6.72 ± 2.00	53.513	0.0001	1.29
Q7	7.75 ± 2.3	4.64 ± 3.02	17.221	0.0001	−0.734
Q8	5.72 ± 2.23	5.34 ± 2.44	1.0639	0.3046	0.182
Q9	5.83 ± 2.35	6.11 ± 2.36	0.6841	0.41	−0.146
Q10	8.44 ± 1.74	7.34 ± 1.87	23.319	0.0001	0.854
Q11	7.64 ± 2.38	6.42 ± 2.79	15.634	0.0001	0.699
Q12	7.83 ± 2.23	6.23 ± 2.15	27.301	0.0001	0.924
Q13	1.78 ± 2.05	5.25 ± 2.77	82.552	0.0001	−1.61
Q14	1.47 ± 2.02	1.33 ± 1.76	0.4714	0.4938	0.121

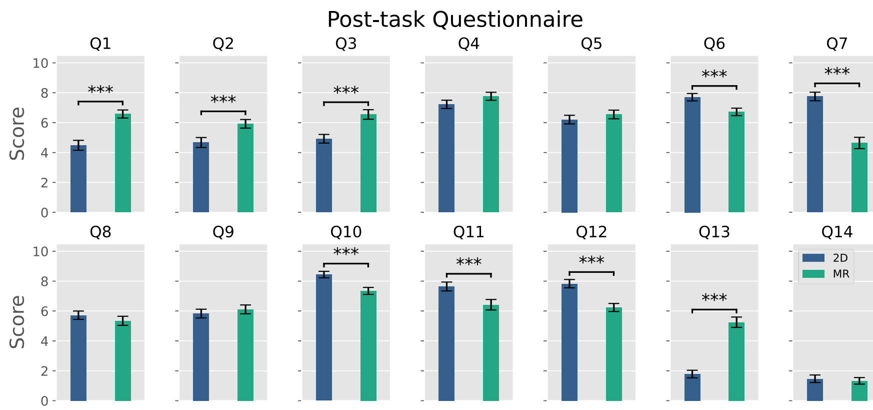

Fig. 12. Mean and standard error for the post-task questionnaire items for the 2D and MR conditions. Legend: *** = p < 0.001.

fitted model, followed by post-hoc tests using pairwise comparisons adjusted with Tukey's correction. A detailed summary of the statistics for each item is shown in Table 5, while the main results are presented in Fig. 12.

The results indicate a significantly higher scoring for MR ($p < 0.001$) in terms of presence and the contribution of the visual representation (Q1–Q3). Conversely, we found a significant effect for 2D ($p < 0.001$) regarding the coherence of the experience for guitar players and the perceived quality of their performance (Q6–Q7). For items related to the sense of immersion (Q10–Q13), we found a significant effect for 2D ($p < 0.001$), indicating that the technical setup used in the MR condition had a negative impact on the experience.

4.5. Thematic analysis

We analyzed the results of the semi-structured interviews using a reflexive thematic analysis (Braun and Clarke, 2019) based on Grounded Theory (Glaser and Strauss, 2017). The analysis was conducted by the authors, who generated codes that were further organized into three themes reflecting shared patterns (User Experience, Technical Challenges, and Future Potential and Recommendations) and in their sub-themes.

- **User Experience:** Thirteen participants commented on the user experience regarding the two systems.
  - *Involvement and Engagement*: Ten participants reported that both systems were engaging and promoted their involvement. Nine participants found that MR led to higher levels of involvement in the task. MR was described as very engaging because the interaction with other musicians through the HMD felt very realistic. It offered an experience of being more immersed in the task, giving the sensation of having the other musicians physically present in the space, thus closely approximating a face-to-face experience (e.g., *"I get a sense of involvement with the holograms, it feels like a normal situation, like when you play with others"*; *"With the HMD, it was very engaging, as it seems like you have the musicians in the same room"*). Five participants found 2D to be less engaging compared to MR. They felt the experience was akin to passively watching a video, reminiscent of a video meeting (e.g., *"It doesn't help to create a sense of connection between musicians, it's difficult, so when I was playing, I was concentrating more on the sound and less on the visual part"*). However, three participants commented that seeing remote musicians through a GUI and a flat monitor was a more reassuring and practical way to use an NMP system, and more suitable for daily use (e.g., *"I already know such systems very well, since we used them to make lessons online"*).
  - *Auditory Feedback*: Eleven participants focused on the audio experience with the two systems. Nine reported that they did not notice a significant difference between the two audio rendering systems, or if there was a difference, it was minor and did not impact their experience (e.g., *"I had the impression that something was changing, but I was focused on playing"*). Three participants commented that binaural audio in MR was not particularly interesting or essential for enhancing the performance and experience (e.g., *"For me, the involvement is made possible by the visual part [...] the normal audio, I mean stereo, was already good for me"*).
  - *User Comfort*: Eleven participants commented on the comfort of the MR system experience. Seven participants found

the headset's field of view (FOV) problematic, particularly the limited lateral and lower vision. This was especially challenging for participants who needed to look at the neck of their guitars during certain parts of the piece, such as chord changes or faster passages. They reported that this limitation caused them to turn their heads more than usual, leading to a perception of increased errors (e.g., *"Especially when playing chords, I have to look at the neck of the guitar to see where I am [...] and with the headset, it's very difficult"*). Four participants noted that the headset's weight made it difficult to immerse themselves in the experience, making playing less comfortable compared to the 2D condition (e.g., *"Because of its weight, I clearly felt that something was on my head, which limited my identification with the situation"*).

- **Issues and Challenges:** Thirteen participants commented on the issues related to performing and playing with other people over the network, along with the technical challenges that should be addressed in both systems, but especially in MR.
  - *The irreplaceable value of physical presence*: Five participants emphasized that, regardless of their opinions on the two systems, the experience of playing face-to-face with other musicians in the same space is irreplaceable (e.g., *"It seems that I am listening to a recording through headphones. Not bad, but playing live is a different thing"*).
  - *Improve MR visual rendering*: Eight participants noted that the visual experience in MR was not optimal. To be more effective, improvements are needed in resolution, color saturation, and the reduction of perceived flickering during movement. Additionally, the point clouds were not detailed enough to convey cues, such as seeing the notes played on the bass at a given moment, like during a chord change, and they should be fully volumetric (e.g., *"The fact that I could not see the hands of the musicians was a problem for coordination"*)
- **Future Potential and Recommendations:** All participants shared their thoughts on future applications and the potential usefulness of these systems. For the 2D system, seven participants mentioned its potential for musical interactions that require a more accessible and straightforward experience. Of these, six participants believed it could be effective for remote learning, such as between a teacher and students. Four participants thought this system was well-suited for group rehearsals, while three suggested its use for remote recordings with other musicians. Regarding the MR system, ten participants commented on its potential applications. They highlighted that MR could be beneficial for experiences requiring more interactivity and immersion. Nine participants noted that MR could be extremely useful for collaborative activities, such as group rehearsals, and could also be interesting for collective jam sessions (e.g., *"If I have to rehearse for a concert, I would use the HMD [...] because being able to fully see the others and potentially interact with them is very important to play well"*). Unlike the 2D system, only one participant suggested MR for music learning and recordings, while two participants pointed out that MR could be intriguing for exploring social and musical interactions that are not merely replicas of existing ones but offer new possibilities.

5. Discussion

This study examined the effects of 2D and MR systems on NMP through a simulated experience using a system for recording and playing back audiovisual stimuli. Through analysis of both quantitative metrics — such as QoM and head orientation — and qualitative feedback from thematic analysis, we gained a comprehensive understanding of how these systems influence user experience in terms of physical engagement and visual attention. These findings have important implications for the design of future NMP technologies for the MM.

5.1. Immersion versus physical constraints in MR condition

Regarding the first research question (RQ1), a key finding of this study was the enhanced sense of immersion and presence reported by participants while using the MR system. This supports previous research showing that MR environments, by simulating a shared physical space, can significantly increase the feeling of being "together" with other performers (Turchet et al., 2022; Schlagowski et al., 2023). The thematic analysis revealed that participants often described the MR environment as "*engulfing*" and "*realistic*", contributing to higher engagement and satisfaction. The qualitative analysis highlighted several challenges related to the immersion provided by the MR system. Some participants found that the experience in the MR condition fell short of their expectations regarding the quality of visual and auditory components. These findings align with the quantitative results, which showed that while MR enhanced immersion, it also imposed physical constraints that negatively affected the overall user experience. The 2D condition, though less immersive, proved easier for participants to adapt to and required less cognitive and physical effort. Participants noted that the familiarity of the 2D setup made it more comfortable for long-term use, even though it did not provide the same level of immersion and engagement as MR. This suggests that while MR systems offer unique advantages, they need careful design to avoid overwhelming users and to support natural, sustained interaction.

5.2. Visual attention and movement patterns

Our findings regarding RQ3 and RQ4 revealed interesting patterns in user movement and attention. The video analysis showed no significant difference in overall body motion between 2D and MR conditions, suggesting that the system type does not substantially affect motion quantity. While QoM is traditionally used as a measure of engagement and expression (Davidson, 1993; Dahl and Friberg, 2007; Bishop et al., 2021), the nature of our task may have influenced participants' behavior, as it did not actively promote movement. The head movement analysis revealed significantly less motion in the MR condition compared to 2D, likely due to the physical limitations of the HMD. The thematic analysis supported this finding, with participants reporting discomfort and restricted movement due to the limited FOV and HMD weight, which affected their natural performance. Importantly, our DFA results showed that head movements in the MR condition exhibited a higher scaling exponent compared to 2D, indicating more persistent and structured movement patterns. This suggests that while MR reduced overall head motion, it led to more coordinated and deliberate head movements. While wearing an HMD, guitar players may need to control their head movements more carefully to maintain visual stability. In addition, players might be more conscious of their head movements in the MR condition, leading to less but more structured head movement patterns. An explanation can be found in the results of the head orientation analysis. It showed a consistent downward tilt across both conditions, with a slight leftward inclination in MR. This aligned with participants' reported difficulties in adapting to the MR system's visual setup. As shown in Fig. 10, guitar players tended to rotate their heads slightly left (Yaw) and tilt downward (Pitch) to view the frets, with a slight increase in these movements in MR. While guitarists primarily use central vision for finger positions, peripheral vision plays a crucial role in maintaining awareness of the fret board, other players, and the environment. The questionnaire results indicated that participants found the visual appearance in MR more problematic, though still acceptable. This was attributed to a combination of FOV, resolution, and point cloud quality. The HMD used had a horizontal FOV of 97° and vertical FOV of 94°, but the ZED Mini camera for passthrough reduced this to 60° horizontally and 90° vertically, with black pixels filling the remaining display area. Similar challenges were reported by participants in Schlagowski et al. (2023), where the HTC

Vive Pro's passthrough had an estimated FOV of 98° horizontally and vertically.

These findings raise important questions about the suitability of current HMDs with passthrough for musicians. While passthrough technology is gaining popularity in consumer devices like the Meta Quest 3, Varjo XR-3, and Apple Vision Pro, our setup's weight of 662 g aligns with typical HMD weights. Previous research showed that increased HMD weight makes distance judgment more difficult (Buck et al., 2018), with HMDs weighing around 570 g or more linked to increased fatigue (Ito et al., 2021) and higher risk of neck injury during rapid head movements (Astrologo et al., 2024). Research has also shown that narrow FOV in VR and AR can impair visual search (Trepkowski et al., 2019) and distance perception (Pfeil et al., 2021). Users with narrow FOV tend to move their heads more quickly during visual tasks (Wille et al., 2014), while broader FOV results in fewer but faster head movements (Venturino and Wells, 1990). While a minimum FOV of 110° is recommended (Pfeil et al., 2021), its impact on musical tasks remains unexplored. Recent studies comparing see-through goggles with hologram-like setups for virtual musician interactions (Weng et al., 2023b,a) found that unrestricted view without goggles is significantly more effective. As research on MR passthrough devices expands into ecological settings (Bailenson et al., 2024), our results emphasize the need for cross-device comparisons and established design thresholds.

5.3. Audio perception and spatial experience

Our investigation of RQ4 revealed unexpected findings regarding binaural audio implementation. Despite previous research showing the benefits of binaural audio in NMP (Tomasetti and Turchet, 2023), participants reported minimal difference between stereo and binaural conditions. This finding warrants deeper analysis. We attribute this result to several factors. First, cross-modal attentional capture (Mast et al., 2017) likely played a significant role, with the visual novelty and cognitive demands of the MR environment dominating participants' attention. This aligns with research showing that congruent visual stimuli can capture attention and reduce auditory information processing (Spence, 2010). While we aligned sound sources with each musician's point cloud, the characteristics of the MR interface (weight and field of view of the HMD) may have overshadowed the perception of subtle audio spatial cues. Audio perception may also have been influenced by the task design. Unlike previous studies showing the benefits of binaural audio (Tomasetti and Turchet, 2023), our participants were actively performing, with the physical and cognitive demands of playing guitar while wearing an HMD potentially limiting attention to spatial audio nuances. Additionally, our participants' professional expertise may have influenced these results, as experienced musicians often rely more on musical timing and visual communication than spatial cues (Davidson and Good, 2002).

Lastly, our findings show similarities with recent research in AR/VR social applications (Arboleda et al., 2024; Immohr et al., 2024b). For instance, multiple studies by Immohr et al. (2023) and Immohr et al. (2024a) consistently found no significant differences in social presence, plausibility, or communication behavior between spatial and non-spatial audio conditions, despite participants' preference for spatial audio in direct listening tests. While these studies did not focus specifically on musical activities and on diadic collaboration, these findings across different domains (AR storytelling, VR communication, and MR music performance) suggest that in focused, task-oriented scenarios, the cognitive demands of the primary activity may overshadow the benefits of spatial audio, particularly in small-group interactions. This pattern however might provide supporting evidence to our interpretation that the visual and physical demands of musical performance in MR may have dominated participants' attention, limiting their ability to perceive and benefit from spatial audio enhancements.

5.4. Design implications

The integrated analysis of QoM, head orientation, and thematic feedback suggests several key recommendations for future immersive NMP systems:

- *Enhancing Ergonomics in MR:* Future MR systems should prioritize reducing HMD physical burden and expanding FOV, either through lighter, more comfortable headsets or alternative display technologies that do not restrict movement. Specifically, our findings suggest that HMDs should aim for weights significantly lower than current ones especially where fatigue becomes problematic, and FOV should extend beyond 110 degrees to support natural peripheral vision. Hardware designers should consider musician-specific requirements, such as the need to glance at instruments while maintaining awareness of other performers. This might involve exploring split-weight designs, adjustable FOV settings, or hybrid display technologies that combine different viewing modes.
- *Balancing Immersion with Usability:* While immersion remains a key strength of MR systems, designers must ensure it does not compromise user comfort and natural interaction. This balance could be achieved through adaptive immersion settings that adjust based on performance context and user preferences. For instance, systems could offer different immersion levels for rehearsal versus performance scenarios, or automatically adjust visual density based on the complexity of the musical task. The goal should be to create environments that maintain engagement while supporting the sustained physical and cognitive demands of musical performance.
- *Improving Visual and Auditory Fidelity:* High-quality audiovisual feedback is essential for maintaining immersion, particularly in MR environments where detailed cues are crucial for effective collaboration. Future systems should focus on enhancing point cloud resolution and stability, reducing visual artifacts, and implementing more sophisticated audio spatialization that adapts to musical activities through a context-dependent approach. This approach might optimize resource allocation by using spatial audio only for complex scenarios — such as situations requiring speaker identification for verbal communication or when musicians need to move in space — while defaulting to stereo for simpler, resource-constrained environments. However, key questions remain unresolved, including whether spatial audio should apply to speech only, musical instrument sounds, or both, as these aspects were not addressed in our study and warrant future research.
- *Supporting Adaptation and Learning:* Given the adaptation challenges participants faced, future designs should incorporate features that facilitate easier learning. This could include interactive tutorials that guide users through basic interactions, virtual rehearsal spaces where musicians can practice with the system before actual performances, and the gradual introduction of advanced features. The interface should offer customizable settings for immersion levels, visual density, and audio spatialization, allowing users to progressively adapt to the system's capabilities. Additionally, the system should provide real-time feedback about optimal positioning and movement ranges, helping users develop comfortable and sustainable performance practices in the MR environment.

5.5. Limitations

A primary limitation of our study is the participant sampling, which focused specifically on expert electric guitar players with 10+ years of professional experience. Out of sixteen recruited participants, two were female. While this ratio aligns with historical demographic patterns in professional electric guitar performance (see Bourdage (2010) and Bayton (2013), Oe and Yamaoka (2023)), it may limit the generalizability of our findings across different musician populations. Future research should examine these interfaces with more diverse instrument groups and skill levels to build a more comprehensive understanding of MR music performance.

A second limitation is our use of simulated experiences through prerecorded content rather than real-time transmission of audio–visual data. This methodological choice was driven by several factors: the current lack of reliable solutions for true NMP combined with MR technologies, and the significant issues with latency and synchronization reported in previous studies (Schlagowski et al., 2023; Turchet et al., 2022). Rather than limiting our experiment to two connected participants, we chose to use pre-recorded material to enable the study of larger ensemble configurations while maintaining consistent experimental conditions. However, we acknowledge that simulation cannot fully capture the dynamics of live musical interaction, particularly the subtle adaptations that occur between performers in real-time.

Additionally, the design of our study (both in terms of settings and exposure to the technologies) may not have captured specific ergonomic effects or adaptation patterns that could emerge over extended use of HMDs, which would be critical for understanding long-term viability of MR-NMP systems. This represents an important aspect that deserves consideration in future studies that should consider different time scales and different types of HMDs for MR.

Furthermore, our focus on electric guitar players, while providing consistency for comparison, may not address the unique requirements of other instruments or musical genres. While our findings provide valuable insights into MR music performance, full validation will require real-time systems, although significant challenges remain in both audio (Rottondi et al., 2016) and visual processing (Cortés et al., 2024; Qiu et al., 2023). Future research should address these limitations through longitudinal studies, diverse instrument types, and varied performance contexts.

6. Conclusion

This study explored the integration of MR technologies into NMP systems within the emerging MM. Our findings revealed a crucial balance between immersion and usability in musical performance contexts. The MR system provided a highly immersive environment that enhanced performers' sense of presence and engagement. However, this immersion often came at the expense of physical comfort and natural movement, as participants reported restricted head movements and discomfort due to hardware limitations, particularly the headset's weight and limited FOV.

Our analysis revealed several key insights. First, while MR technology offers promising capabilities for musical collaboration, current hardware constraints significantly impact performers' physical expression and comfort. The reduction in head movement and restricted field of view particularly affected guitarists' ability to view their instruments while maintaining awareness of other performers. Second, contrary to expectations, participants reported limited perceptual benefit from spatial audio features, suggesting that visual aspects of MR may dominate attention in performance contexts. This finding challenges assumptions about the importance of spatial audio in immersive musical experiences and suggests the need for more nuanced audio implementation strategies.

These results have important implications for future MR-NMP system design. Rather than pursuing full immersion at any cost, developers should prioritize ergonomic considerations that support natural movement and extended use; balance audiovisual presentation to enhance rather than hinder performance, and flexible systems that can adapt to different musical contexts and user preferences.

While our simulation-based approach provided valuable insights, future research should focus on developing real-time MR-NMP systems that address current technical limitations in both audio and visual processing. Additionally, studies involving diverse musical instruments, genres, and performer demographics will be crucial for understanding how MR can best serve the broader musical community. As the Musical Metaverse continues to evolve, our findings suggest that successful integration of MR technology in musical performance will require careful attention to both technical capabilities and human factors.

CRediT authorship contribution statement

Alberto Boem: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matteo Tomasetti:** Writing – review & editing, Resources, Conceptualization. **Luca Turchet:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relation-ships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the European Innovation Council can be held responsible for them.

We thank Chiharu Seki for her assistance with the illustrations, and Pietro Gola for his help during the development of the systems used in the study. We also thank Luca Vignati and Leonardo Collizzolli for assisting with the recordings.

Data availability

Data will be made available on request.

References

Abdallah, M., Prabhakaran, B., Cai, W., Hsu, C.-H., 2023. Recent advances in immersive multimedia. IEEE MultiMedia 30 (2), 5–7.
Arboleda, S.A., Kunert, C., Hartbrich, J., Schneiderwind, C., Diao, C., Gerhardt, C., Surdu, T., Weidner, F., Broll, W., Werner, S., et al., 2024. Beyond looks: a study on agent movement and audiovisual spatial coherence in augmented reality. In: 2024 IEEE Conference Virtual Reality and 3D User Interfaces. VR, IEEE, pp. 502–512.
Argyle, M., 2013. Bodily Communication. Routledge.
Astrologo, A.N., Nano, S., Klemm, E.M., Shefelbine, S.J., Dennerlein, J.T., 2024. Determining the effects of AR/VR HMD design parameters (mass and inertia) on cervical spine joint torques. Appl. Ergon. 116, 104183.
Bailenson, J.N., Beams, B., Brown, J., DeVeaux, C., Han, E., Queiroz, A.C.M., Ratan, R., Santoso, M., Srirangarajan, T., Tao, Y., Wang, P., 2024. Seeing the world through digital prisms: Psychological implications of passthrough video usage in mixed reality. Technol. Mind Behav. 5 (2: Summer 2024).
Bayton, M., 2013. Women and the electric guitar. In: Sexing the Groove. Routledge, pp. 37–49.
Behne, K.-E., Wöllner, C., 2011. Seeing or hearing the pianists? A synopsis of an early audiovisual perception experiment and a replication. Music. Sci. 15 (3), 324–342.
Billinghurst, M., Kato, H., 1999. Collaborative mixed reality. In: Proceedings of the First International Symposium on Mixed Reality. pp. 261–284.
Bishop, L., 2024. Togetherness in musical interaction. Routledge Open Res. 3 (16), 16.
Bishop, L., Goebl, W., 2015. When they listen and when they watch: Pianists' use of nonverbal audio and visual cues during duet performance. Music. Sci. 19 (1), 84–110.
Bishop, L., Gonzalez Sanchez, V., Laeng, B., Jensenius, A.R., Høffding, S., 2021. Move like everyone is watching: Social context affects head motion and gaze in string quartet performance. J. New Music Res. 50 (4), 392–412.

Boem, A., Tomasetti, M., Turchet, L., 2024. Harmonizing the Musical Metaverse: unveiling needs, tools, and challenges from experts' point of view. In: Proceedings of the International Conference on New Interfaces for Musical Expression.
Boem, A., Tomasetti, M., Turchet, L., 2025. Issues and Challenges of Audio Technologies for the Musical Metaverse. J. Audio Eng. Soc. 73 (3), 94–114.
Bourdage, M., 2010. "A Young girls dream": Examining the barriers facing female electric guitarists. IASPM J. 1 (1), 1–16.
Braun, V., Clarke, V., 2019. Reflecting on reflexive thematic analysis. Qual. Res. Sport. Exerc. Heal. 11 (4), 589–597.
Broughton, M., Stevens, C., 2009. Music, movement and marimba: An investigation of the role of movement and gesture in communicating musical expression to an audience. Psychol. Music. 37 (2), 137–153.
Buck, L.E., Young, M.K., Bodenheimer, B., 2018. A comparison of distance estimation in HMD-based virtual environments with different HMD-based conditions. ACM Trans. Appl. Percept. (TAP) 15 (3), 1–15.
Burger, B., Wöllner, C., 2023. Drumming action and perception: how the movements of a professional drummer influence experiences of tempo, time, and expressivity. Music. Sci. 6, 20592043231186870.
Cáceres, J., Chafe, C., 2010. JackTrip: Under the hood of an engine for network audio. J. New Music Res. 39 (3), 183–187.
Cairns, P., Hunt, A., Johnston, D., Cooper, J., Lee, B., Daffern, H., Kearney, G., 2023. Evaluation of metaverse music performance with bbc maida vale recording studios. J. Audio Eng. Soc. 313–325.
Camurri, A., Mazzarino, B., Volpe, G., 2005. A tool for analysis of expressive gestures: The EyesWeb expressive gesture processing library. Ret 18 (2).
Carôt, A., Schuller, G., 2011. Towards a telematic visual-conducting system. J. Audio Eng. Soc. (16).
Cortés, C., Viola, I., Gutiérrez, J., Jansen, J., Subramanyam, S., Alexiou, E., Pérez, P., García, N., César, P., 2024. Delay threshold for social interaction in volumetric extended reality communication. ACM Trans. Multimed. Comput. Commun. Appl. 20 (7), 1–22.
Dahl, S., Friberg, A., 2007. Visual perception of expressiveness in musicians' body movements. Music. Percept. 24 (5), 433–454.
D'Amario, S., Daffern, H., Bailes, F., 2018. Synchronization in singing duo performances: The roles of visual contact and leadership instruction. Front. Psychol. 9, 1208.
Davidson, J.W., 1993. Visual perception of performance manner in the movements of solo musicians. Psychol. Music. 21 (2), 103–113.
Davidson, J.W., 1994. What type of information is conveyed in the body movements of solo musician performers. J. Hum. Mov. Stud. 6, 279–301.
Davidson, J.W., Good, J.M., 2002. Social and musical co-ordination between members of a string quartet: An exploratory study. Psychol. Music. 30 (2), 186–201.
Delle Monache, S., Comanducci, L., Buccoli, M., Zanoni, M., Sarti, A., Pietrocola, E., Berbenni, F., Cospito, G., 2019. A presence-and performance-driven framework to investigate interactive networked music learning scenarios. Wirel. Commun. Mob. Comput. 2019 (1), 4593853.
Drioli, C., Allocchio, C., Buso, N., 2013. Networked performances and natural interaction via LOLA: Low latency high quality A/V streaming system. In: International Conference on Information Technologies for Performing Arts, Media Access, and Entertainment. Springer, pp. 240–250.
2025. Elk live studio. https://studio.elk.live/. (Accessed 31 July 2025).
Fussell, S.R., Kraut, R.E., Siegel, J., 2000. Coordination of communication: Effects of shared visual context on collaborative work. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. pp. 21–30.
Gabrielli, L., Squartini, S., Gabrielli, L., Squartini, S., 2016. Networked music performance. Wirel. Netw. Music. Perform. 3–19.
Glaser, B., Strauss, A., 2017. Discovery of Grounded Theory: Strategies for Qualitative Research. Routledge.
González Sánchez, V., Żelechowska, A., Jensenius, A.R., 2020. Analysis of the movement-inducing effects of music through the fractality of head sway during standstill. J. Mot. Behav. 52 (6), 734–749.
Gunkel, S.N., Dijkstra-Soudarissanane, S., Stokking, H.M., Niamut, O.A., 2023. From 2d to 3D video conferencing: Modular RGB-D capture and reconstruction for interactive natural user representations in immersive extended reality (XR) communication. Front. Signal Process. 3, 1139897.
Gurevich, M., Donohoe, D., Bertet, S., 2011. Ambisonic spatialization for networked music performance. In: Proceedings of the 17th International Conference on Auditory Display. ICAD2011.
Hevner, K., 1936. Experimental studies of the elements of expression in music. Am. J. Psychol. 48 (2), 246–268.
Hopkins, T., Weng, S.C.C., Vanukuru, R., Wenzel, E.A., Banic, A., Gross, M.D., Do, E.Y.-L., 2022. Ar drum circle: Real-time collaborative drumming in ar. Front. Virtual Real. 3, 847284.
Hunt, A., Daffern, H., Kearney, G., 2023. Avatar representation in extended reality for immersive networked music performance. In: Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio. Audio Engineering Society.
Hupke, R., Preihs, S., Peissig, J., 2022. Immersive room extension environment for networked music performance. In: Audio Engineering Society Convention 153. Audio Engineering Society.
Hupke, R., Preihs, S., Peissig, J., 2024. Immersive networked music performance: Impact of extended reality on the quality of experience. In: 2024 IEEE 5th International Symposium on the Internet of Sounds. IS2, IEEE, pp. 1–9.
Immohr, F., Rendle, G., Kehling, C., Lammert, A., Göring, S., Froehlich, B., Raake, A., 2024a. Subjective evaluation of the impact of spatial audio on triadic communication in virtual reality. In: 2024 16th International Conference on Quality of Multimedia Experience. QoMEX, IEEE, pp. 262–265.
Immohr, F., Rendle, G., Lammert, A., Neidhardt, A., Zur Heyde, V.M., Froehlich, B., Raake, A., 2024b. Evaluating the effect of binaural auralization on audiovisual plausibility and communication behavior in virtual reality. In: 2024 IEEE Conference Virtual Reality and 3D User Interfaces. VR, IEEE, pp. 849–858.
Immohr, F., Rendle, G., Neidhardt, A., Göring, S., Ramachandra Rao, R.R., Arevalo Arboleda, S., Froehlich, B., Raake, A., 2023. Proof-of-concept study to evaluate the impact of spatial audio on social presence and user behavior in multi-modal VR communication. In: Proceedings of the 2023 ACM International Conference on Interactive Media Experiences. pp. 209–215.
Ito, K., Tada, M., Ujike, H., Hyodo, K., 2021. Effects of the weight and balance of head-mounted displays on physical load. Appl. Sci. 11 (15).
2025. JackTrip virtual studio. https://jacktrip.github.io/jacktrip/#virtual-studio. (Accessed 31 July 2025).
2025. Jamulus. https://jamulus.io/. (Accessed 31 July 2025).
Jansen, J., Subramanyam, S., Bouqueau, R., Cernigliaro, G., Cabré, M.M., Pérez, F., Cesar, P., 2020. A pipeline for multiparty volumetric video conferencing: transmission of point clouds over low latency DASH. In: Proceedings of the 11th ACM Multimedia Systems Conference. pp. 341–344.
Jensenius, A.R., Godøy, R.I., Wanderley, M.M., 2005. Developing tools for studying musical gestures within the max/MSP/Jitter environment. In: Proceedings of the 2005 International Computer Music Conference.
Kantonen, T., Woodward, C., Katz, N., 2010. Mixed reality in virtual world teleconferencing. In: 2010 IEEE Virtual Reality Conference. VR, IEEE, pp. 179–182.
Liu, Z., Li, Q., Chen, X., Wu, C., Ishihara, S., Li, J., Ji, Y., 2021. Point cloud video streaming: Challenges and solutions. IEEE Netw. 35 (5), 202–209.
Massie-Laberge, C., Cossette, I., Wanderley, M., 2016. Motion kinematics and auditors' perception of expressive gesture in romantic piano performances. In: 14th International Conference for Music Perception and Cognition.
Mast, F., Frings, C., Spence, C., 2017. Crossmodal attentional control sets between vision and audition. Acta Psychol. 178, 41–47.
Moumdjian, L., Maes, P.-J., Dalla Bella, S., Decker, L.M., Moens, B., Feys, P., Leman, M., 2020. Detrended fluctuation analysis of gait dynamics when entraining to music and metronomes at different tempi in persons with multiple sclerosis. Sci. Rep. 10 (1), 12934.
Nusseck, M., Wanderley, M.M., 2009. Music and motion—how music-related ancillary body movements contribute to the experience of music. Music. Percept. 26 (4), 335–353.
Oe, H., Yamaoka, Y., 2023. Exploring socioeconomic factors for the sustainability of electric guitar brands: Focusing on traditional values and psychological attachment. SocioEcon. Chall. 7 (4), 40–57.
Olmos, A., Bouillot, N., Knight, T., Mabire, N., Redel, J., Cooperstock, J.R., 2012. A high-fidelity orchestra simulator for individual musicians' practice. Comput. Music J. 36 (2), 55–73.
Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., Kim, D., Davidson, P.L., Khamis, S., Dou, M., et al., 2016. Holoportation: Virtual 3d teleportation in real-time. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology. pp. 741–754.
Palmer, C., Spidle, F., Koopmans, E., Schubert, P., 2019. Ears, heads, and eyes: When singers synchronise. Q. J. Exp. Psychol. 72 (9), 2272–2287.
Pfeil, K., Masnadi, S., Belga, J., Sera-Josef, J.-V.T., LaViola, J., 2021. Distance perception with a video see-through head-mounted display. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–9.
Qiu, H., Streli, P., Luong, T., Gebhardt, C., Holz, C., 2023. ViGather: Inclusive virtual conferencing with a joint experience across traditional screen devices and mixed reality headsets. Proc. ACM Hum.- Comput. Interact. 7 (MHCI).
Redman, B., 2020. The potential of videoconferencing and low-latency (LoLa) technology for instrumental music teaching. Music. Pr. 6, 15.
Regenbrecht, H., Ott, C., Park, N., Duncan, S., Collins, J., 2021. Voxelvideos for entertainment, education, and training. IEEE Access 9, 68185–68196.
Renaud, A., Carôt, A., Rebelo, P., 2007. Networked music performance: State of the art. In: Proceedings of the AES 30th International Conference. Finland Saariselkä, pp. 16–22.
Rottondi, C., Chafe, C., Allocchio, C., Sarti, A., 2016. An overview on networked music performance technologies. IEEE Access 4, 8823–8843.
Russell, J.A., 1980. A circumplex model of affect. J. Pers. Soc. Psychol. 39 (6), 1161.
Schiavio, A., Maes, P.-J., van der Schyff, D., 2022. The dynamics of musical participation. Music. Sci. 26 (3), 604–626.
Schlagowski, R., Nazarenko, D., Can, Y., Gupta, K., Mertes, S., Billinghurst, M., André, E., 2023. Wish you were here: Mental and physiological effects of remote music collaboration in mixed reality. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–16.
Schubert, T., Friedmann, F., Regenbrecht, H., 2001. The experience of presence: Factor analytic insights. Presence: Teleoperators Virtual Environ. 10 (3), 266–281.

Schütz, A., 1951. Making music together: A study in social relationship. Soc. Res. 76–97.

Smith, H.J., Neff, M., 2018. Communication behavior in embodied virtual reality. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Son, J., Gül, S., Bhullar, G.S., Hege, G., Morgenstern, W., Hilsmann, A., Ebner, T., Bliedung, S., Eisert, P., Schierl, T., Buchholz, T., Hellge, C., 2020. Split rendering for mixed reality: Interactive volumetric video in action. In: SIGGRAPH Asia 2020 XR. SA '20, Association for Computing Machinery, New York, NY, USA.

2025. Sonobus. https://sonobus.net/. (Accessed 31 July 2025).

Speicher, M., Hall, B.D., Nebeling, M., 2019. What is mixed reality? In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–15.

Spence, C., 2010. Crossmodal spatial attention. Ann. New York Acad. Sci. 1191 (1), 182–200.

Timmers, R., Endo, S., Bradbury, A., Wing, A.M., 2014. Synchronization and leadership in string quartet performance: a case study of auditory and visual cues. Front. Psychol. 5, 645.

Tomasetti, M., Turchet, L., 2023. Playing with others using headphones: Musicians prefer binaural audio with head tracking over stereo. IEEE Trans. Hum.-Mach. Syst. 53 (3), 501–511.

Trepkowski, C., Eibich, D., Maiero, J., Marquardt, A., Kruijff, E., Feiner, S., 2019. The effect of narrow field of view and information density on visual search performance in augmented reality. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces. VR, IEEE, pp. 575–584.

Tsay, C.-J., 2014. The vision heuristic: Judging music ensembles by sight alone. Organ. Behav. Hum. Decis. Process. 124 (1), 24–33.

Turchet, L., 2023. Musical Metaverse: vision, opportunities, and challenges. Pers. Ubiquitous Comput. 27 (5), 1811–1827.

Turchet, L., Garau, N., Conci, N., 2022. Networked musical XR: where's the limit? A preliminary investigation on the joint use of point clouds and low-latency audio communication. In: Proceedings of the 17th International Audio Mostly Conference. AM '22, Association for Computing Machinery, New York, NY, USA, pp. 226–230.

Turchet, L., Rinaldi, C., Centofanti, C., Vignati, L., Rottondi, C., 2024. 5G-enabled internet of musical things architectures for remote immersive musical practices. IEEE Open J. Commun. Soc..

Van Kerrebroeck, B., Crombé, K., de Leymarie, S.M., Leman, M., Maes, P.-J., 2024. The virtual drum circle: polyrhythmic music interactions in mixed reality. J. New Music Res. 1–21.

Venturino, M., Wells, M.J., 1990. Head movements as a function of field-of-view size on a helmet-mounted display. In: Proceedings of the Human Factors Society Annual Meeting. Vol. 34, SAGE Publications Sage CA: Los Angeles, CA, pp. 1572–1576.

Vines, B.W., Krumhansl, C.L., Wanderley, M.M., Dalca, I.M., Levitin, D.J., 2011. Music to my eyes: Cross-modal interactions in the perception of emotions in musical performance. Cogn. 118 (2), 157–170.

Viola, I., Jansen, J., Subramanyam, S., Reimat, I., Cesar, P., 2023. VR2Gather: A collaborative, social virtual reality system for adaptive, multiparty real-time communication. IEEE MultiMedia 30 (2), 48–59.

Weng, S.C.-C., Hopkins, T., Ma, S.-Y., Tobin, C., Banic, A., Do, E.Y.-L., 2023a. Investigating the effects of limited field of view on jamming experience in extended reality. In: 2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct. ISMAR-Adjunct, IEEE, pp. 493–495.

Weng, S.C.-C., Hopkins, T., Vanukuru, R., Tobin, C., Banic, A., Leithinger, D., Do, E.Y.-L., 2023b. How field of view affects awareness of an avatar during a musical task in augmented reality. In: 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops. VRW, IEEE, pp. 633–634.

Wille, M., Adolph, L., Grauel, B., Wischniewski, S., Theis, S., Alexander, T., 2014. Prolonged work with head mounted displays. In: Proceedings of the 2014 ACM International Symposium on Wearable Computers: Adjunct Program. pp. 221–224.

Williamon, A., Davidson, J.W., 2002. Exploring co-performer communication. Music. Sci. 6 (1), 53–72.

Witmer, B.G., Singer, M.J., 1998. Measuring presence in virtual environments: A presence questionnaire. Presence 7 (3), 225–240.

Wright, M., 2005. Open sound control: an enabling technology for musical networking. Organised Sound 10 (3), 193–200.

Young, G.W., O'Dwyer, N., Moynihan, M., Smolic, A., 2022. Audience experiences of a volumetric virtual reality music video. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces. VR, IEEE, pp. 775–781.

Zioulis, N., Alexiadis, D., Doumanoglou, A., Louizis, G., Apostolakis, K., Zarpalas, D., Daras, P., 2016. 3D tele-immersion platform for interactive immersive experiences between remote users. In: 2016 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 365–369.

Zotter, F., Frank, M., 2019. Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality. Springer Nature.